Vol.2013-DBS-157 No.12
Vol.2013-IFAT-111 No.12
2013/7/23

How Intuitive Are Diversified Search Metrics? Concordance Test Results for the Diversity U-measures

Tetsuya Sakai[1,a)]

Abstract: For the past few decades, ranked retrieval (e.g. web search) has been evaluated using *rank-based* evaluation metrics such as Average Precision and normalised Discounted Cumulative Gain (nDCG). These metrics discount the value of each retrieved relevant document based on its rank. The situation is similar with diversified search which has gained popularity recently: diversity metrics such as α-nDCG, Intent-Aware Expected Reciprocal Rank (ERR-IA) and D♯-nDCG are also rank-based. These widely-used evaluation metrics just regard the system output as a list of document IDs, and ignore all other features such as snippets and document full texts of various lengths. The recently-proposed U-measure framework of Sakai and Dou uses the *amount of text read by the user* as the foundation for discounting the value of relevant information, and can take into account the user's snippet reading and full text reading behaviours. The present study compares the diversity versions of U-measure (*D-U* and *U-IA*) with state-of-the-art diversity metrics in terms of how "intuitive" they are: given a pair of ranked lists, we quantify the ability of each metric to favour the *more diversified and more relevant* list by means of the concordance test. Our results show that while D♯-nDCG is the overall winner in terms of simultaneous concordance with diversity and relevance, D-U and U-IA statistically significantly outperform other state-of-the-art metrics. Moreover, in terms of concordance with relevance alone, D-U and U-IA significantly outperform all rank-based diversity metrics. These results suggest that D-U and U-IA are not only more realistic than rank-based metrics but also intuitive, i.e., that they measure what we want to measure.

Keywords: diversity, evaluation, intents, TREC, subtopics, web search.

1. Introduction

For the past few decades, ranked retrieval (e.g. web search) has been evaluated using *rank-based* evaluation metrics such as *Average Precision* [7] and *normalised Discounted Cumulative Gain* (nDCG) [8]. These metrics discount the value of each retrieved relevant document based on its rank. The situation is similar with *diversified search* which has gained popularity recently: diversity metrics such as *α-nDCG* [3], *Intent-Aware Expected Reciprocal Rank* (ERR-IA) [2] and *D♯-nDCG* [15] are also rank-based. These widely-used evaluation metrics just regard the system output as a list of document IDs, and ignore all other features such as snippets and document full texts of various lengths.

The recently-proposed *U-measure* framework of Sakai and Dou [12] uses the *amount of text read by the user* as the foundation for discounting the value of relevant information, and can take into account the user's snippet reading and full text reading behaviours. The present study compares the diversity versions of U-measure (*D-U* and *U-IA*) with state-of-the-art diversity metrics in terms of how "intuitive" they are: given a pair of ranked lists, we quantify the ability of each metric to favour the *more diversified and more relevant* list by means of the *concordance test* [11]. Our results show that while D♯-nDCG is the overall winner in terms of simultaneous concordance with diversity and relevance, D-U and U-IA statistically significantly outperform other state-of-the-art metrics. Moreover, in terms of concordance with relevance alone, D-U and U-IA significantly outperform all rank-based diversity metrics. These results suggest that D-U and U-IA are not only more realistic than rank-based metrics but also intuitive, i.e., that they measure what we want to measure.

2. Prior Art

This section discusses existing studies on evaluation metrics for diversified search, which, given an ambiguous and/or underspecified query, aims to satisfy different user intents with a single search engine result page[*1]. While traditional ranked retrieval only considers relevance, diversified search systems are expected to find the right balance between diversity and relevance. In diversified search evaluation, it is assumed that the following are available [15], [16]:

- A set of ambiguous and/or underspecified topics (i.e., queries) $\{q\}$;
- A set of *intents* $\{i\}$ for each topic;
- The *intent probability* $Pr(i|q)$ for each intent;
- Per-intent (possibly graded) relevance assessments for each topic.

Because diversity metrics need to consider the above different factors to evaluate systems, they tend to be more complex than traditional ranked retrieval metrics. However, since the ultimate

[1] Microsoft Research Asia, China
[a)] tetsuyasakai@acm.org

[*1] An example of an ambiguous query would be "office": does the user mean "workplace" or "Microsoft software"? An example of an underspecified query would be "harry potter": *Harry Potter books*? *Harry Potter films*? Or perhaps *Harry Potter the main character*?

goal of Information Retrieval (IR) researchers is to satisfy the user's information need, we want to make sure that the metrics are intuitive, i.e., that the metrics are measuring what we want to measure. This is the focus of the present study.

