Modelling A User Population for Designing Information Retrieval Metrics

Tetsuya Sakai† Stephen Robertson*
†NewsWatch, Inc., tetsuyasakai@acm.org
*Microsoft Research Cambridge, ser@microsoft.com

Abstract

Although Average Precision (AP) has been the most widely-used retrieval effectiveness metric since the advent of Text Retrieval Conference (TREC), the general belief among researchers is that it lacks a user model. In light of this, Robertson recently pointed out that AP can be interpreted as a special case of Normalised Cumulative Precision (NCP), computed as an expectation of precision over a population of users who eventually stop at different ranks in a list of retrieved documents. He regards AP as a crude version of NCP, in that the probability distribution of the user's stopping behaviour is uniform across all relevant documents.

In this paper, we generalise NCP further and demonstrate that AP and its graded-relevance version Q-measure are in fact reasonable metrics despite the above uniform probability assumption. From a probabilistic perspective, these metrics emphasise long-tail users who tend to dig deep into the ranked list, and thereby achieve high reliability. We also demonstrate that one of our new metrics, called $NCU_{gu,\beta=1}$, maintains high correlation with AP and *shows the highest* discriminative power, *i.e., the proportion of statistically significantly different system pairs given a confidence level, by utilising graded relevance in a novel way. Our experimental results are consistent across NTCIR and TREC.*

Keywords: *evaluation metrics, average precision, graded relevance, user model, normalised cumulative utility.*

1 Introduction

After the advent of Text Retrieval Conference (TREC), evaluating ranked retrieval systems using test collections with *Average Precision* (AP), or its Mean across topics (MAP), has become the *de facto* standard. In words, the meaning of AP is as follows: Examine a ranked list from the top and, every time you find a relevant document, compute *precision* at this point, i.e., the proportion of relevant documents among the documents seen so far. Take the average of the precision values over all relevant documents: For relevant documents that are not retrieved, let the precision values be zero.

AP has received some criticisms, one of them being that it "lacks a user model." For example, Buckley and Voorhees remark that "there is no single user application that directly motivates MAP" ([2], p. 59). Moffat, Webber and Zobel [11] argue that "there is no plausible search model that corresponds to MAP, because no user knows in advance the number of relevant answers present in the collection they are addressing." However, Robertson has recently pointed out that AP can be interpreted as a special case of *Normalised Cumulative Precision* (NCP), computed as an expectation of precision over a population of users who eventually stop at different ranks in a ranked list of retrieved documents. He regards AP as a crude version of NCP, in that the probability distribution of the user's stopping behaviour is uniform across all relevant documents.

In this paper, we generalise NCP to introduce a family of metrics called *Normalised Cumulative Utility* (NCU), some of which are arguably more "realistic" than AP. First, in addition to the uniform probability distribution of AP, we consider a *rank-biased* distribution that reflects the assumption that users tend to stop at a relevant document near the top of the ranked list rather than one near the bottom, and a *graded-uniform* distribution that reflects the assumption that users tend to stop at a highly relevant document rather than at a partially relevant document. Second, to generalise precision which AP uses as the utility function given the user's stopping point, we use an alternative that can handle graded relevance: the *blended ratio* [14]. Using data from both NTCIR and TREC, we examine the family of NCU metrics in terms of rank correlation and *discriminative power*, i.e., the proportion of statistically significantly different system pairs given a confidence level [15].

Our main conclusion is that AP and its graded-relevance version *Q-measure* [14] are reasonable metrics despite the fact that they rely on a uniform distribution across all relevant documents, as most of our new variants do not demonstrate any perceivable advantages. In particular, using a rank-biased distribution over relevant documents substantially hurts discriminative power, which suggests that it is a good idea to look beyond the stopping point of an ordinary user for obtaining reliable conclusions from experiments. From a probabilistic perspective, AP and Q emphasise long-tail users who tend to dig deep into the ranked list, and thereby achieve high reliability. In addition, we show that one of our new metrics, called $NCU_{gu,\beta=1}$, maintains high correlation with AP *and* shows the highest discriminative power among our metrics, by utilising graded relevance in a novel way.

The remainder of this paper is organised as follows. Section 2 discusses previous work: First, we define Robertson's NCP; second, we discuss related work that examines alternatives to AP; third, we describe existing methods we use for comparing evaluation metrics, namely, Kendall's rank correlation, *Yilmaz/Aslam/Robertson rank correlation* [26] that is arguably more suitable than Kendall's rank correlation for our purpose, and *discriminative power* [15]. Section 3 formally defines our proposed metrics and provide some simple examples. Section 4 describes our experiments using NTCIR and TREC data for comparing our NCU metrics, including AP, in terms of rank correlation and discriminative power. Finally, Section 5 concludes this paper and discusses some possible future work.

2 Previous Work

2.1 Normalised Cumulative Precision

Robertson [12] defined *Normalised Cumulative Precision* (NCP) in order to provide a user model for AP and to generalise it.

