Which Diversity Evaluation Measures Are "Good"?

Tetsuya Sakai
Waseda University, Tokyo, Japan
tetsuyasakai@acm.org

Zhaohao Zeng
Waseda University, Tokyo, Japan
zhaohao@fuji.waseda.jp

ABSTRACT

This study evaluates 30 IR evaluation measures or their instances, of which nine are for adhoc IR and 21 are for diversified IR, primarily from the viewpoint of whether their preferences of one SERP (search engine result page) over another actually align with users' preferences. The gold preferences were contructed by hiring 15 assessors, who independently examined 1,127 SERP pairs and made preference assessments. Two sets of preference assessments were obtained: one based on a relevance question "Which SERP is more relevant?" and the other based on a diversity question "Which SERP is likely to satisfy a higher number of users?" To our knowledge, our study is the first to have collected diversity preference assessments in this way and evaluated diversity measures successfully. Our main results are that (a) Popular adhoc IR measures such as nDCG actually align quite well with the gold relevance preferences; and that (b) While the D♯-measures align well with the gold diversity preferences, intent-aware measures perform relatively poorly. Moreover, as by-products of our analysis of existing evaluation measures, we define new adhoc measures called iRBU (intentwise Rank-Biased Utility) and EBR (Expected Blended Ratio); we demonstrate that an instance of iRBU performs as well as nDCG when compared to the gold relevance preferences. On the other hand, the original RBU, a recently-proposed diversity measure, underperforms the best D♯-measures when compared to the gold diversity preferences.

KEYWORDS

evaluation measures; search result diversification; user preferences

ACM Reference Format:
Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331215

1 INTRODUCTION

Test-collection-based IR experiments are generally inexpensive compared to user-based ones. Hence many IR researchers consider them to be valuable for developing and optimising effective IR algorithms. However, while the goal of IR is to satisfy the user's information need, such laboratory IR experiments require "offline" evaluation measures as surrogate objective functions: those that can be computed from a canned set of relevance assessments and a SERP. For traditional *adhoc* IR, the relevance assessments are given for each topic; for the more recent *diversified* IR tasks, they are given for each pre-identified *intent* [31] (or *subtopic* [38]) for a topic. Since we routinely rely on these offline measures instead of actually asking users, we need to make sure that the measures actually align well with the user's perception. This is perhaps one of the reasons why new IR measures keep getting invented: we feel we need to fill the gap.

The present study primarily concerns offline diversity IR evaluation measures that can utilise intentwise graded relevance assessments and the intent probabilities for each topic in the diversity test collection. As diversity IR measures are generally more complex than adhoc IR measures, the question of whether they align well with user perception is particularly important. The present study was directly motivated by two existing studies: the recent proposal of a diversity measure called *Rank Biased Utility* (RBU) by Amigó, Spina, and Carrillo de Albornoz [6], and the work of Sanderson *et al.* [32] who tried to examine whether adhoc and diversity measures agree with crowd workers' SERP preferences. The approach of Amigó *et al.* was to first define a set of axioms that IR measures should satisfy and then design a measure that satisfy all of them; natural questions that arise are: Is each axiom *necessary* and does it matter to the user? Are the set of axioms *sufficient* for good alignment with user perception? On the other hand, while the crowdsourced approach of Sanderson *et al.* seems like a straighforward one to address the above questions, their gold data do not, in our opinion, tell us *which SERP is more diversified* (See Section 2.2). Hence we revisit the question of whether diversity measures align with users' SERP preferences for a suite of modern diversity measures, including the recently-proposed RBU.

This study evaluates 30 IR evaluation measures or their instances, of which 21 are for diversified IR, primarily from the viewpoint of whether their preferences of one SERP over another actually align with users' preferences. The gold preferences were contructed by hiring 15 assessors, who independently examined 1,127 SERP pairs and made preference assessments. Two sets of preference assessments were obtained: one based on a relevance question "Which SERP is more relevant?" and the other based on a diversity question "Which SERP is likely to satisfy a higher number of users?" To our knowledge, our study is the first to have collected diversity preference assessments in this way and evaluated diversity measures successfully. Our main results are that (a) Popular adhoc IR measures such as nDCG [18] actually align quite well with the gold relevance preferences; and that (b) While the *D♯-measures* [30] align well with the gold diversity preferences, *intent-aware measures* [1, 10] perform relatively poorly. Moreover, as by-products of our analysis of existing evaluation measures, we define new adhoc measures called iRBU (intentwise RBU) and EBR (Expected Blended Ratio); we demonstrate that an instance of iRBU performs as well

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331215

as nDCG when compared to the gold relevance preferences. On the other hand, the original RBU, a diversity measure, underperforms the best D♯-measures when compared to the gold diversity preferences.

2 RELATED WORK

2.1 Axiomatic Approaches

Amigó *et al.* [6] proposed RBU by first proposing a set of axioms and then deliberately designing a measure that inherits properties from different existing measures (including the intent-aware versions of *Expected Reciprocal Rank* (ERR) [10, 11] and *Rank-Biased Precision* (RBP) [23]) to satisfy all of the axioms. Their axioms are composed of five relevance-oriented constraints from prior art [5] plus five new diversity-oriented constraints. For example, their *redundancy* constraint dictates that *if a SERP examined so far contains more documents relevant to intent i than those relevant to intent i', and if there is a new document d relevant to i and another new document d' relevant to i', then it is better to append d' to the SERP than to append d*. However, this constraint relies on three prerequisites: (1) intentwise relevance assessments are binary; (2) intent probabilities are all equal; and (3) a document is never relevant to multiple intents. (Albahem *et al.* [4] make the same assumptions.) Hence the aforementioned research questions: how important are these axioms in practice? Do measures that satisfy many axioms actually align well with user perceptions?

There is no doubt that axiomatic approaches to designing evaluation measures are useful for understanding the inherent properties of the measures, and Amigó *et al.* [5, 6] are not alone in taking this path. For example, Moffat *et al.* [22] discussed five "requirements for a user model" to propose a measure called *INST* which takes into account how many relevant documents the user expects to find for a particular topic. However, they remark: "*We acknowledge that these desiderata present an idealized situation, and that "all other factors being equal" is a caveat that might be difficult to satisfy, let alone study and measure.*" Our question is: how do IR evaluation measures behave in *practical* situations?

2.2 Agreement with Users/Participants

One approach to examining the gap between offline IR measures and user experience/performance is to give search tasks to hired participants. Thus, correlations between the evaluation measure scores of the SERPs and how well the participants did can be measured. Mixed results have been reported in this line of research. Turpin and Scholer [34] found "*little relationship between performance of systems as measured by MAP and the performance of users*" as measured by the time taken to find the first relevant document and the number of relevant documents that could be identified in five minutes. The results of Al-Maskari *et al.* [2] were more positive: they found that "*the users saved more relevant documents, took less time to save the first relevant document and hence were more satisfied with systems of higher retrieval effectiveness than those with lower effectiveness.*" Furthermore, it is also possible to obtain explicit feedback about the participants' search experience and directly compare it with relevance-based measures (e.g. [21]). However, we are more interested in whether a measure's *preference* of one SERP over another actually reflects the preference of real users: we want to first examine whether *relative* assessments of SERPs by an evaluation measure agree with those by real users, before addressing the question of whether a measure's *absolute* assessment of a SERP is a good predictor of that of the user. Moreover, our primary concern is diversity measures. Hence, the aforementioned work of Sanderson *et al.* [32], which tried to examine how adhoc and diversity measures align with crowd workers' SERP *preferences*, is the work that is most closely related to our study.