The TREC[*2] Web Track ran the Diversity Task from 2009 to 2012 [6]. In the present study, we use the TREC 2011 diversity data [4]: only the 2011 and 2012 data have *graded* relevance assessments, and the number of participating teams was higher in 2011 (9 vs. 8). At the TREC 2009 Diversity Task, the primary metric used for ranking the runs was α-nDCG; ERR-IA was used in the subsequent years.

NTCIR[*3] ran the INTENT task [13][*4] at NTCIR-9 and -10, which also evaluated diversified search. The primary evaluation metric used there was D♯-nDCG, which is a simple linear combination of *intent recall* (I-rec) and *D-nDCG* [15]. The NTCIR-10 INTENT-2 task also used additional metrics called *DIN-nDCG* and *P+Q* to evaluate the systems' ability to handle *informational* and *navigational* intents in diversified search. However, this *intent-type-sensitive* evaluation is beyond the scope of this paper, as very few teams have tackled this particular problem so far [13], [19].

In this study, we compare D-U and U-IA with these official diversity metrics from TREC and NTCIR, namely, α-nDCG, ERR-IA and D(♯)-nDCG, from the viewpoint of how "intuitive" they are. We use the *official* α-nDCG and ERR-IA performance values that were computed with the `ndeval` software[*5], as well as D(♯)-nDCG values computed with `NTCIREVAL`[*6]. Below, we formally define these rank-based diversity metrics from TREC and NTCIR.

First, let us define the original nDCG for traditional IR, given graded relevance assessments per topic, where the relevance level x varies from 0 to H. In the present study, $H = 3$ (See Table 1 in Section 5), and $x = 0$ means "nonrelevant." Following previous work (e.g. [1], [2]), we let the *gain value* of each x-relevant document be $gv_x = (2^x - 1)/2^H$: hence $gv_1 = 1/8$, $gv_2 = 3/8$ and $gv_3 = 7/8$. For a given ranked list, the gain at rank r is defined as $g(r) = gv_x$ if the document at r is x-relevant. Moreover, let $g^*(r)$ denote the gain at rank r in an *ideal* ranked list, obtained by sorting all relevant documents by the relevance level [8], [10]. A popular version of nDCG [1] is defined as:

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

where l is the *measurement depth* or *document cutoff*.

In diversified IR evaluation where each topic q has a set of possible intents $\{i\}$, (graded) relevance assessments are obtained for each i rather than for each q. Let $I_i(r)$ be one if the document at rank r is relevant to intent i and zero otherwise; let $C_i = \sum_{k=1}^{r} I_i(k)$. α-nDCG is defined by replacing the gains in Eq. 1 with the following *novelty-biased gain* [3]:

$$ng(r) = \sum_i I_i(r)(1-\alpha)^{C_i(r-1)} \quad (2)$$

where α is a parameter, set to $\alpha = 0.5$ at TREC. Thus it discounts the value of each relevant document based on redundancy within each intent (Eq. 2) and then further discounts it based on the rank (Eq. 1). Although this definition requires the novelty-biased gains for the ideal list ($ng^*(r)$), the problem of obtaining the ideal list for α-nDCG is NP-complete, and therefore a greedy approximation is used in practice [3]. Note that α-nDCG cannot handle per-intent graded relevance: it defines the graded relevance of a document based solely on the number of intents it covers.

In contrast, ERR-IA utilises per-intent graded relevance assessments: let $g_i(r)$ denote the gain at rank r *with respect to intent* i, using the aforementioned gain value setting (i.e., $1/7, 3/8, 7/8$). This may be interpreted as the probability that the user with intent i is satisfied with this particular document at r. Then the ERR for this particular intent, ERR_i, is computed as:

$$ERR_i = \sum_r \prod_{k=1}^{r-1} (1 - g_i(k)) g_i(r) \frac{1}{r} . \quad (3)$$

This is an intuitive metric: the user with intent i is dissatisfied with documents between ranks 1 and $r-1$, and is finally satisfied at r; the utility at this satisfaction point is measured by the reciprocal rank $1/r$. Finally, ERR-IA is computed as the expectation over the intents:

$$ERR\text{-}IA = \sum_i Pr(i|q) ERR_i . \quad (4)$$

The D♯ framework [15] used at the NTCIR INTENT task also utilises per-intent graded relevance assessments. First, for each document at rank r, the *global gain* is defined as:

$$GG(r) = \sum_i Pr(i|q) g_i(r) . \quad (5)$$

Then, by sorting all relevant documents by the global gain, a "globally ideal list" is defined for a given topic, so that the *ideal global gain* $GG^*(r)$ can be obtained. Note that unlike α-nDCG, there is no NP-complete problem involved here, and that, unlike ERR-IA, there is exactly one ideal list for a given topic. By replacing the gains in Eq. 1 with these global gain values, a *D-measure* version of nDCG, namely, *D-nDCG* is obtained. This is further combined with *intent recall*, defined as $I\text{-}rec = |\{i'\}|/|\{i\}|$ where $\{i'\}$ is the set of intents covered by the system output:

$$D\sharp\text{-}nDCG = \gamma I\text{-}rec + (1-\gamma) D\text{-}nDCG . \quad (6)$$

Here, γ is a parameter ($0 \le \gamma \le 1$), simply set to 0.5 at NTCIR. D♯-nDCG is a single-value summary of the *I-rec/D-nDCG graph* used at the NTCIR INTENT task [13], which visualises the trade-off between diversity and overall relevance.

As we shall demonstrate later, α-nDCG and ERR-IA behave very similarly, as they both possess the per-intent *diminishing return* property [2]: whenever a relevant document is found, the value of the next relevant document is discounted for each intent. Because redundancy within each intent is penalised, diversity across intents is rewarded. Whereas, D-nDCG does not have this property, so it is combined with I-rec, a pure diversity metric,

[*2] Text Retrieval Conference: `http://trec.nist.gov/`
[*3] NII Testbeds and Community for Information access Research: `http://research.nii.ac.jp/ntcir/`
[*4] `http://research.microsoft.com/INTENT/`
[*5] `http://trec.nist.gov/data/web/11/ndeval.c`
[*6] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`

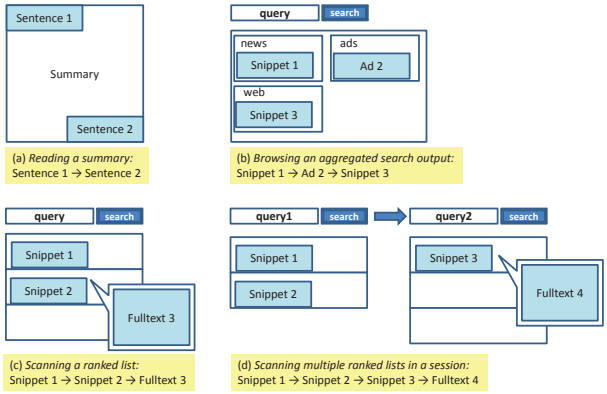

Fig. 1 Constructing trailtexts for various tasks [12].

to compensate for this. However, none of these metrics used at TREC or NTCIR reflects the real user behaviours such as reading snippets and visiting the full text of a relevant document. The new diversity metrics that we advocate in this study, called D-U and U-IA, do just that.

The recently-proposed *Time-Biased Gain* (TBG) evaluation framework [18] is similar to the U-measure framework in that it can also take into account the user's snippet and full text reading behaviours: while U discounts the value of relevant information based on the *amount of text read so far*[*7], TBG does this based on the *time spent so far*. The idea is basically equivalent if the user's reading speed is constant. However, TBG as formulated by Smucker and Clarke [18] relies on the linear traversal assumption, that is, that the user reads the full texts of some of the documents while scanning the ranked list from top to bottom. In contrast, U can handle nonlinear traversals if click information is available [12]. Also, unlike U, TBG as formulated by Smucker and Clarke [18] does not guarantee diminishing return for traditional IR [12]. While TBG is probably more suitable for evaluating non-textual information seeking activities than U is, it has not been extended to diversity evaluation, which is the focus of this study.

3. U-measure, D-U and U-IA

This section defines U-measure and its diversity versions D-U and U-IA as described by Sakai and Dou [12].