Let $I(n)$ be a flag indicating whether the document retrieved at rank n is relevant or not, and let $C(n) = \sum_{i=1}^{n} I(i)$. Clearly, precision at rank n is given by $P(n) = C(n)/n$. Moreover, let $AP_n = I(n)P(n)$. That is, $AP_n = P(n)$ if the document at rank n is relevant, and $AP_n = 0$ otherwise.

Following Cooper in his proposal for the Expected Search Length (ESL) measure [4, 5], let us envisage a user stepping down a ranked list of documents until some stopping point. Unlike ESL, let us assume fully-ranked output with no ties, so that the reason that Cooper introduced an expectation, which was to deal with ties, no longer applies to us. However, let us assume instead that we do *not* know the number of documents the user will examine before he eventually stops. More specifically, let us assume that with probability $p_s(n)$, the user's stopping point is the document at rank n in the list.

Robertson further assumed that the user stops due to *satisfaction*, and that satisfaction can only occur at a relevant document. Thus, according to these assumptions, $p_s(n) = 0$ for every rank n where there is a nonrelevant document. But more generally, the only requirement for $p_s(n)$ is that it must sum to one: The user's stopping behaviour may be due to satisfaction, frustration, a combination of the two, or possibly some other reason (e.g., exhaustion).

The original definition of NCP is as follows:

$$NCP = \sum_{n=1}^{\infty} p_s(n) AP_n \tag{1}$$

NCP is designed to be an expectation of "utility" over a population of users with different stopping behaviours, where "utility" at each given rank is measured by $AP_n = I(n)P(n)$. That is, "utility" at each given stopping point with a relevant document is measured by $P(n)$, which relates to the effort on the user's part in reaching this satisfaction point.

We note that one of the reasons for choosing $P(n)$ as the utility measure is that it does not in itself have any discount based on rank. If we *know* that the user has stopped / will stop at rank n, it does not matter where above rank n any particular good or bad document is located. The resulting expected utility NCP is nevertheless 'top-heavy' in the sense that it takes more account of earlier than of later ranks. This top-heaviness arises entirely from the probabilistic stopping point – if some users are expected to stop earlier than others, then the earlier ranks become more important, simply because they affect more users.

Robertson provided two simple versions of NCP called NCP_u and NCP_1. Let R denote the number of relevant documents for a particular topic. NCP_u employs a uniform probability distribution $p_u(n)$ over all the relevant documents for this topic. That is, $p_s(n) = p_u(n) = 1/R$ for all n s.t. $I(n) = 1$, while $p_s(n) = 0$ for all n s.t. $I(n) = 0$. Let n_1 denote the rank of the first relevant document found in the ranked list/ NCP_1 uses $p_s(n_1) = 1$ and $p_s(n) = 0$ for all $n(= n_1)$. Hence,

$$NCP_u = \frac{1}{R} \sum_{n=1}^{\infty} I(n)P(n) \tag{2}$$

$$NCP_1 = I(n_1)P(n_1) = P(n_1) = 1/n_1 \, . \tag{3}$$

Robertson points out that NCP_u is none other than AP and that NCP_1 is none other than Reciprocal Rank (RR), both of which are used widely in the information retrieval research community. (Note that the above definitions assume that all documents in the document collection are ranked: In practice, however, we approximate them by using *truncated* ranked lists containing, for example, up to 1000 documents.) We also observe that even with the uniformity assumption, the resulting measure (AP) is top-heavy, for precisely the reason given above.

Hereafter, we use a slightly generalised form of NCP:

$$NCP = \sum_{n=1}^{\infty} p_s(n)P(n) \; . \qquad (4)$$

That is, we omit the flag $I(n)$ in order to let the probability distribution $p_s(n)$ handle whether the document at rank n (whether it is relevant or not) should contribute to NCP or not. If we follow Robertson in assuming that the user's stopping point is at a relevant document and never at a nonrelevant document, then we just let $p_s(n) = 0$ for every rank n s.t. $I(n) = 0$. Hence explictly including $I(n)$ in the formula for NCP is not necessary for our purpose. We will also consider a more general utility function than $P(n)$ below.

2.2 Alternatives to AP

Here, we discuss some retrieval effectiveness metrics other than AP, some of which are closely related to the present study.

Popular binary-relevance metrics that are often used alongside with AP include precision at k and *R-precision*, i.e., precision at rank R, where R is the number of relevant documents for a given topic. However, unlike AP, these metrics are by definition totally insensitive to document swaps within top k (R). Precision at k also has a normalisation problem: its maximum value may be less than one for some topics.

The NTCIR test collections, as well as recent TREC test collections, provide graded relevance assessment data. However, being a binary-relevance metric, AP cannot directly utilise such data. Hence, as long as AP is used for optimisation, it is difficult for researchers to develop a system that can retrieve highly relevant documents on top of partially relevant documents. In light of this, graded-relevance metrics are in order.

The most popular graded-relevance metric to date is probably nDCG [7]. Although the original version of nDCG had a parameter for reflecting the user's patience, this version was a counterintuitive metric because of this very feature [17]. Thus the version of nDCG that is in fact widely used is the "Microsoft version" first introduced in [3]. This version does not have the aforementioned parameter and is free from the "bug" of the original nDCG, and is the one we use in our experiments. Another version of nDCG that is also bug-free, though not yet as widely-used as the Microsoft version, is described in [8].