The work of Sanderson *et al.* [32], however, has a few limitations from the viewpoint of our own research questions. Firstly, their experiments utilised the TREC 2009 web track diversity task data, which lacks intent probabilities and intentwise *graded* relevance assessments, even though many modern diversity measures are capable of utilising them. Secondly, only four measures, namely, α-nDCG [12, 14], NRBP [15], intent-aware precision, and *cluster recall* (i.e., *intent recall* described in Section 3.2; a.k.a. *subtopic recall* [38], were considered in their study, and the results were rather inconclusive: the measures were similar in terms of preference agreement with the crowd workers, and the best measure was the simplest intent recall, which agreed with the workers for 55 SERP pairs and disagreed with them for 21 pairs. We wanted to obtain a better insight into a wider variety of modern diversity measures, including RBU from SIGIR 2018 [6]. Thirdly, and most importantly, the gold data of Sanderson *et al.* [32] do not, in our opinion, tell us *which SERP is more diversified.* They showed a SERP pair together with a query and a particular *subtopic* (i.e., intent) to each worker, and asked them to select a better SERP in relation to that particular subtopic. Then, these preferences were aggregated via majority vote for the entire topic. Given a SERP pair (A, B), suppose that some crowd workers preferred A for Intent 1, and that some preferred B for Intent 2. Taking a majority vote across these intents will give us a preference for the entire topic, but that preference does not reflect whether the preferred SERP covers more diverse intents than the other. Another concern is that real users do not have a subtopic explicitly shown on their screen; all they have is the query.

Sanderson *et al.* actually report that their initial attempt at using the topic title (i.e., query) alone was not successful; we follow up on this approach. Since a major objective of a diversified SERP is to satisfy as many users' needs as possbile, one ideal approach would be to collect, for each query, a different group of users who issued that same query, and ask each of them to conduct a preference judgement from her personal point of view. However, as we do not find this practical, we take a new approach (Section 6.1). We show that our gold SERP preferences for both relevance and diversity are are reliable *and* useful for comparing evaluation measures.

2.3 Advanced User Models

The diversity evaluation measures considered in this study utilise adhoc IR evaluation measures as their components. Many of the component measures belong to the *Normalised Cumulative Utility* (NCU) family of Sakai and Robertson [29]. An NCU measure assumes a user population, and that different users stop and leave the ranked list at different ranks, and that for each group of users stopping at a particular rank, a utility score of the ranked list can be computed. Subsequently, similar formulations have been discussed by Yang and Lad [37], Zhang, Park, and Moffat [40], Carterette [8], Chandar and Carterette [9] Zhang *et al.* [39] amongst others.

There are user models for evaluation measures that are more complex than NCU. For example, the Bejewled Player Model [39] models users who stop examining the SERP not only due to satisfaction but also frustration; Azzopardi, Thomas, and Craswell [7] proposed a related measure, by building on the aforementioned INST measure [22]. While these new measures are outside the scope of the present study, interested researchers can easily evaluate them by utilising our user preference data (Section 7).

3 MEASURES CONSIDERED IN OUR STUDY

Our primary objective is to investiage how diversity IR measures align with users' SERP preferences. We first consider six *adhoc* ranked retrieval measures that can handle graded relevance, including two new measures: *Expected Blended Ratio* (EBR) and *intentwise Rank Biased Utility* (iRBU). The adhoc measures are then used as components for defining sixteen *diversity* IR measures. Our comparison of adhoc IR measures also include *Precision* as it is the most crude relevance measure, while our comparison of diversity IR measures also include *Intent Recall*, (i.e., the number of intents covered by the SERP) as it is the most crude diversity measure. All measures are based on the document cutoff of 10, as we are interested in diversifying the first page of the web search results.

3.1 Adhoc Measures

As adhoc IR measures, we primarily consider a family of measures known as *Normalised Cumulative Utility* (NCU) [29], defined as:

$$NCU = \sum_{r=1}^{\infty} P_S(r)NU(r)\ , \tag{1}$$

where $P_S(r)$ is the *stopping probability* at rank r and $NU(r)$ is the *normalised utility* at rank r. Given a population of users, it is assumed that $100P_S(r)\%$ of those users will stop and abandon the ranked list at rank r due to satisfaction; the utility of the ranked list for this particular group of users is given by $NU(r)$. Hence, NCU is the expectation of the normalised utility over a population of users. Let $I(r)$ be a flag such that $I(r) = 1$ iff the document at rank r is relevant (at least to some degree). Sakai and Robertson [29] discuss three instances of $NU(r)$, namely, *reciprocal rank*: $RR(r) = 1/r$ (which is based on the view that only the document at rank r was of use to the user), *precision*: $Prec(r) = C(r)/r$ where $C(r) = \sum_{k=1}^{r} I(k)$ (which is based on the view that all relevant documents seen in the top r were of some use to the user), and the *blended ratio*:

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

Here, $\beta(\geq 0)$ is a *patience parameter* [29] akin to a feature in the original definition of nDCG [18] (we let $\beta = 1$ throughout this study); $cg(r)$ is the *cumulative gain* [18] at rank r in the system output, and $cg^*(r)$ is that for the ideal output. In short, the blended ratio is a graded relevance version of precision.

Q-measure [24, 26] is an instance of NCU that uses $P_S(r) = P_{UNIFORM}(r) = I(r)/R$ as the stopping probability, which means that the users are equally likely to stop at any of the relevant documents in the (infinite) ranked list. Also, it uses the blended ratio as the normalised utility to handle graded relevance.[1] In practice, we use a cutoff-based Q-measure, to ensure the [0, 1] range:

$$Q@l = \frac{1}{\min(l, R)} \sum_{r=1}^{l} I(r)BR(r)\ . \tag{3}$$

Chapelle *et al.* [11] point out that their *Expected Reciprocal Rank* (ERR) measure is also an instance of NCU. For ERR, the stopping probability is given by:

$$P_S(r) = P_{ERR}(r) = P_{sat}(r)dsat(r-1) = P_{sat}(r)\prod_{k=1}^{r-1}(1-P_{sat}(k))\ , \tag{4}$$

where $P_{sat}(r)$ is the probability that the users will be satisfied with the document at rank r and abandon the ranked list; if the relevance level of the document at r is L and the maximum relevance level is L_{max}, $P_{sat}(r)$ is assumed to be $(2^L - 1)/2^{L_{max}}$ [11]. As for the normalised utility, ERR uses $RR(r) = 1/r$. Unlike other measures, ERR's stopping probability incorporates *diminishing return* [11, 26], which means that a relevant document found near the top of the ranked list will reduce the value of relevant ones found below. For this reason, here we consider a new hybrid of Q and ERR as yet another instance of NCU, which we call *Expected Blended Ratio*:

$$EBR@l = \sum_{r=1}^{l} P_{ERR}(r)BR(r)\ . \tag{5}$$

Thus, EBR utilises graded relevance for both stopping probability and utility functions. Whereas ERR assumes that only the document at r was relevant to those who abandoned the list at r, EBR assumes that all of the relevant documents seen were of some use.

Rank Biased Precision (RBP) is another popular measure which may also be seen as an instance of NCU. It uses $P_S(r) = P_{RBP} = (1-p)p^{r-1}$ as the stopping probability for some constant $p(< 1)$, and $NU(r) = g(r)/g_{max}$ as the normalised utility; $g(r)$ is the gain value of the document at rank r and g_{max} is the maximum gain value. In contrast to ERR, RBP assumes that the users will proceed from rank r to rank $(r + 1)$ with a constant probability p regardless of the relevance of the document at r.

We also introduce here a measure which we call *intentwise RBU* (iRBU), which is actually a component of RBU [6], a diversity measure. iRBU is an NCU measure where the stopping probability is given by $P_S(r) = P_{ERR}(r)$ and the utility function is given by $NU(r) = p^r$ for some constant[2] p. That is,

$$iRBU@l = \sum_{r=1}^{l} P_{ERR}(r)p^r\ . \tag{6}$$

Note that $NU(r) = p^r$ ignores document relevance: the rank is all that matters; it is more like "inverse user effort" than utility.