First, we present the general U-measure framework. Figure 1 introduces *trailtexts*, the foundation of the U-measure framework. Part (a) shows a single textual query-biased summary being shown to the user. Suppose that we have observed (by means of, say, eyetracking or mousetracking) that the user read only the first and the last sentences of this summary. In this case, we define the trailtext as a simple concatenation of these two sentences: "`Sentence1 Sentence2`." Part (b) shows an aggregated search output: the user reads a snippet in the *news* panel, then reads an *ad*, and finally reads a snippet in the *web* panel. In this case, the trailtext is defined as "`Snippet1 Ad2 Snippet3`." Part (c) is a more traditional search engine result page: the user reads the first two snippets, and then visits the second URL to read the full text. In this case, the trailtext is "`Snippet1 Snippet2 Fulltext3`." Finally, Part (d) shows a session that involves one query reformulation: the user reads two snippets in the original ranked list, reformulates the query, reads one snippet in the new ranked list, and finally visits the actual document. The trailtext is then "`Snippet1 Snippet2 Snippet3 Fulltext4`." Thus, the trailtext is a concatenation of all texts read by the user during her information seeking activity. If evidence from eyetracking/mousetracking etc. is unavailable, the trailtext can alternatively be constructed systematically under a certain user model, using document relevance assessments and or click data [12].

The general U-measure framework comprises two steps:

Step 1 Generate a trailtext, or multiple possible trailtexts, by either observing the actual user or assuming a user model;

Step 2 Evaluate the trailtext(s), based on relevant *information units* (e.g. documents, passages, nuggets) found within it, while discounting the value of each information unit based on its *position* within the trailtext.

Formally, a trailtext tt is a concatenation of n strings: $tt = s_1 s_2 \ldots s_n$. Each string $s_k (1 \leq k \leq n)$ could be a document title, snippet, full text, or even some arbitrary part of a text (e.g. nugget). We assume that the trailtext is exactly what the user actually read, in the exact order, during an information seeking process. We define the offset position of s_k as $pos(s_k) = \sum_{j=1}^{k} |s_j|$. We measure lengths in terms of the number of *characters* [14]. Each s_k in a trailtext tt is considered either x-relevant or nonrelevant. In the present study, we assume that s_k is either a web search engine snippet of 200 characters or a part of a relevant web page; we also assume that the user examines the snippets starting from the top of the list, and that she reads exactly $F = 20\%$ of every relevant web page that she sees[*8]. We define the *position-based gain* as $g(pos(s_k)) = 0$ if s_k is considered nonrelevant, and $g(pos(s_k)) = gv_x$ if it is considered x-relevant, where the gain value setting is the same as those for the rank-based metrics. In the present study where s_k is either a snippet or a part of a full text, $g(pos(s_k)) = gv_x$ if and only if s_k is a part of a full text of an x-relevant document; an example will be discussed later.

The general form of U-measure is given by:

$$U = \frac{1}{\mathcal{N}} \sum_{pos=1}^{|tt|} g(pos) D(pos) \tag{7}$$

where $\mathcal{N}$ is a normalisation factor, which we simply set to $\mathcal{N} = 1$ in this study, pos is an offset position within tt, and $D(pos)$ is a *position-based* decay function. Following the S-measure framework [14], here we assume that the value of a relevant information unit decays linearly with the amount of text the user has read:

$$D(pos) = \max(0, 1 - \frac{pos}{L})) \,. \tag{8}$$

Here, L is the amount of text at which all relevant information units become worthless, which we set to $L = 132000$ based on statistics from 21,802,136 sessions from Bing [12].

The U-measure framework can be extended to handle diversified IR evaluation in two ways. The first is to take the D-measure approach: as shown in Figure 2(a) and (b), given a ranked list, a single trailtext can be built by adding a 200-character snippet for

[*7] The U-measure framework is text-oriented because it is an extension of the *S-measure* framework designed for summarisation evaluation [14].

[*8] The snippet length for Microsoft's Bing is approximately 200 characters on average. Sakai and Dou [12] have discussed the choice of F.

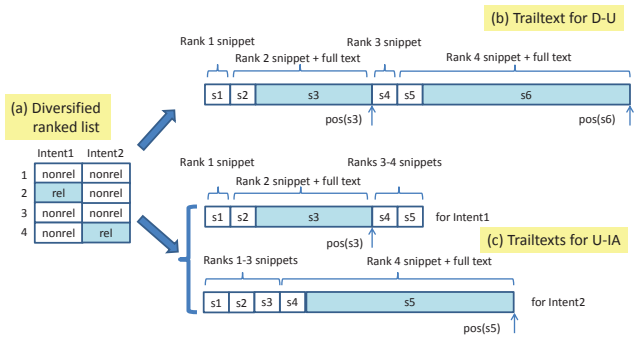

Fig. 2 Constructing trailtexts for D-U and U-IA [12].