Another well-studied graded-relevance metric is Q-measure, or simply Q [14]. This is a generalised version of AP and correlates very highly with it: the only difference between the two is that while AP relies on precision, Q relies on the aforementioned blended ratio. The NTCIR-6 crosslingual task has used Q and (a version of) nDCG along with AP for evaluating the participating systems [10]. The NTCIR-7 ACLIA IR4QA task uses AP, Q and the Microsoft version of nDCG [19].

The present study considers the blended ratio as the utility function of our NCU metrics. The NCU metrics subsume AP, Q, and something close to nCG (normalised cumulative gain), while the rank-based discounting is handled somewhat differently from nDCG.

Rank-biased precision (RBP) [11] can also handle graded relevance. It models a single user examining a document at rank i and then moving to one at rank $(i + 1)$. Based on the argument that the user usually does not know the number of relevant documents R, RBP does not have a recall component. However, it has a normalisation problem similar to that of simple precision, and lacks discriminative power due to lack of a recall component [18]. The present paper borrows the idea of rank-bias for considering a non-uniform probability distribution for the user's stopping behaviour. We will discuss the key differences between the idea of RBP and our rank-biased NCU metrics in Section 3.2.

Sakai [16] proposed some variants of Q for reflecting different stopping behaviours of users. *O-measure*, a graded-relevance version of RR, assumes that the user stops at the first relevant document found, regardless of its relevance level. *P-measure* and *P*$+$ assume that the user keeps going until he finds one of the most relevant documents in the ranked list. However, it is known that these metrics are not as discriminative as Q (just as RR is not as discriminative as AP) as they ignore all of the retrieved relevant documents below the assumed stopping point. In contrast, our graded-uniform NCU metrics do consider all relevant documents, while taking into account the effect of relevance level on the user's stopping behaviour.

Kazai, Piwowarski and Robertson [9] have discussed a probabilistic user model and proposed an effectiveness metric for Web search and structured document retrieval, but their study focusses on the user's post-query navigation. This issue is beyond the scope of the present study.

2.3 Criteria for Assessing Metrics

This section briefly describes existing methods that we use for comparing different retrieval effectiveness metrics.

The present study examines metrics from two perspectives: (1) How two system rankings produced by two different metrics resemble each other; and (2) How statistically *reliable* the metrics are.

Regarding Perspective (1), since AP is currently the *de facto* standard, we compare the system ranking of a metric with that of AP. For this purpose, we use Kendall's rank correlation and Yilmaz/Aslam/Robertson (YAR) rank correlation [26][1].

[1] We refrain from using its original name, *AP correlation*, to

Kendall's rank correlation is a monotonic function of the probability that a *randomly chosen* pair of ranked systems is ordered identically in the two rankings. Hence a swap near the top of a ranked list and that near the bottom of the same list has equal impact. However, for the purpose of ranking retrieval systems, for example, in a competition-style workshop such as NTCIR and TREC, the ranks near the top of the list are arguably more important than those near the bottom. In light of this, the recently-proposed Yilmaz/Aslam/Robertson rank correlation is a monotonic function of the probability that a randomly chosen system *and one ranked above it* are ordered identically in the two rankings. Like Kendall's rank correlation, YAR rank correlation lies between -1 and 1, but unlike Kendall's, it is not symmetrical. Yilmaz, Aslam and Robertson also provide a symmetric version, but we use the raw asymmetic YAR rank correlation by taking AP as the gold standard. When the errors (i.e., pairwise swaps with respect to the gold standard) are uniformly distributed over the ranked list being examined, YAR rank correlation is equivalent to Kendall's rank correlation.

Formally, let the size of the ranked lists be L. Let C be the number of system pairs that are ranked in the same order in both rankings, and let D be the number of system pairs that are ranked in opposite order in the two rankings. Kendall's rank correlation is given by:

$$Kendall = \frac{C - D}{L(L-1)/2} \; . \qquad (5)$$

For a given ranked list to be examined, let $n(i)$ be the number of systems *correctly* ranked above rank i in the list with respect to a gold-standard ranked list. YAR correlation is given by:

$$YAR = \frac{2}{L-1} \sum_{i=2}^{L} \frac{n(i)}{i-1} - 1 \; . \qquad (6)$$

Regarding the aforementioned Perspective (2), we measure the reliability of effectiveness metrics using Sakai's discriminative power, which represents the overall ability to detect pairwise statistical significance while guaranteeing that the probability of Type I Error is below a given threshold [15].

If there are L systems to be evaluated, then there are $L(L-1)/2$ system pairs. For each pair, we conduct a two-sided, paired *bootstrap hypothesis test* using $B = 1000$ *bootstrap samples* of the original topic set, obtained by sampling with replacement [6]. This yields $L(L-1)/2$ *achieved significance level* (ASL) values, also known as p-values. For a given threshold α, Sakai's discriminative power is defined as the proportion of system pairs with a statistically significant difference, i.e., those that satisfy $ASL < \alpha$.