For both RBP and iRBU, we let $p = 0.99$ by default, as this achieved the highest *unanimity* (Section 5.2) for RBU in the experiments reported by Amigó *el al.* [6]; we also consider $p = 0.85$ as this was the choice by Moffat *et al.* [22] in their recent work in

[1] Note that letting $\beta = 0$ in Eq. 2 reduces Q-measure to Average Precision.

[2] Note that we use p for both RBP and RBU since RBU inherits the idea of p from RBP. However, while RBP uses p to define a stopping probability distribution, iRBU uses it to define the utility at rank r.

which measures such as nDCG, Q, RBP, and their new measures were compared.[3]

In addition to the above five adhoc NCU measures, we also consider nDCG (the most widely used "Microsoft" version [26]), and Precision at $l = 10$ (Prec). The latter can be interpreted as yet another instance of NCU, where *all* users are assumed to stop and abandon the ranked list at l, with $NU(r) = Prec(r)$.

3.2 Diversity Measures

There are two major approaches to extending adhoc IR measures for evaluating diversified search.[4] The first is the *intent-aware* (IA) approach [1] which combines intentwise evaluation measures; the second is the *D♯-measure* approach [30] which combines intentwise graded relevance values. Hereafter, we assume that each topic t is associated with a set of intents $\{i\}$, where the intent probability given t is estimated to be $Pr(i|t)$ and the intentwise relevance level of a document d for Intent i is known to be $g_i(d)$.

In the D♯-measure approach, the *global gain* (i.e., overall relevance) of a document d for a topic t is computed as $G(d) = \sum_i Pr(i|t)g_i(d)$. Based on this, the ideal list for the entire *topic* can be defined, so that measures such as nDCG can be computed; the resultant measures are called *D-measures*. Furthermore, to evaluate diversified SERPs, diversity can be emphasised as follows:

$$D\sharp\text{-}M@l = \gamma I\text{-}rec@l + (1-\gamma)M@l \ , \tag{7}$$

where *I-rec@l* is the *intent recall* at cutoff l; M is any D-measure, e.g., D-nDCG; γ is a parameter for balancing relevance and diversity (set to $\gamma = 0.5$ as in previous work throughout this study). Several extensions to the D♯ approach exist [16, 17, 35, 36].

In the IA approach, an adhoc evaluation measure M_i is computed for each intent i, based on intentwise relevance levels. If the measure requires an ideal list, it must be constructed *for each intent*. Finally, the intentwise scores are combined as follows:

$$M\text{-}IA = \sum_i Pr(i|t)M_i \ . \tag{8}$$

In our comparison of diversity measures, we consider the aforementioned nDCG, Q, ERR, EBR, RBP as the component measure M for constructing D-measures, D♯-measures, and IA measures.

Finally, we also consider the recently-proposed RBU:

$$RBU@l = iRBU\text{-}IA@l - e\sum_{r=1}^{l} p^r = \sum_i Pr(i|t)iRBU_i@l - e\sum_{r=1}^{l} p^r \ , \tag{9}$$

where e is a small positive constant for estimating the user effort. We let $e = 0.01$ throughout this study [6]: while it is not clear how e should be determined in practice, our initial setting of $e = 0.1$ resulted in many negative RBU scores for our data, which we chose to avoid. As before, we let $p = 0.99$ by default, and also consider $p = 0.85$.

4 NTCIR-9 INTENT-1 DATA

This study utilises the Japanese subtask data from the NTCIR-9 INTENT-1 task [31, 33], which was similar to the TREC web track diversity tasks [13]. The data comprises the Japanese portion of the clueweb09 collection, with 100 topics and 10.91 intents per topic on average, and 15 runs submitted to the task. The main reasons we chose this data set are: (a) Unlike the TREC diversity test collections, the INTENT test collection has the intent probabilities collected based on assessor voting; (b) As the topics and the documents are in Japanese, it was convenient for us to hire assessors in our university; (c) The properties of this data set are well-documented [31, 33]. The INTENT-1 data comes with 5-point intentwise relevance levels; we use an exponential gain value setting for computing the diversity measures: each L-relevant document (for a particular intent) is given a gain value of $2^L - 1$, following the approach of ERR (Section 3.1).

In addition to computing the diversity measures by utilising the INTENT-1 data "as is," we also computed the aforementioned adhoc measures by converting the intentwise relevance assessments into artificial topicwise relevance assessments in order to study their characteristics as component measures. The "pseudo" adhoc qrel file was created as follows. For each document, let S be the sum of its intentwise relevance levels. Its topicwise relevance level is then obtained as $\log_2(S + 1)$, truncated to an integer. For example, if a document is 4-relevant to one intent and 3-relevant to another, then $S = 7$ and therefore the topicwise relevance level is 3. As a result, the highest relevance level in this pseudo adhoc qrels was also 4. Again, exponential gain value setting was used for computing the adhoc measures. It should be noted that our adhoc measures rely on the pseudo adhoc qrels thus created, and also that they are actually similar to the D-measures: in contrast the global gain formula (Section 3.2), these measures disregard the intent probabilities, and sum up the raw relevance levels rather than the exponential gain values (before taking the log).

5 OFFLINE COMPARISON OF MEASURES

Before discussing our user preference-based results in Section 6, here we present "offline" evaluation of the aforementioned measures as in prior art, mainly to illustrate the point that these approaches do not provide the answer to our question: "Which measures are *good*?" *Rank correlation* [26] compares two system rankings produced by two different measures; it merely quantifies which measures are similar to each other. *Unanimity* [6] quantifies how a measure agrees with the SERP preferences according to all other measures (See also Albahem *et al.* [3] for a recent study on offline comparison of measures involving unanimity and RBU.) *Discriminative power* [25, 26] compares the statistical stability of measures by means of obtaining the p-value for every system pair.

5.1 Rank Correlation

Table 1 compares the system rankings according to different measures in terms of Kendall's τ [26]. Each point estimate is accompanied by a 95% '*Fisher's CI*' as described in Long and Cliff [20].[5] We can observe that:

[3] It is known that RBP with a small p (e.g., $p = 0.5$) hurts discriminative power substantially due to its excessive "top-heaviness" [28].

[4] α-nDCG represents an early approach to evaluating diversified search, but we do not discuss it as it cannot handle intentwise graded relevance [26]. The original α-nDCG [14] could not handle intent probabilities either, but this was addressed in a later formulation [12].

[5] Not all pairs of measures are discussed here due to lack of space, but the missing values can easily be obtained from our data (Section 7).

Table 1: Kendall's τ with 95%CIs for ranking 15 systems with different measures (with $p = 0.99$ for RBP, D-RBP, RBP-IA, and (i)RBU)] (a) Adhoc measures; (b) Components of D♯-measures; (c) IA measures and RBU.