each rank and 20% of each document full text which is relevant to at least one intent; then, the *global gain* at the end position of each relevant document is computed as:

$$g(pos(s_k)) = \sum_i P(i|q) g_i(pos(s_k)) \quad (9)$$

where $g_i(pos(s_k)) = gv_x$ if s_k is x-relevant to the i-th intent. This is then plugged in to Eq. 7, to obtain *D-U*. The second approach is to follow the Intent-Aware approach: as shown in Figure 2(a) and (c), a trailtext is built *for each intent*, and a U value (U_i) is computed independently for each i. Finally, the Intent-Aware U is given as:

$$U\text{-}IA = \sum_i Pr(i|q) U_i \; . \quad (10)$$

D-U and U-IA are in fact very similar. Let $\{i'\}(\subseteq \{i\})$ be the set of intents covered by the system output; a document in this output is *strictly locally relevant* if it is relevant to at least one intent from $\{i'\}$ *and* nonrelevant to at least one intent from $\{i'\}$. It is easy to show that *if there is no strictly locally relevant document in the system output, then D-U = U-IA holds*. A corollary is that *if the system output covers only one intent, then D-U = U-IA holds* [12].

4. Concordance Test

Sakai and Dou [12] compared D-U and U-IA with D(♯)-nDCG and a version of ERR-IA in terms of *discriminative power*: the ability of a metric to find statistically significant differences with high confidence for many system pairs. They reported that D-U, U-IA and ERR-IA underperform D(♯)-nDCG in terms of discriminative power, probably because D(♯)-nDCG does not possess the diminishing return property: it does not penalise "redundant" relevant documents, so it relies on more data points and is statistically more stable. However, discriminative power is only a measure of stability: it does not tell us whether the metrics are measuring what they are supposed to measure. To evaluate diversity metrics from this "intuitiveness" point of view, we adopt Sakai's concordance test [11].

Because diversity IR metrics are complex, the concordance test tries to examine how "intuitive" they are by using some very simple "gold-standard" metrics. Since we want both high diversity and high relevance in diversified search, it is possible to regard *intent recall* and/or *precision* (where a document relevant to at least one intent is counted as relevant) as the gold standard. Note that these gold-standard metrics themselves are not good enough for diversity evaluation: these merely represent the basic properties of the more complex diversity metrics that should be satisfied.

Disagreements = 0; $Correct_1$ = 0; $Correct_2$ = 0;
foreach pair of runs (X, Y)
  foreach topic q
    $\Delta M_1 = M_1(q, X) - M_1(q, Y)$;
    $\Delta M_2 = M_2(q, X) - M_2(q, Y)$;
    $\Delta M^* = M^*(q, X) - M^*(q, Y)$;
    if($\Delta M_1 \times \Delta M_2 < 0$) **then** // M_1 and M_2 strictly disagree
      Disagreements + +;
      if($\Delta M_1 \times \Delta M^* \geq 0$)) **then**// M_1 is concordant with M^*
        $Conc_1$ + +;
      if($\Delta M_2 \times \Delta M^* \geq 0$)) **then** // M_2 is concordant with M^*
        $Conc_2$ + +;
    end if
  end foreach
$Conc(M_1|M_2, M^*) = Conc_1/Disagreements$;
$Conc(M_2|M_1, M^*) = Conc_2/Disagreements$;

Fig. 3 Concordance test algorithm for a pair of metrics M_1 and M_2, given the gold-standard metric M^* [11].

Table 1 TREC 2011 Web Track Diversity Task data.

Documents	ClueWeb09 (one billion web pages)
Intent probabilities	Not available
#intents/topic	3.3
#Topics	50
Pool depth	25
#runs	17 Category A runs
Per-intent relevance	graded (0, 1, 2, 3)
#Unique relevant/topic	100.6

Figure 3 shows a simple algorithm for comparing two candidate metrics M_1 and M_2 given a gold standard metric M^*: concordance with multiple gold standards may be computed in a similar way. Here, for example, $M_1(q, X)$ denotes the value of metric M_1 computed for the output of system X obtained in response to topic q. Note that this algorithm focusses on the cases where M_1 and M_2 disagree with each other, and then turn to M^* which serves as the judge. While the condordance test relies on the assumption that the gold-standard metrics represent the real users' preferences[*9], it is useful to be able to quantify exactly how often the metrics satisfy the basic properties that we expect them to satisfy, given many pairs of ranked lists. In our case, the specific questions we address are: (a) How often does a diversity metric agree with intent recall (i.e., prefer the more diversified list)?; (b) How often does it agree with precision (i.e., prefer the more relevant list)?; and (c) How often does it agree with intent recall and precision at the same time?