Sakai's method also provides an estimate of the absolute performance difference required between two systems in order to detect a statistical significance. For each bootstrap hypothesis test concerning a particular system pair, we look at the $B * \alpha$-th largest absolute value among the studentised versions of the B *bootstrap replicates* of the performance difference under the null hypothesis. We then record the absolute value of the raw bootstrap replicate, i.e., the performance difference that corresponds to a particular bootstrap sample of topics. For example, if $B = 1000$ and $\alpha = 0.05$, we examine the fiftieth largest absolute value among the 1000 studentised values. The corresponding raw value represents a borderline between a significant difference and a nonsignificant one. Finally, we take the maximum of the $L(L-1)/2$ values in order to be conservative. More details can be found elsewhere [15].

For the purpose of comparing the reliability of different evaluation metrics, Sakai's method is known to yield results that are similar to those obtained by the more ad hoc method proposed earlier by Voorhees and Buckley, which empirically examines three degree of consistency between two experiments for determining which of two systems is better in absolute terms [22].

3 Proposed Metrics

$P(n)$ assumes binary relevance, but as discussed, some more general utility measures such as nDCG consider graded relevance judgements. We could incorporate graded relevance into Robertson's NCP in two different ways. One is to make the stopping probability depend on relevance grade – we would assume that the user is more likely to reach satisfaction, and therefore to stop, on encountering a more highly relevant document. One interpretation of this idea is that the stopping probabilities arise in a population of users – each individual user has a binary notion of relevance, but they disagree on where the boundary between relevant and nonrelevant sits.

The other is to include it in the utility part of the function: we would assume that more highly relevant documents are more useful. We can interpret this as being a statement about any individual user – that each user gets more benefit from documents of higher grade than those of lower grade. Both these two ideas are plausible, and they are complementary – therefore we can also consider combining them.

We also generalise NCP in another way. In between the uniformly-distributed stopping probability version NCP_u and the completely top-heavy NCP_1, we could consider a probability distribution which is somewhat top-heavy. We propose a formulation inspired by the RBP model.

avoid confusion.

3.1 Further notes on utility

The argument of this paper is based on a separation of the stopping-point issue (variable over a user population) from the utility (to an individual user with a given stopping point) of the ranking. As indicated above, this utility should not discount gains internally, because for a given stopping point, it does not matter where in the ranking up to this point any benefit occurs. Nevertheless, we still subsume into the utility measure a combination of benefit and cost (or effort). Thus precision as a utility measure is suitable, because it takes a very simple ratio of benefit (number of relevant retrieved) to effort (total number retrieved).

It would be possible to make a further separation of the utility measure into cost and benefit – this is the line taken in some of the work on effectiveness for XML retrieval (see e.g. [9]). However, in the present paper we restrict ourselves to separating out the stopping-point issue, and seek a utility measure which combines individual user effort and benefit.

3.2 Definitions

Utility

We begin by generalising the utility component of NCP as defined by Eq. (4). We replace $P(n)$ by *normalised utility* $NU(n)$, which should lie between 0 and 1, to obtain *Normalised Cumulative Utility* (NCU):

$$NCU = \sum_{n=1}^{\infty} p_s(n) NU(n) \; . \tag{7}$$

$NU(n)$ could be precision $P(n)$, but alternatively we can use a measure that is based on cumulative gain (following [7]) in order to handle graded relevance. Let $\mathcal{L}$ be a relevance level, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an $\mathcal{L}$-relevant document. For the NTCIR data, for example, let $gain(S) = 3$ for each S-relevant (highly relevant) document, $gain(A) = 2$ for each A-relevant (relevant) document, and $gain(B) = 1$ for each B-relevant (partially relevant) document. Let $R(\mathcal{L})$ denote the number of known $\mathcal{L}$-relevant documents for a topic, so that $\sum_{\mathcal{L}} R(\mathcal{L}) = R$. Let $g(n) = gain(\mathcal{L})$ if the document at rank n is $\mathcal{L}$-relevant and let $g(n) = 0$ otherwise. In particular, let $g^*(n)$ denote the gain at rank n of an *ideal* ranked output, where an ideal ranked output for a particular topic is one that satisfies $I(n) = 1$ for $1 \leq n \leq R$ and $g(n) \leq g(n-1)$ for $n > 1$. For the NTCIR data, this can be achieved by listing up all S-relevant documents, then all A-relevant documents, and then all B-relevant documents. (Whether some nonrelevant documents are included below these exhaustive list of relevant documents is of no consequence.) Using the above notations, for $NU(n)$ we use the *blended ratio* $BR(n)$:

$$BR(n) = \frac{C(n) + \beta \sum_{i=1}^{n} g(i)}{n + \beta \sum_{i=1}^{n} g^*(i)} \tag{8}$$

$C(n)$, as before, is the number of relevant documents seen by rank n, irrespective of relevance levels.