(a)	ERR	nDCG	Prec	Q	RBP	iRBU
EBR	.886 [0.767, 0.946]	.924 [0.841, 0.964]	.493 [0.148, 0.731]	.957 [0.908, 0.980]	.914 [0.821, 0.960]	.886 [0.767, 0.946]
ERR	-	.848 [0.695, 0.927]	.414 [0.050, 0.681]	.880 [0.755, 0.943]	.837 [0.675, 0.922]	.848 [0.695, 0.927]
nDCG	-	-	.572 [0.254, 0.778]	.919 [0.831, 0.962]	.894 [0.782, 0.950]	.962 [0.919, 0.982]
Prec	-	-	-	.545 [0.217, 0.762]	.567 [0.247, 0.775]	.552 [0.227, 0.767]
Q	-	-	-	-	.928 [0.849, 0.966]	.880 [0.755, 0.943]
RBP	-	-	-	-	-	.856 [0.710, 0.931]

(b)	D-ERR	D-nDCG	D-Q	D-RBP	I-rec
D-EBR	.861 [0.719, 0.934]	.848 [0.695, 0.927]	.829 [0.661, 0.918]	.817 [0.639, 0.912]	.752 [0.528, 0.878]
D-ERR	-	.785 [0.583, 0.895]	.804 [0.616, 0.905]	.773 [0.563, 0.889]	.689 [0.426, 0.845]
D-nDCG	-	-	.943 [0.880, 0.973]	.971 [0.938, 0.987]	.638 [0.349, 0.816]
D-Q	-	-	-	.952 [0.898, 0.978]	.657 [0.377, 0.827]
D-RBP	-	-	-	-	.625 [0.330, 0.809]

(c)	ERR-IA	nDCG-IA	Q-IA	RBP-IA	RBU
EBR-IA	.981 [0.959, 0.991]	.900 [0.794, 0.953]	.733 [0.496, 0.868]	.748 [0.521, 0.876]	.924 [0.841, 0.964]
ERR-IA	-	.880 [0.755, 0.943]	.714 [0.466, 0.858]	.748 [0.521, 0.876]	.943 [0.880, 0.973]
nDCG-IA	-	-	.842 [0.684, 0.924]	.859 [0.716, 0.933]	.823 [0.650, 0.915]
Q-IA	-	-	-	.961 [0.917, 0.982]	.657 [0.377, 0.827]
RBP-IA	-	-	-	-	.689 [0.426, 0.845]

Table 2: Unanimity scores based on all 30 measures. Adhoc measures are indicated in bold.

Measure	Unanimity	Measure	Unanimity
RBP ($p = 0.99$)	.0378	RBU ($p = 0.85$)	.0185
RBP-IA ($p = 0.99$)	.0291	ERR-IA	.0185
D♯-RBP	.0289	D♯-nDCG	.0182
D-RBP ($p = 0.99$)	.0282	D-nDCG	0181
iRBU ($p = 0.85$)	.0242	D-RBP ($p = 0.85$)	.0181
EBR	.0242	D♯-RBP ($p = 0.85$)	.0180
iRBU ($p = 0.99$)	.0240	D♯-Q	.0178
nDCG	.0239	D-Q	.0178
RBP ($p = 0.85$)	.0238	EBR-IA	.0153
Q	.0238	Q-IA	.0090
D♯-ERR	.0194	RBU ($p = 0.99$)	.0057
RBP-IA ($p = 0.85$)	.0193	D-ERR	.0053
D♯-EBR	.0189	**ERR**	−.0192
D-EBR	.0186	I-rec	−.2429
nDCG-IA	.0185	**Prec**	−.6763

- From Part (a), among the adhoc measures, Prec (the only set-retrieval, binary-relevance measure) ranks systems substantially differently compared to the other measures (τ: .414-.572). ERR (a measure suitable for navigational intents unlike the others) is also slightly different from the other measures (τ: .414-.886). The rankings by the other adhoc measures are very similar.
- From Part (b), among the D♯-measure components, I-rec (the only set-retrieval, pure diversity measure) ranks systems substantially differently compared to the other measures (τ: .625-.752). D-ERR is also slightly different from the other measures (τ: .689-.861).
- From Part (c), among the IA measures and RBU (an IA measure that considers effort (Eq. 9)), the rankings by RBU, EBR-IA, and ERR-IA are very similar (τ: .924-.981). This was as expected since all three measures rely on the stopping probability $P_{ERR}(r)$ (Section 3.1) for each intent. Also, Q-IA and RBP-IA yield remarkably similar rankings despite their different principles in the component measures (τ: .961).

5.2 Unanimity

We implemented the unanimity method of Amigó *et al.* [6] as follows. First, from the INTENT-1 data, we constructed $N = 10,500$ *topic-SERP1-SERP2* triplets (100 topics times 105 system pairs), where each SERP contains the top 10 documents from a run. Let $\mathcal{M}$ be a set of measures. Let $U(\mathcal{M})$ be a set where each element represents one of the above triplets tagged with a *unanimous decision* $ud \in \{GT, LT, EQ\}$; GT means that all of the measures in $\mathcal{M}$ agree that SERP1 outperforms SERP2; LT means that all of them agree that SERP1 underperforms SERP2; EQ means that all of them agree that the two are equally effective. Let $size[U(\mathcal{M})]$ denote the size of $U(\mathcal{M})$, where each element tagged with a GT or LT contributes 1 while each element tagged with an EQ contributes 0.5 to the size [6]. The unanimity of a measure M wrt a set of measures $\mathcal{M}$ (where $M \in \mathcal{M}$) is given by:

$$\log_2 \frac{size[U(\{M\}) \cap U(\mathcal{M} - \{M\})]/N}{(size[U(\{M\})]/N) * (size[U(\mathcal{M} - \{M\})]/N)} . \qquad (10)$$

A high-unanimity measure is one that agrees with many other measures as to which SERP is better.

Table 2 shows the unanimity scores when the universal set $\mathcal{M}$ contains all 30 measures, where two settings for the parameter p are considered for all RBP-based and RBU-based measures. It can be observed that RBP-based measures achieve high unanimity

Table 3: Discriminative power at the 5% significance level for 105 run pairs. Adhoc measures are indicated in bold.

Measure	Significantly diff. pairs	Minimum Δ
D♯-RBP ($p = 0.85$)	50/105= 47.6%	.05
D♯-EBR	48/105= 45.7%	.07
D♯-nDCG	47/105= 44.8%	.06
D♯-ERR	46/105= 43.8%	.07
I-rec	46/105= 43.8%	.08
RBU ($p = 0.85$)	44/105= 41.9%	.05
RBU ($p = 0.99$)	44/105= 41.9%	.07
D♯-RBP ($p = 0.99$)	43/105= 41.0%	.04
nDCG-IA	42/105= 40.0%	.04
RBP ($p = 0.99$)	42/105= 40.0%	.01
nDCG	40/105= 38.1%	.08
iRBU ($p = 0.85$)	40/105= 38.1%	.05
iRBU ($p = 0.99$)	40/105= 38.1%	.06
RBP-IA ($p = 0.85$)	39/105= 37.1%	.02
RBP-IA ($p = 0.99$)	39/105= 37.1%	.01
D♯-Q	39/105= 37.1%	.06
D-RBP ($p = 0.85$)	39/105= 37.1%	.05
RBP ($p = 0.85$)	39/105= 37.1%	.03
Q	39/105= 37.1%	.07
D-RBP ($p = 0.99$)	38/105= 36.2%	.01
Q-IA	35/105= 33.3%	.03
D-nDCG	35/105= 33.3%	.07
EBR	35/105= 33.3%	.09
D-Q	33/105= 31.4%	.07
EBR-IA	30/105= 28.6%	.05
D-EBR	27/105= 25.7%	.09
D-ERR	26/105= 24.8%	.09
ERR	26/105= 24.8%	.07
ERR-IA	24/105= 22.9%	.05
Prec	14/105= 13.3%	.11

scores: that is, they agree often with many other measures. In contrast, ERR (a measure suitable for navigational intents), I-rec (a set retrieval measure which only considers how many intents are covered by the SERP), and Prec (a set retrieval, binary relevance measure) have negative unanimity scores, reflecting the fact that they are very different from the other measures. More importantly, however, it can be observed that the unanimity scores of RBU are rather mediocre: RBU with $p = 0.99$ is near the bottom of the table. This is in sharp contrast to the results of Amigó *et al.* [6] where RBU outperformed many other measures. We also computed unanimity scores by restricting $\mathcal{M}$ to the 21 diversity measures, but the unanimity scores of RBU were still unremarkable. The discrepancy suggests that unanimity scores should be interpreted with caution since they depend heavily (by definition) on the choice of the universal set $\mathcal{M}$.