5. Experiments

Table 1 shows some statistics of the TREC 2011 Diversity Task data which we used for conducting the concordance tests. Note that as we have $17*16/2 = 136$ run pairs, we have $50*136 = 6800$ pairs of ranked lists for the tests. The diversity evaluation metrics, D-U, U-IA, D(♯)-nDCG, α-nDCG and ERR-IA use the measurement depth of $l = 10$ as diversified search mainly concerns the *first* search engine result page. Computation of D-U and U-IA

[*9] Sanderson *et al.* [17] reported on experiments similar to the concordance test where Amazon Mechanical Turkers were used instead of the gold-standard metrics. However, they had to treat each intent of a topic as an independent topic, and hence it is probably difficult for this method to evaluate the ability of a diversity metric to actually reward diversity.

Table 2 Concordance test results with the TREC 2011 Web Track Diversity Task data (50 topics; 17 runs). Statistically significant differences with the sign test are indicated by ‡'s ($\alpha = 0.01$) and †'s ($\alpha = 0.05$).

(a) gold standard: intent recall

	D-nDCG	D-U	U-IA	ERR-IA	α-nDCG
D♯-nDCG	**100%**/0%‡ (415)	**100%**/48%‡ (771)	**100%**/49%‡ (745)	**99%**/61%‡ (1106)	**99%**/71%‡ (913)
D-nDCG	-	81%/84% (562)	79%/**86%**‡ (568)	82%/82% (1044)	75%/**91%**‡ (974)
D-U	-	-	54%/**94%**‡ (54)	83%/81% (1472)	78%/**89%**‡ (1323)
U-IA	-	-	-	**84%**/80%† (1463)	79%/**89%**‡ (1299)
ERR-IA	-	-	-	-	42%/**100%**‡ (292)

(b) gold standard: precision

	D-nDCG	D-U	U-IA	ERR-IA	α-nDCG
D♯-nDCG	48%/**71%**‡ (415)	47%/**76%**‡ (771)	46%/**77%**‡ (745)	**71%**/53%‡ (1106)	**69%**/55%‡ (913)
D-nDCG	-	51%/**74%**‡ (562)	51%/**75%**‡ (568)	**77%**/49%‡ (1044)	**75%**/52%‡ (974)
D-U	-	-	74%/85% (54)	**77%**/48%‡ (1472)	**76%**/50%‡ (1323)
U-IA	-	-	-	**77%**/48%‡ (1463)	**76%**/49%‡ (1299)
ERR-IA	-	-	-	-	48%/**76%**‡ (292)

(c) gold standard: intent recall AND precision

	D-nDCG	D-U	U-IA	ERR-IA	α-nDCG
D♯-nDCG	**48%**/0%‡ (415)	**47%**/38%‡ (771)	**45%**/39%† (745)	**70%**/29%‡ (1106)	**68%**/35%‡ (913)
D-nDCG	-	42%/**65%**‡ (562)	40%/**67%**‡ (568)	**66%**/40%‡ (1044)	**58%**/48%‡ (974)
D-U	-	-	33%/**80%**‡ (54)	**66%**/40%‡ (1472)	**62%**/45%‡ (1323)
U-IA	-	-	-	**67%**/38%‡ (1463)	**63%**/43%‡ (1299)
ERR-IA	-	-	-	-	19%/**76%**‡ (292)

requires the document length statistics for all the relevant documents retrieved above top $l = 10$: the estimated lengths are available at `http://research.microsoft.com/u/`.

As the TREC diversity data lack intent probabilities $Pr(i|q)$, we follow TREC and simply assume that the probability distribution across intents is uniform[*10].

Table 2 summarises our concordance test results. Part (a) shows condordance with intent recall (i.e., the ability to prefer the more diversified result); (b) shows concordance with precision (i.e., the ability to prefer the more relevant result); and (c) shows simultaneous concordance with intent recall and precision. For example, Part (a) contains the following information for the comparison between U-IA and ERR-IA in terms of concordance with intent recall:

- U-IA and ERR-IA disagree with each other for 1,463 out of the 6,800 ranked list pairs;
- Of the above disagreements, U-IA is concordant 84% of the time, while ERR-IA is concordant 80% of the time;
- U-IA is significantly better than ERR-IA according to the sign test ($\alpha = 0.05$)[*11].