$BR(n)$ can be seen as a mixture of precision $P(n)$ and nCG, normalised cumulative gain. For $\beta = 0$ it reduces to $P(n)$, and for large β it approaches nCG. (Note that $C(n)$ and n in Eq. (8) are bounded above by the size of the ranked list, which in practice is no greater than 1000.) We do not at this point consider nDCG, normalised discounted cumulative gain, which applies a rank-based discount to the utility. The reason is that any desired top-heaviness or rank-based discount is provided by the probabilistic stopping rule; there is no reason to apply it also to the utility part of the measure.

In the experiments reported in this paper we use $BR(n)$ with $\beta = 0$ (which is equivalent to using $P(n)$), $\beta = 1$, and $\beta = 10000$ (approximating nCG: But see the note at Section 3.3). Sakai [17] has reported on the effect of varying β between 0 and 1000 for Q.

The formulation of NCU using $BR(n)$ reveals a link to another measure. Just as replacing $p_s(n)$ with the uniform probability distribution $p_u(n)$ and replacing $NU(n)$ with $P(n)$ reduces NCU to AP, replacing $p_s(n)$ with $p_u(n)$ and replacing $NU(n)$ with $BR(n)$ reduces NCU to Q:

$$Q\text{-}measure = \frac{1}{R} \sum_{n=1}^{\infty} I(n) BR(n) \; . \tag{9}$$

Stopping Probability

The assumption behind Robertson's $p_u(n)$ is that the user eventually stops at a relevant document with probability $1/R$ *regardless of the rank or the relevance level of the document*. Hence AP and Q-measure can also be interpreted as metrics based on this assumption. Below, we consider two alternative probability distributions, $p_{rb}(n)$ and $p_{gu}(n)$.

Robertson [12] notes that "*it is probably much more likely that a user would stop after few relevant documents than after many.*" Our first non-uniform probability distribution is based on this assumption. Let $\gamma (\leq 1)$ be a positive constant. We can define a *rank-biased* probability distribution $p_{rb}(n)$ over all relevant documents as follows. For each rank n where there is a nonrelevant document, let $p_{rb}(n) = 0$. Otherwise, let

$$p_{rb}(n) = \frac{\gamma^{C(n)-1}}{\sum_{i=1}^{R} \gamma^{i-1}} \tag{10}$$

The numerator decreases the stopping probability as the user goes down the ranked list. For example, for

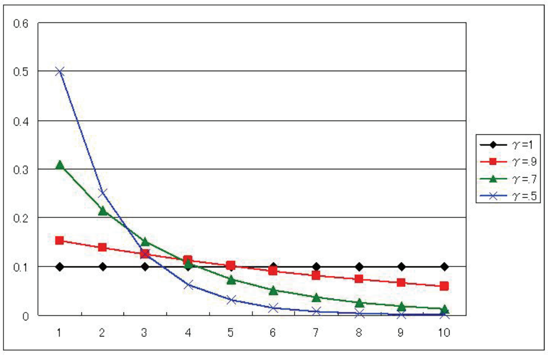

Figure 1. Rank-biased probability distribution over relevant documents ($R = 10$) for different values of γ.

the first relevant document found in the ranked list ($C(n) = 1$), the numerator would be one; for the second relevant document, it would be γ; for the third, it would be γ^2. The denominator is a constant for a given topic, ensuring that the probabilities sum to one. This function resembles the definition of RBP [11], which is based on the following model: the user persistence parameter p, which is the probability of the user continuing beyond any rank, *given that they have reached that rank*, is fixed. Thus the probability that a user will both reach rank n and continue from it is p^n.

The differences between our definition and RBP are as follows:

- We assume that the user will stop only at a *relevant* document;
- In order to normalise our measure over a finite number of relevant documents, we introduce the denominator of Eq. (10).

We also note that the definition of RBP does not specify how the probability arises, in other words over what population of events it is defined. In our model, we specifically assume a population of users, making it clear that the top-heaviness of our measure arises because of differences between users regarding the stopping point. In fact a model like that represented by RBP implies top-heaviness to a very high degree: Our model of Eq. (10) is slightly less top-heavy for the same value of γ and p, because in our model the user does not stop on nonrelevant documents.

Figure 1 illustrates, for a topic with $R = 10$ relevant documents, the p_{rb} curves for $\gamma = 1, 0.9, 0.7, 0.5$. Note that $\gamma = 1$ reduces $p_{rb}(n)$ to $p_u(n)$.

Another probability distribution that we consider, $p_{gu}(n)$, is based on the assumption that *it is much more likely that a user would stop (due to satisfaction) after a highly relevant document than a partially relevant document*. For simplicity, we assume that the stopping probability is uniform *within each relevance level*. For each relevance level $\mathcal{L}$, we define the *stopping weight*, $stop(\mathcal{L})$, which reflects how likely it is for the user to eventually stop at an $\mathcal{L}$-relevant document. For example, we can let $stop(B) : stop(A) : stop(S) = 1 : 2 : 3$, representing the assumption that the user is three times as likey to stop at an S-relevant document than at a B-relevant document, and so on. Moreover, let $S(n) = stop(\mathcal{L})$ whenever the document at rank n is $\mathcal{L}$-relevant and $S(n) = 0$ otherwise. Our *graded-uniform* probability distribution $p_{gu}(n)$ is defined as follows. For all n with a nonrelevant document, let $p_{gu}(n) = 0$. Otherwise, let

$$p_{gu}(n) = \frac{S(n)}{\sum_{\mathcal{L}} R(\mathcal{L}) stop(\mathcal{L})} \; . \tag{11}$$

Again, the denominator is a constant for a given topic, ensuring that the probabilities sum to one. Note that when the stopping weight is the same for all relevance levels, $p_{gu}(n)$ reduces to $p_u(n)$.