5.3 Discriminative Power

Table 3 shows our discriminative power results for all 30 measures: we used the Discpower toolkit [26][6] to compute a randomised Tukey HSD test p-value with $B = 10,000$ trials [27] for every system pair ($15 * 14/2 = 105$ pairs). The middle column shows the number of statistically significant pairs obtained at $\alpha = 0.05$; the right column shows the observed minimum delta between two system means that was found to be statistically significant. Regarding adhoc measures, RBP, iRBU (with $p = 0.85, 0.99$ for both measures), nDCG, and Q are similar in terms of discriminative power: 39-42 significant differences at $\alpha = 0.05$ are obtained among the 105 comparisons. They are closely followed by EBR (35 significant differences); ERR is substantially less discriminative due to its focus on the first relevant document seen; Prec substantially underperforms the others.[7]

As for the diversity measures, D♯-measures are generally more discriminative than others. Note, for example, that while the discriminative powers of D-RBP ($p = 0.85$) and I-rec are 37.1% and 43.8%, respectively, D♯-RBP, which is a simple linear combination of these measures, achieves 47.6%. In this respect, the linear combination is successful for every D-measure (D-{RBP, nDCG, Q, EBR, ERR}), at least for this data set. Whereas, the IA measures underperform the corresponding D♯-measures in terms of discriminative power; for example, while the discriminative power of D♯-ERR is 43.8%, that of ERR-IA is only 22.9%. Recall that while D♯-measures combine intent recall and a D-measure (which in turn combines intent probabilities and intentwise graded document relevance), IA measures combine intent probabilities and intentwise *evaluation measure scores*; the former approach seems more promising. As for RBU, it also does well in terms of discriminative power (41.9%).

6 AGREEMENT WITH USERS' SERP PREFERENCES

The offline results reported in Section 5 are, like similar results reported in prior art, rather inconclusive. The rank correlation and unanimity results just show that ERR, Prec, and I-rec behave differently from other measures, and that some measures resemble each other more than others do. Similarly, the discriminative power results show that D♯-measures are more stable than other diversity measures, but not that they are *correct*.

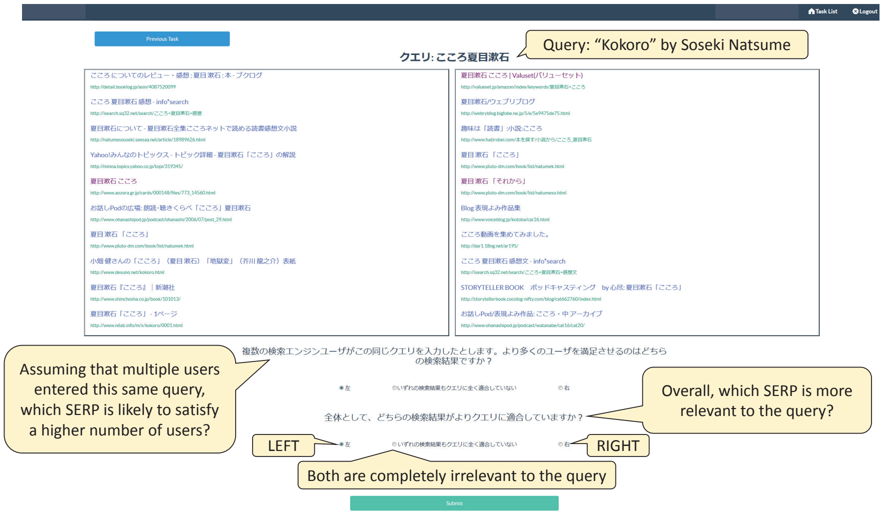
In this section, we evaluate how well each of the aforementioned measures agrees with users' SERP preferences in terms of Kendall's τ: note that τ is computed by subtracting the number of preference disagreements (B) from the number of preference agreements (A), and by normalising by the number of pairs ($P = L(L-1)/2$) where L is the number of ranked items. The 95%CIs are computed as described in Section 5.1, but with L obtained from P backwards as $L = (1 + \sqrt{1 + 8P})/2$, since we are computing the τ directly from pairwise preferences, not from actual ranked lists of size L.

6.1 Gold Data Construction

To construct our SERP preference data, we first obtained a set of topic-SERP-SERP triplets as follows. From the 10,500 triplets described in Section 5.2, we selected those where the performance delta was greater than 0.1 in terms of Prec *and* in terms of I-rec, since we do not want SERP pairs that are very similar in quality. This resulted in 1,258 triplets, but we removed those for Topic 0198 as several runs returned fewer than 10 documents for this topic. Thus we were left with 1,247 triplets. Subsequently, we further filtered out triplets whose SERPs involved web page HTML files without a TITLE field and those with character encoding problems. In the end, 1,127 triplets were left.

[6] http://research.nii.ac.jp/ntcir/tools/discpower-en.html

[7] The minimum Δ is very small for RBP (0.01 with $p = 0.99$ and 0.03 with $p = 0.85$ because RBP with a large p generally takes very small values, due to the $(1 - p)$ factor in P_{RBP} (Section 3.1).



クエリ: こころ夏目漱石

Query: “Kokoro” by Soseki Natsume

こころ についてのレビュー・感想：夏目 漱石：本 - ブクログ
http://detail.booklog.jp/asin/4087520099

こころ 夏目漱石 感想 - info*search
http://isearch.sq32.net/search/こころ+夏目漱石+感想

夏目漱石について - 夏目漱石全集こころネットで読める読書感想文小説
http://natumesouseki.seesaa.net/article/18989626.html

Yahoo!みんなのトピックス - トピック詳細 - 夏目漱石「こころ」の解説
http://minna.topics.yahoo.co.jp/topi/319345/

夏目漱石 こころ
http://www.aozora.gr.jp/cards/000148/files/773_14560.html

お話しPodの広場: 朗読･聴きくらべ「こころ」夏目漱石
http://www.ohanashipod.jp/podcast/ohanashi/2006/07/post_29.html

夏目 漱石「こころ」
http://www.pluto-dm.com/book/list/natumek.html

小畑 健さんの「こころ」（夏目 漱石）「地獄変」（芥川 龍之介）表紙
http://www.desuno.net/kokoro.html

夏目漱石『こころ』｜新潮社
http://www.shinchosha.co.jp/book/101013/

夏目漱石「こころ」- 1ページ
http://www.nilab.info/m/x/kokoro/0001.html

夏目漱石 こころ | Valuset(バリューセット)
http://valueset.jp/amazon/index/keywords/夏目漱石+こころ

夏目漱石/ウェブリブログ
http://webryblog.biglobe.ne.jp/5/e/5e9475de75.html

趣味は「読書」:小説:こころ
http://www.hatirobei.com/本を探す/小説から/こころ_夏目漱石

夏目 漱石「こころ」
http://www.pluto-dm.com/book/list/natumek.html

夏目 漱石「それから」
http://www.pluto-dm.com/book/list/natumeso.html

Blog 表現よみ作品集
http://www.voiceblog.jp/kotoba/car26.html

こころ動画を集めてみました。
http://dar1.18ng.net/ar195/

こころ 夏目漱石 感想文 - info*search
http://isearch.sq32.net/search/こころ+夏目漱石+感想文

STORYTELLER BOOK　ポッドキャスティング　by 心尽: 夏目漱石「こころ」
http://storytellerbook.cocolog-nifty.com/blog/cat662760/index.html

お話しPod/表現よみ作品: こころ・中 アーカイブ
http://www.ohanashipod.jp/podcast/watanabe/cat16/cat20/

複数の検索エンジンユーザがこの同じクエリを入力したとします。より多くのユーザを満足させるのはどちら
の検索結果ですか？

Assuming that multiple users entered this same query, which SERP is likely to satisfy a higher number of users?