Let "$M_1 \gg M_2$" denote the relationship: "M_1 statistically significantly outperforms M_2 in terms of concordance with a given gold-standard metric." Then our results can be summarised as follows[*12]:

(a) Concordance with I-rec (pure diversity): D♯-nDCG $\gg$ α-nDCG $\gg$ U-IA $\gg$ D-U, D-nDCG, ERR-IA;
(b) Concordance with Prec (pure relevance): U-IA, D-U $\gg$ D-nDCG $\gg$ D♯-nDCG $\gg$ α-nDCG $\gg$ ERR-IA;
(c) Simultaneous concordance with I-rec and Prec : D♯-nDCG $\gg$ U-IA $\gg$ D-U $\gg$ D-nDCG $\gg$ α-nDCG $\gg$ ERR-IA.

Recall that D-U and U-IA are more realistic than the other metrics, *including the gold-standard metrics*, in that they consider the snippet and full text reading activities. Thus, we can conclude that D-U and U-IA are not only more realistic than other diversity metrics but also intuitive.

Finally, it can be observed that D-U and U-IA disagree with each other for only 54 out of the 6800 ranked list pairs: thus, in practice, it is not necessary to use both of these metrics at the same time. Based on the above concordance test results, we recommend the use of U-IA. It can also be observed that α-nDCG and ERR-IA also behave similarly: they disagree with eath other for only 292 out of the 6800 ranked list pairs.

[*10] Sakai and Song [16] have discussed the effect of utilising intent probabilities in diversity evaluation.
[*11] Though not shown in the table, U-IA "wins" 273 times while ERR-IA "wins" 250 times.
[*12] In general, note that pairwise statistical significance is not transitive. However, it turns out that our results do not violate transitivity.

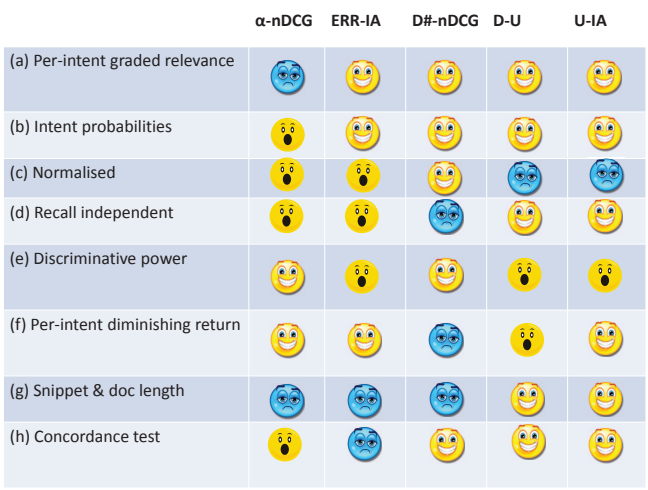

Fig. 4 Comparison of Diversified IR Metrics.

6. Conclusions and Future Work

Our results show that while D♯-nDCG is the overall winner in terms of simultaneous concordance with diversity and relevance, D-U and U-IA statistically significantly outperform other state-of-the-art metrics. Moreover, in terms of concordance with relevance alone, D-U and U-IA significantly outperform all rank-based diversity metrics. These results suggest that D-U and U-IA are not only more realistic than rank-based metrics but also quite intuitive, i.e., that they measure what we want to measure. Moreover, as D-U and U-IA in fact behave extremely similarly, we recommend the use of of U-IA, which outperformed D-U according to the concordance tests.

Figure 4 summarises the various properties of existing diversity evaluation metrics. Below, we provide additional comments for each row in this figure:

(a) As was mentioned in Section 2, α-nDCG lacks a mechanism for directly handling per-intent graded relevance.

(b) The original α-nDCG [5] did not consider $Pr(i|q)$, but it was incorporated later [3].

(c) and (d)These are two sides of the same coin. α-nDCG requires an approximation of an ideal ranked list; there is a version of ERR-IA used at TREC that is normalised in a way similar to α-nDCG [3]. Normalisation generally implies the knowledge of all relevant documents: in this sense, the normalised metrics are recall-dependent. D(♯)-nDCG requires a globally ideal list which also implies the knowledge of all relevant documents. While normalised metrics assume that every topic is of equal importance, unnormalised mercics such as D-U and U-IA assume that every *user effort* is of equal importance: the user needs to spend more effort for topics that have more relevant information.