We now have a family of NCU metrics, with two fundamental parameters, namely, the stopping probability distribution $p_s(n) \in \{p_u(n), p_{rb}(n), p_{gu}(n)\}$ and the β parameter of $BR(n)$. AP, which uses $p_u(n)$ with $\beta = 0$, can be expressed as $NCU_{u,\beta=0}$; similarly, Q-measure with the default $\beta = 1$ can be expressed as $NCU_{u,\beta=1}$. We can also define a measure based on taking nCG as the utility function by $NCU_{u,\beta=\infty}$ (but see the note below). In practice we have run our experiments as $NCU_{u,\beta=10000}$.

3.3 A Note on nCG and Stopping

Actually the combination of pure nCG and the satisfaction-point stopping model is somewhat flawed, for the reason given in section 3.1. Pure nCG takes no account of effort (in the form of nonrelevant documents seen) beyond rank R, the total number of relevant documents. That is, the nCG values achieved at each relevant document beyond this point are independent of the number of nonrelevant documents preceding them. From this point of view, it is not a good cost-benefit measure.

This deficiency could actually be compensated by a suitable stopping-point model. However, the assumption of the satisfaction-point stopping model, that the user will only stop on a relevant document, means that $NCU_{*,\beta=\infty}$ itself also takes no account of nonrelevant documents beyond rank R. Thus the use of nCG would really only be consistent with a model which included some other stopping rule, such as frustration or exhaustion. The introduction of such a rule is beyond the scope of the present paper. Neither precision nor the blended ratio with any other β value suffers from this problem, although either might be made more realistic with more complex stopping rules.

3.4 Examples

Consider an NTCIR topic with $R(S) = 3, R(A) = 3, R(B) = 4$ and therefore $R = 10$ relevant docu-

Table 1. Computing $p_s(n)$ and $BR(n)$ for a topic with $R = 10$ ($R(S) = 3, R(A) = 3, R(B) = 4$) relevant documents: An example.

rank n	rel. level	$p_u(n)$	$p_{rb}(n)$ with $\gamma = 0.7$	$p_{gu}(n)$ (1:2:3)	$BR(n)$ with $\beta = 0$ ($P(n)$)	$BR(n)$ with $\beta = 1$
2	S	.1	1/3.2392=.3087	3/19=.1579	1/2=.5000	(1+3)/(2+6)=.5000
5	A	.1	.7/3.2392=.2161	2/19=.1053	2/5=.4000	(2+5)/(5+13)=.3889
8	S	.1	.49/3.2392=.1513	3/19=.1579	3/8=.3750	(3+8)/(8+17)=.4400
12	B	.1	.343/3.2392=.1059	1/19=.0526	4/12=.3333	(4+9)/(12+19)=.4194
15	A	.1	.2401/3.2392=.0741	2/19=.1053	5/15=.3333	(5+11)/(15+19)=.4706

Table 2. Computing NCU metrics for a topic with $R = 10$ ($R(S) = 3, R(A) = 3, R(B) = 4$) relevant documents, using Table 1.

	$\beta = 0$	$\beta = 1$
p_u	.1942 (AP)	.2219 (Q)
p_{rb}	.3575	.3842
p_{gu}	.2329	.2610

Table 3. TREC and NTCIR data used in our experiments.

	NTCIR-6J	TREC03
#topics	50	50
#documents	858,400	approx. 528,000
pool depth	100	125
average N	1157.9	925.5
range N	[480, 2732]	[292, 2050]
average R	95.3	33.2
range R	[4, 311]	[4, 115]
S-relevant	2.5	8.1
A-relevant	61.1	-
B-relevant	31.7	25.0
#teams	10	16
#all runs	74	78

ments. Now, consider a ranked list of documents that has an S-relevant document at ranks 2 and 8, an A-relevant document at ranks 5 and 15, and a B-relevant document at rank 12. Suppose that the other five relevant documents were not retrieved. Our NCU metrics can be computed using the values of stopping probabilities ($p_u(n), p_{rb}(n), p_{gu}(n)$) and the β parameter of $BR(n)$ shown in Table 1. For handling graded relevance, here we use $gain(B) : gain(A) : gain(S) = stop(B) : stop(A) : stop(S) = 1 : 2 : 3$. The parameter for the rank-biased probability distribution p_{rb} is set to $\gamma = 0.7$.