● 左 ○いずれの検索結果もクエリに全く適合していない ○ 右

全体として、どちらの検索結果がよりクエリに適合していますか？

Overall, which SERP is more relevant to the query?

● 左 ○いずれの検索結果もクエリに全く適合していない ○ 右

LEFT

RIGHT

Both are completely irrelevant to the query

Figure 1: Preference judgement interface used in our experiments.

To collect gold SERP preferences, we hired 15 Japanese-course computer science students at Waseda University as judges. Each judge used a judgement interface shown in Figure 1 on a desktop computer and independently examined all 1,127 triplets. The presentation order of the triplets was randomised for each judge, although it was possible for them to choose any triplet from the "triplet list" screen (separate from the side-by-side screen) to start or restart the judgements in any order. As Figure 1 shows, the interface shows the *query* (not the *intents*[8]) at the top, and two SERPs side by side; beneath the SERPs, two questions are shown, and the judge is required to answer each of them. The two questions (originally in Japanese) were: the relevance question "*Overall, which SERP is more relevant to the query?*" and the diversity question "*Assuming that multiple users entered this same query, which SERP is likely to satisfy a higher number of users?*" The presentation order of these two questions was also randomised. As for the choice of answers, the judges were required to choose from: "*LEFT is better*," "*RIGHT is better*," and "*Both are completely irrelevant to the query*." Unlike Sanderson *et al.* [32], we avoided the "*Both are equally good*" option and thereby forced the assessors to take a stand.

Recall that, while Sanderson *et al.* [32] asked their crowd workers to imagine that they were searching for a given subtopic with the query, we ask each judge to imagine a group of users who issued the query. While we acknowledge that our approach represents but one way to investigate whether diversity measures align with user preferences, we argue that this tells us much more about evaluation measures than offline approaches do.

Figure 1 shows that our SERP panels provide TITLEs and URLs, but not snippets. The reasons are as follows. Firstly, none of our measures model snippet reading behaviours. If we include snippets in the gold preference contruction step, the τ would reflect not only the the gap between the measures and the user preferences but also the effect of considering snippets (from a particular search engine). Secondly, Sanderson *et al.* [32] mentioned that that "missing snippets did not appear to influence user preferences" at least in their experiments. Our judges were allowed to access the full text of each document by clicking on its TITLE or URL, although our instruction explained that this was not mandatory.

After examining several sample side-by-side instances on the interface, we judged that spending about one minute on the left SERP, one minute on the right one, and then one minute to answer the two questions is a reasonable guideline. Hence the judges were told in advance that they were expected to process one triplet in three minutes on average and that they would be paid on that basis. The *actual* time spent per triplet was 50.1 seconds on average.

6.2 Gold Data Reliability

The procedure described above yielded a matrix with 15 rows and 1,127 columns for each of the two questions described above, where each cell is either 1 (LEFT), 2 (both irrelevant), or 3 (RIGHT). To quantify the reliability of this data set, we employed *Krippendorff's* α [19] by treating the above three answers as three nominal categories; the α was found to be .406 for the relevance question and .356 for the diversity question, both indicating substantial agreement beyond chance. Hence, we believe that we managed to obtain reasonable user preference data not only for relevance but also for diversity. To see if every assessor worked conscientiously, we also computed Krippendorff's α for leave-one-out matrices as well, and the results are shown in Table 4. For example, Column 05 in the Relevance row shows the α after removing the labels given by Judge 05: this is the highest α among the leave-one-out α's for Relevance, which implies that Judge 05 was the one who disagreed with others the most. However, it can be observed that the leave-one-out α's vary only slightly for both Relevance and Diversity, and hence that all of our judges are reliable to a reasonable degree.

To check the relationship between our Relevance and Diversity preference assessments, we computed, for each of our 1,127 triplets, the counts of LEFT *minus* the counts of RIGHT for each of the two questions: clearly, the range of this score is $[-15, 15]$. The Pearson correlation between the scores for Relevance and those for Diversity is .898 (95%CI[.886, .909]): the assessors tended to prefer the same SERP for both relevance and diversity questions, but not always.

From each of the two full matrices, we also created a *high-agreement* matrix by filtering the columns (i.e., triplets) so that

[8] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/Evaluations/INTENT/ntcir9intent.0101-0200.full.xml

Table 4: Krippendorff's α for all 15 judges and leave-one out data (1,127 topic-SERP-SERP triplets).

		Leave One Out														
	All15	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
Relevance	.406	.402	.399	.393	.399	.406	.429	.403	.420	.418	.394	.407	.400	.405	.422	.401
Diversity	.356	.357	.348	.360	.345	.362	.375	.354	.380	.354	.344	.357	.342	.358	.365	.347

Table 5: Agreement with gold relevance preference data. Adhoc measures are indicated in bold.

(a) full gold data ($P = 1,115$ SERP pairs)				(b) high-agreement gold data ($P = 894$ SERP pairs)			
Measure	agree	disagree	τ with 95%CI	Measure	agree	disagree	τ with 95%CI
iRBU ($p = 0.99$)	952	163	.708 [.596, .793]	**nDCG**	804	90	.799 [.710, .863]
nDCG	945	170	.695 [.579, .783]	**iRBU** ($p = 0.99$)	803	91	.796 [.706, .861]
RBP ($p = 0.85$)	935	180	.677 [.556, .770]	D-Q	796	98	.781 [.686, .850]
Q	935	180	.677 [.556, .770]	**RBP** ($p = 0.85$)	792	102	.772 [.674, .844]
D-Q	933	182	.674 [.553, .767]	**RBP** ($p = 0.99$)	791	103	.770 [.671, .842]
RBP ($p = 0.99$)	932	183	.672 [.550, .766]	**Q**	791	103	.770 [.671, .842]
iRBU ($p = 0.85$)	922	193	.654 [.527, .752]	**iRBU** ($p = 0.85$)	787	107	.761 [.659, .836]
EBR	917	198	.645 [.516, .745]	**EBR**	785	109	.756 [.652, .832]
D♯-Q	892	223	.600 [.460, .711]	D♯-nDCG	778	116	.740 [.631, .821]
D♯-nDCG	888	227	.593 [.451, .706]	D♯-Q	778	116	.740 [.631, .821]
Prec	887	228	.591 [.449, .704]	D♯-RBP ($p = 0.85$)	766	128	.714 [.596, .802]
D♯-RBP ($p = 0.85$)	881	234	.580 [.435, .695]	D♯-EBR	760	134	.700 [.578, .791]
D♯-EBR	872	243	.564 [.416, .683]	D-RBP ($p = 0.99$)	759	135	.698 [.575, .790]
D-RBP ($p = 0.99$)	867	248	.555 [.405, .676]	RBU ($p = 0.99$)	752	142	.682 [.554, .778]
D-nDCG	864	251	.550 [.399, .672]	RBP-IA ($p = 0.99$)	752	142	.682 [.554, .778]
RBU ($p = 0.99$)	860	255	.543 [.391, .666]	D-nDCG	751	143	.680 [.552, .777]
RBP-IA ($p = 0.99$)	859	256	.541 [.388, .665]	**Prec**	750	144	.678 [.549, .775]
I-rec	859	256	.541 [.388, .665]	RBU ($p = 0.85$)	746	148	.669 [.538, .769]
D♯-RBP ($p = 0.99$)	859	256	.541 [.388, .665]	I-rec	746	148	.669 [.538, .769]
RBU ($p = 0.85$)	852	263	.528 [.373, .655]	D♯-RBP ($p = 0.99$)	746	148	.669 [.538, .769]
nDCG-IA	851	264	.526 [.370, .653]	D♯-ERR	742	152	.660 [.526, .762]
D♯-ERR	848	267	.521 [.364, .649]	D-RBP ($p = 0.85$)	742	152	.660 [.526, .762]
D-RBP ($p = 0.85$)	846	269	.517 [.359, .646]	RBP-IA ($p = 0.85$)	741	153	.658 [.524, .760]
D-EBR	845	270	.516 [.358, .645]	nDCG-IA	738	156	.651 [.515, .755]
RBP-IA ($p = 0.85$)	845	270	.516 [.358, .645]	D-EBR	738	156	.651 [.515, .755]
Q-IA	844	271	.514 [.356, .643]	Q-IA	735	159	.644 [.506, .750]
ERR	841	274	.509 [.350, .639]	EBR-IA	725	169	.622 [.478, .734]
EBR-IA	833	282	.494 [.332, .627]	**ERR**	718	176	.606 [.458, .721]
ERR-IA	824	291	.478 [.314, .615]	ERR-IA	717	177	.604 [.455, .720]
D-ERR	811	304	.455 [.287, .596]	D-ERR	706	188	.579 [.424, .701]