(e) In terms of discriminative power, D(♯)-nDCG and α-nDCG outperform ERR-IA [15]; D(♯)-nDCG outperform D-U, U-IA and ERR-IA [12][*13].

(f) α-nDCG, ERR-IA and U-IA possess the per-intent diminishing return property; as we have seen, D-U behaves similarly to U-IA, as the original U-measure already has the *per-topic* diminishing return property.

(g) To date, D-U and U-IA are the only diversity metrics that take the user's snippet and full text reading behaviour into account.

(h) This row summarises the findings from the present study.

Our future work for diversity evaluation includes the following:

- Exploring different (possibly nonlinear) decay functions $D(pos)$ with U-IA for different types of search intents (e.g. navigational and informational);
- Comparing different information access styles (e.g. direct answers vs. diversified list of URLs) given ambiguous and/or underspecified queries on the U-measure framework;
- Exploring diversity evaluation methods without using explicit set of intents $\{i\}$, for example, based on relevant *information units* [9], [14] rather than relevant documents.

[*13] These two studies [12], [15] used a version of ERR-IA, which is an "IA version of normalised ERR," not the official ERR-IA from TREC.

References

[1] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G.: Learning to Rank Using Gradient Descent, *Proceedings of ICML 2005*, pp. 89–96 (2005).

[2] Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L. and Wu, S.-L.: Intent-based Diversification of Web Search Results: Metrics and Algorithms, *Information Retrieval*, Vol. 14, No. 6, pp. 572–592 (2011).

[3] Clarke, C. L., Craswell, N., Soboroff, I. and Ashkan, A.: A Comparative Analysis of Cascade Measures for Novelty and Diversity, *Proceedings of ACM WSDM 2011*, pp. 75–84 (2011).

[4] Clarke, C. L., Craswell, N., Soboroff, I. and Voorhees, E.: Overview of the TREC 2011 Web Track, *Proceedings of TREC 2011* (2012).

[5] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S. and MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation, *Proceedings of ACM SIGIR 2008*, pp. 659–666 (2009).

[6] Clarke, C. L., Craswell, N. and Voorhees, E.: Overview of the TREC 2012 Web Track, *Proceedings of TREC 2012* (2013).

[7] Harman, D.: Overview of the Second Text REtrieval Conference (TREC-2), *Proceedings of TREC-2* (1994).

[8] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422–446 (2002).

[9] Kato, M. P., Sakai, T., Yamamoto, T. and Iwata, M.: Report from the NTCIR-10 1CLICK-2 Japanese Subtask: Baselines, Upperbounds and Evaluation Robustness, *Proceedings of ACM SIGIR 2013* (2013).

[10] Pollock, S. M.: Measures for the Comparison of Information Retrieval Systems, *American Documentation*, Vol. 19, No. 4, pp. 387–397 (1968).

[11] Sakai, T.: Evaluation with Informational and Navigational Intents, *Proceedings of WWW 2012*, pp. 499–508 (2012).

[12] Sakai, T. and Dou, Z.: Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation, *Proceedings of ACM SIGIR 2013* (2013).

[13] Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., Kato, M. P., Song, R. and Iwata, M.: Summary of the NTCIR-10 INTENT-2 Task: Subtopic Mining and Search Result Diversification, *Proceedings of ACM SIGIR 2013* (2013).

[14] Sakai, T., Kato, M. P. and Song, Y.-I.: Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access, *Proceedings of ACM CIKM 2011*, pp. 621–630 (2011).

[15] Sakai, T. and Song, R.: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, *Proceedings of ACM SIGIR 2011* (2011).

[16] Sakai, T. and Song, R.: Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT Task, *Information Retrieval* (2013).

[17] Sanderson, M., Paramita, M. L., Clough, P. and Kanoulas, E.: Do user preferences and evaluation measures line up?, *Proceedings of ACM SIGIR 2010*, pp. 555–562 (2010).

[18] Smucker, M. D. and Clarke, C. L. A.: Time-Based Calibration of Effectiveness Measures, *Proceedings of ACM SIGIR 2012*, pp. 95–104 (2012).

[19] Tsukuda, K., Dou, Z. and Sakai, T.: Microsoft Research Asia at the NTCIR-10 Intent Task, *Proceedings of NTCIR-10* (2013).