Table 2 shows the values of the NCU metrics computed based on Table 1. For example, $NCU_{u,\beta=0}$ (i.e., AP) is .1942.

4 Experiments

4.1 Data

Table 3 shows some statistics of the two data sets we used for comparing our effectiveness metrics. Our first data set, which we call NTCIR-6J, is from the Stage 1 Japanese document retrieval subtask of the NTCIR-6 crosslingual task [10]. The data contains 74 runs including monolingual and crosslingual runs. Our second data set, which we call TREC03, is from the TREC 2003 robust track using the 50 new topics [23]. N and R represent the number of judged nonrelevant and relevant documents, respectively. The TREC03 relevance assessments contain "highly relevant" and "relevant" documents, but we treated the former as S-relevant (highly relevant) and the latter as B-relevant (partially relevant). This is because it is known that many TREC relevant documents are in fact partially or marginally relevant [13, 20].

Some of our metrics require parameter values for utilising the above graded relevance data: the gain values $gain(\mathcal{L})$ and the stopping weights $stop(\mathcal{L})$. Recall that the former represents the utility for obtaining an $\mathcal{L}$-relevant document, while the latter represents the likelihood of the user eventually stopping at an $\mathcal{L}$-relevant document. However, for simplicity, we use the same set of values for $gain(\mathcal{L})$ and $stop(\mathcal{L})$. In this paper, we consider two cases: $gain(B) : gain(A) : gain(S) = stop(B) : stop(A) : stop(S) = 1 : 2 : 3$, and $gain(B) : gain(A) : gain(S) = stop(B) : stop(A) : stop(S) = 1 : 5 : 10$. These parameter settings will be denoted simply by 1:2:3 and 1:5:10, respectively.

For computing rank correlations between two system rankings according to two different metrics, we used all runs shown in Table 3. For computing discriminative power, which is based on *pairs* of runs, we randomly selected one run from each team. For NTCIR-6J, we selected one *monolingual* run from each team.

4.2 Results and Discussions

First, we discuss the resemblance of two system rankings according to two different metrics: We compare the ranking according to an NCU metric with that according to AP, as AP is the *de facto* standard. Tables 4 and 5 show the Kendall's rank correlation and YAR rank correlation results, respectively. For simplicity, $NCU_{rb,\beta=0}$ is represented by "$rb, \beta = 0$", and so on. For the rank-biased NCU metrics ($NCU_{rb,*}$), we tried $\gamma = 1, 0.9, 0.7, 0.5$ but recall that $\gamma = 1$ reduces p_{rb} to the original uniform distribution p_u. As for the graded-uniform NCU metrics ($NCU_{gu,*}$), the results using the parameter settings 1:2:3 and 1:5:10 are shown in the top half and the bottom half of each table, respectively. The results for $NCU_{u,\beta=0}$ (i.e., AP) are omitted in (c) and (d) because using the graded-relevance parameter setting does not af-

Table 4. Kendall's rank correlation with AP (i.e., $NCU_{u,\beta=0}$).

	(a) NTCIR6J, 74 runs (1:2:3)				(b) TREC03, 78 runs (1:2:3)			
	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$
$rb, \beta = 0$	**1** (AP)	.843	.743	.685	**1** (AP)	.843	.707	.652
$rb, \beta = 1$	**.967** (Q)	.833	.724	.673	**.936** (Q)	.855	.710	.639
$rb, \beta = 10000$	.862 (Q)	.822	.710	.670	.857 (Q)	.819	.690	.624
$gu, \beta = 0$	**.977**	-	-	-	**.957**	-	-	-
$gu, \beta = 1$	**.961**	-	-	-	**.951**	-	-	-
$gu, \beta = 10000$	.896	-	-	-	.848	-	-	-
	(c) NTCIR6J, 74 runs (1:5:10)				(d) TREC03, 78 runs (1:5:10)			
	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$
$rb, \beta = 1$	**.938** (Q)	.818	.702	.654	.890 (Q)	.843	.682	.609
$rb, \beta = 10000$	.841 (Q)	.796	.688	.634	.818 (Q)	.776	.657	.592
$gu, \beta = 0$	**.957**	-	-	-	**.913**	-	-	-
$gu, \beta = 1$	**.939**	-	-	-	**.913**	-	-	-
$gu, \beta = 10000$	.893	-	-	-	.826	-	-	-

Table 5. Yilmaz/Aslam/Robertson rank correlation with AP (i.e., $NCU_{u,\beta=0}$).