at least *nine* judges out of 15 agreed as to which was the SERP was better. As a result, the relevance matrix was reduced to 894 columns, while the diversity matrix was reduced to 897 columns. The α rose to .518 for the high-agreement relevance matrix, and to .453 for the high-agreement diversity matrix. Finally, the gold preference data were constructed from the above four matrices by taking a majority vote for each column; for the full relevance and diversity matrices, columns were removed if LEFTs and RIGHTs were tied, resulting in 1,115 triplets and 1,119 triplets, respectively. In other words, there were only 12 and 8 tied triplets in the full relevance and diversity matrices, respectively.

6.3 Main Results

Table 5 summarises the results based on the gold *relevance* preferences. First, as indicated in bold, the top performing measures are adhoc measures, not diversity measures, which is as expected. iRBU with $p = 0.99$ and nDCG perform best: for example, in Table 5(b), nDCG agrees with 804 of the user preferences, while disagreeing for only 90 cases. Q, RBP, and EBR also do well. On the other hand, Prec performs worse than the aforementioned measures (as it is only a set retrieval measure), and ERR performs very poorly (as it relies heavily on the *first* relevant document). This suggests that our judges, when asked about the relevance of the two SERPs, cared about the amount of relevant information in the SERPs *and* the ranks of relevant pages. Moreover, among the diversity measures, note that the D♯-measures align better with the gold *relevance* preferences than the IA measures do.

The fact that iRBU with $p = 0.99$ outperforms many other adhoc measures, especially EBR, was unexpected. Recall that iRBU and EBR differ only in how utility is defined: EBR's utility is based on the relevance of the documents examined so far, but iRBU's utility completely ignores relevance (Section 3.1). This suggests that, to design an adhoc measure that aligns well with user's SERP preferences, it may be a good idea to encode the document relevance information in the stopping probability distribution, but not in the utility function. This topic is left for future work.

Tables 6 summarises the results based on the gold *diversity* preferences. Note that the D♯-measures and its I-rec component are

Table 6: Agreement with gold diversity preference data. Adhoc measures are indicated in bold.

(a) full gold data ($P = 1,119$ SERP pairs)				(b) high-agreement gold data ($P = 897$ SERP pairs)			
Measure	agree	disagree	τ with 95%CI	Measure	agree	disagree	τ with 95%CI
D♯-nDCG	956	163	.709 [.598, .794]	D♯-nDCG	825	72	.839 [.766, .891]
D♯-RBP ($p = 0.85$)	948	171	.694 [.578, .782]	D♯-RBP ($p = 0.85$)	822	75	.833 [.757, .887]
D♯-EBR	943	176	.685 [.567, .776]	I-rec	813	84	.813 [.730, .873]
I-rec	940	179	.680 [.560, .772]	D♯-RBP ($p = 0.99$)	813	84	.813 [.730, .873]
D♯-RBP ($p = 0.99$)	940	179	.680 [.560, .772]	D♯-EBR	813	84	.813 [.730, .873]
D♯-Q	933	186	.668 [.545, .763]	D♯-Q	808	89	.802 [.715, .865]
RBU ($p = 0.99$)	931	188	.664 [.540, .760]	RBU ($p = 0.99$)	807	90	.799 [.710, .863]
D♯-ERR	923	196	.650 [.522, .749]	D♯-ERR	798	99	.779 [.683, .848]
EBR	919	200	.643 [.514, .744]	RBU ($p = 0.85$)	781	116	.741 [.632, .821]
iRBU ($p = 0.85$)	909	210	.625 [.491, .730]	**EBR**	779	118	.737 [.627, .818]
RBU ($p = 0.85$)	907	212	.621 [.486, .727]	nDCG-IA	771	126	.719 [.603, .805]
nDCG-IA	906	213	.619 [.484, .725]	**iRBU** ($p = 0.85$)	767	130	.710 [.591, .799]
iRBU ($p = 0.99$)	890	229	.591 [.449, .704]	D-RBP ($p = 0.99$)	759	138	.692 [.568, .786]
nDCG	889	230	.589 [.447, .702]	EBR-IA	758	139	.690 [.565, .784]
EBR-IA	887	232	.585 [.442, .699]	D-EBR	755	142	.683 [.556, .779]
D-RBP ($p = 0.99$)	886	233	.584 [.440, .698]	**nDCG**	755	142	.683 [.556, .779]
D-EBR	886	233	.584 [.440, .698]	RBP-IA ($p = 0.99$)	752	145	.677 [.548, .774]
Q-IA	881	238	.575 [.429, .691]	**iRBU** ($p = 0.99$)	752	145	.677 [.548, .774]
D-nDCG	881	238	.575 [.429, .691]	Q-IA	748	149	.668 [.537, .768]
RBP ($p = 0.85$)	881	238	.575 [.429, .691]	D-nDCG	746	151	.663 [.530, .764]
RBP-IA ($p = 0.99$)	877	242	.567 [.420, .685]	D-RBP ($p = 0.85$)	745	152	.661 [.528, .763]
D-RBP ($p = 0.85$)	876	243	.566 [.418, .684]	**RBP** ($p = 0.85$)	745	152	.661 [.528, .763]
RBP-IA ($p = 0.85$)	875	244	.564 [.416, .683]	RBP-IA ($p = 0.85$)	744	153	.659 [.525, .761]
ERR-IA	870	249	.555 [.405, .676]	ERR-IA	742	155	.654 [.519, .757]
ERR	866	253	.548 [.397, .670]	D-ERR	726	171	.619 [.474, .731]
RBP ($p = 0.99$)	856	263	.530 [.375, .656]	**RBP** ($p = 0.99$)	725	172	.616 [.470, .729]
D-ERR	855	264	.528 [.373, .654]	**ERR**	723	174	.612 [.465, .726]
D-Q	841	278	.503 [.343, .635]	D-Q	715	182	.594 [.443, .712]
Q	833	286	.489 [.327, .623]	**Q**	703	194	.567 [.410, .692]
Prec	770	349	.376 [.197, .531]	**Prec**	656	241	.463 [.285, .610]

clearly the best measures here. In particular, D♯-nDCG and D♯-RBP with $p = 0.85$ agree most often with both full and high-agreement gold data. In contrast, the IA measures do substantially poorly. For example, in Table 6(b), while D♯-nDCG agrees with the gold diversity preferences for 825 SERP pairs of out 897, the IA counterpart nDCG-IA agrees with the gold data for only 771 pairs. Moreover, recall that D♯-measures generally outperformed the IA measures even with the gold *relevance* preferences, and that the D♯-measures demonstrated higher discriminative power than the IA counterparts (Section 5.3). These results suggest that the D♯-measure approach should be preferred over the IA approach.

According to Amigó *et al.* [6], D♯-measures satisfy fewer of their axioms than RBU. However, Tables 6 shows that, while RBU correlates well with the gold diversity preferences, it underperforms the best D♯-measures. Hence, the answer to our research questions (*Is each axiom necessary and does it matter to the user? Are the set of axioms sufficient for good alignment with user perception?*) appears to be: *Not necessarily.*

As for the adhoc measures, EBR has the highest τ with the gold *diversity* data, while ERR, RBP, Q, and Prec perform poorly.

7 CONCLUSIONS AND FUTURE WORK

We evaluated 30 IR evaluation measures, of which 21 are for diversified IR, primarily from the viewpoint of whether their preferences of one SERP over another actually align with users' preferences. Our main results are that (a) Popular adhoc IR measures such as nDCG actually align quite well with the gold relevance preferences; and that (b) While the D♯-measures align well with the gold diversity preferences, IA measures perform relatively poorly. Moreover, the proposed iRBU (with $p = 0.99$) measure performs as well as nDCG when compared to the gold relevance preferences; on the other hand, the original RBU underperforms the best D♯-measures when compared to the gold diversity preferences. These results are arguably more informative than offline results like the ones presented in Section 5 or in prior art.

The following from our experiments are publicly available[9]: (a) topic-by-run matrices for all 30 evaluation measures; (b) 1,127 topic-SERP-SERP triplets (and the INTENT-1 run names from which the SERPs come from); (c) two 15 × 1,127 user preference matrices. We encourage other researchers to validate their favourite adhoc or diversity measures by utilising the above data along with the NTCIR-9 INTENT-1 test collections and runs[10].

Clearly, the correlations between the SERP preferences of our best measures and those of users are not perfect. Therefore, designing new evaluation measures that outperform these best measures is an important direction for future research. Moreover, devising offline evaluation methods that can accurately predict user-oriented results such as ours would be highly challenging but useful.

[9] http://waseda.box.com/SIGIR2019PACK

[10] http://research.nii.ac.jp/ntcir/data/data-en.html

REFERENCES

[1] Rakesh Agrawal, Gollapudi Sreenivas, Alan Halverson, and Samuel Leong. 2009. Diversifying Search Results. In *Proceedings of ACM WSDM 2009*. 5–14.

[2] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. 2008. The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness. In *Proceedings of ACM SIGIR 2018*. 59–66.

[3] Ameer Albahem, Damiano Spina, Falk Scholer, and Lawrence Cavedon. 2019. Meta-evaluation of Dynamics Search: How Do Metrics Capture Topical Relevance, Diversity, and User Effort?. In *Proceedings of ECIR 2019*. to appear.

[4] Ameer Albahem, Damiano Spina, Falk Scholer, Alistair Moffat, and Lawrence Cavedon. 2018. Desirable Properties for Diversity and Truncated Effectiveness Metrics. In *Proceedings of ADCS 2018*.

[5] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks. In *Proceedings of ACM SIGIR 2013*. 643–652.

[6] Enrique Amigó, Damiano Spina, and Jorge Carrillo de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducting the Rank-Biased Utility Metric. In *Proceedings of ACM SIGIR 2018*. 625–634.

[7] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages. In *Proceedings of ACM SIGIR 2018*. 605–614.

[8] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of ACM SIGIR 2011*. 903–912.

[9] Praveen Chandar and Ben Carterette. 2013. Preference Based Evaluation Measures for Novelty and Diversity. In *Proceedings of ACM SIGIR 2013*. 413–422.

[10] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based Diversification of Web Search Results: Metrics and Algorithms. *Information Retrieval* 14, 6 (2011), 572–592.

[11] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of ACM CIKM 2009*. 621–630.

[12] Charles L.A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proceedings of ACM WSDM 2011*. 75–84.

[13] Charles L.A. Clarke, Nick Craswell, and Ellen M. Voorhees. 2013. Overview of the TREC 2012 Web Track. In *Proceedings of TREC 2012*.

[14] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2009. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of ACM SIGIR 2008*. 659–666.

[15] Charlies L.A. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Proceedings of ICTIR 2009 (LNCS 5766)*. 188–199.

[16] Zhicheng Dou, Xue Yang nad Diya Li, Ji-Rong Wen, and Tetsuya Sakai. 2019. Low-cost, Bottom-up Measures for Evaluating Search Result Diversification. *Information Retrieval Journal*.

[17] Peter B. Golbus, Javed A. Aslam, and Charles L.A. Clarke. 2013. Increasing Evaluation Sensitivity to Diversity. *Information Retrieval* 16, 4 (2013), 530–555.

[18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[19] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology (Fourth Edition)*. SAGE Publications.

[20] Jeffrey D. Long and Norman Cliff. 1997. Confidence Intervals for Kendall's tau. *Brit. J. Math. Statist. Psych.* 50 (1997), 31–41.

[21] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When Does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *Proceedings of ACM SIGIR 2016*. 463–472.

[22] Alistair Moffat, Peter Bailey adn Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM TOIS* 35, 3 (2017).

[23] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS* 27, 1 (2008).

[24] Tetsuya Sakai. 2005. Ranking the NTCIR Systems based on Multigrade Relevance. In *Proceedings of AIRS 2004 (LNCS 3411)*. 251–262.

[25] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of ACM SIGIR 2007*. 71–78.

[26] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*. 116–163.

[27] Tetsuya Sakai. 2018. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer.

[28] Tetsuya Sakai and Noriko Kando. 2008. On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments. *Information Retrieval* 11 (2008), 447–470.

[29] Tetsuya Sakai and Stephen Robertson. 2008. Modelling A User Population for Designing Information Retrieval Metrics. In *Proceedings of EVIA 2008*. 30–41.

[30] Tetsuya Sakai and Ruihua Song. 2011. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In *Proceedings of ACM SIGIR 2011*.

[31] Tetsuya Sakai and Ruihua Song. 2013. Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT Task. *Information Retrieval* 16, 4 (2013), 504–529.

[32] Mark Sanderson, Monica L. Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of ACM SIGIR 2010*. 555–562.

[33] Ruihua Song, Min Zhang, Tetsuya Sakai, Makoto P. Kato, Yiqun Liu, Miho Sugimoto, Qinglei Wang, and Naoki Orii. 2011. Overview of the NTCIR-9 INTENT Task. In *Proceedings of NTCIR-9*. 82–105.

[34] Andrew Turpin and Falk Scholer. 2006. User Performance versus Precision Measures for Simple Search Tasks. In *Proceedings of ACM SIGIR 2006*. 11–18.

[35] Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang. 2018. Search Result Diversity Evaluation Based on Intent Hierarchies. *IEEE TKDE* 30 (2018), 156–169. Issue 1.

[36] Takehiro Yamamoto, Yiqun Liu, Min Zhang, Zhicheng Dou, Ke Zhou, Ilya Markov, Makoto P. Kato, Hiroaki Ohshima, and Sumio Fujita. 2016. Overview of the NTCIR-12 IMine-2 Task. In *Proceedings of NTCIR-12*. 8–26.

[37] Yiming Yang and Abhimanyu Lad. 2009. Modeling Expected Utility of Multi-session Information Distillation. In *Proceedings of ICTIR 2009 (LNCS 5766)*. 164–175.

[38] ChengXiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *Proceedings of ACM SIGIR 2003*. 10–17.

[39] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating Web Search with a Bejeweled Player Model. In *Proceedings of ACM SIGIR 2017*. 425–434.

[40] Yuye Zhang, Laurence A.F. Park, and Alistair Moffat. 2010. Click-based Evidence for Decaying Weight Distributions in Search Effectiveness Metrics. *Information Retrieval* 13 (2010), 46–69.