	(a) NTCIR6J, 74 runs (1:2:3)				(b) TREC03, 78 runs (1:2:3)			
	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$
$rb, \beta = 0$	**1** (AP)	.773	.660	.604	**1** (AP)	.761	.601	.535
$rb, \beta = 1$	**.954** (Q)	.740	.628	.589	.893 (Q)	.776	.595	.524
$rb, \beta = 10000$	.788 (Q)	.729	.613	.584	.786 (Q)	.744	.569	.507
$gu, \beta = 0$	**.960**	-	-	-	**.925**	-	-	-
$gu, \beta = 1$	.890	-	-	-	**.909**	-	-	-
$gu, \beta = 10000$	.808	-	-	-	.766	-	-	-
	(c) NTCIR6J, 74 runs (1:5:10)				(d) TREC03, 78 runs (1:5:10)			
	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 1\,(u)$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$
$rb, \beta = 1$	**.908** (Q)	.743	.608	.559	.807 (Q)	.749	.564	.493
$rb, \beta = 10000$	.764 (Q)	.721	.597	.542	.729 (Q)	.677	.536	.473
$gu, \beta = 0$	**.925**	-	-	-	.865	-	-	-
$gu, \beta = 1$	**.927**	-	-	-	.829	-	-	-
$gu, \beta = 10000$	.815	-	-	-	.735	-	-	-

fect them. For convenience, values higher than 0.9 are shown in bold.

The following observations can be made from Tables 4 and 5:

(1) Heavy rank bias over relevant documents yields metrics that are substantially different from AP. For example, Table 4(b) shows that the Kendall's rank correlation between $NCU_{rb,\beta=0}$ with $\gamma = 0.5$ and AP are only .652 for TREC03. The corresponding YAR rank correlation in Table 5(b) is even lower: .535.

(2) The system rankings according to $NCU_{u,\beta=1}$ (i.e., Q with $\beta = 1$), $NCU_{gu,\beta=0}$ and $NCU_{gu,\beta=1}$ are generally very similar to that according to AP. For example, Table 4(a) shows that the Kendall's rank correlation between $NCU_{gu,\beta=1}$ and AP is .961 for NTCIR-6J. Whereas, the rankings according to $NCU_{*,\beta=10000}$ are quite different from that according to AP.

(3) The YAR rank correlation values in Table 5 are generally lower than the corresponding Kendall' rank correlation values in Table 4, from which it follows that the ranking "errors" (See Section 2.3) are *not* evenly distributed across the ranked list.

(4) The results are generally consistent across NTCIR and TREC.

Observation (1) means that AP is not consistent with heavy rank bias over relevant documents, i.e., small γ. Or in other words, the small γ metric measures something different from AP. We might conclude that AP is not as top-heavy as some user models would suggest. Observation (2) means that it is possible to utilise graded relevance in the form of gain values and/or stopping weights *and* maintain consistency with AP, if a small β is chosen. Observation (3) demonstrates that the recently-proposd YAR rank correlation is indeed useful.

Next, we discuss discriminative power, the overall ability of a metric to detect statistical significance given a significance level. Since we use 10 runs from NTCIR-6J, we have 10*9/2=45 run pairs for this data set. Similarly, with TREC03, we have 16*15/2=120 run pairs. Table 6 summarises the results at $\alpha = 0.05$, i.e., 95% confidence. For example, Table 6(a) shows that, for the NTCIR-6J data set, the discriminative power of AP ($NCU_{u,\beta=0}$) at $\alpha = 0.05$ is 57.8%: It manages to detect a statisitical significant difference for 26 run pairs out of 45. Moreover, given 50 topics, the estimated overall performance difference required to achieve statistical significance is 0.08. That is, if two systems differ by at least 0.08 in average performance, this difference is usually statistically significant.

The following observations can be made from Table 6:

(i) Heavy rank bias over relevant documents hurts discriminative power. For example, Table 6(b) shows that, at $\alpha = 0.05$, while the discriminative power of AP is 64.2%, that of $NCU_{rb,\beta=0}$

Table 6. Discriminative power at $\alpha = 0.05$.

	(a) NTCIR6J, 10 teams (1:2:3)				(b) TREC03, 16 teams (1:2:3)			
	$\gamma = 1$ (u)	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 1$ (u)	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$
$rb, \beta = 0$	26/45=57.8 (AP) 0.08	25/45=55.6 0.09	24/45=53.3 0.10	24/45=53.3 0.13	77/120=64.2 (AP) 0.07	65/120=54.2 0.11	56/120=46.7 0.14	50/120=41.7 0.14
$rb, \beta = 1$	28/45=62.2 (Q) 0.07	27/45=60.0 0.08	22/45=48.9 0.10	22/45=48.9 0.11	80/120=66.7 (Q) 0.07	75/120=62.5 0.09	55/120=45.8 0.11	49/120=40.8 0.12
$rb, \beta = 10000$	29/45=64.4 (Q) 0.09	29/45=64.4 0.07	21/45=46.7 0.10	18/45=40.0 0.11	70/120=58.3 (Q) 0.07	64/120=53.3 0.08	52/120=43.3 0.10	46/120=38.3 0.12
$gu, \beta = 0$	26/45=57.8 0.07	-	-	-	77/120=64.2 0.08	-	-	-
$gu, \beta = 1$	29/45=64.4 0.08	-	-	-	82/120=68.3 0.08	-	-	-
$gu, \beta = 10000$	29/45=64.4 0.08	-	-	-	72/120=60.0 0.08	-	-	-
	(c) NTCIR6J, 10 teams (1:5:10)				(d) TREC03, 16 teams (1:5:10)			
	$\gamma = 1$ (u)	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$	$\gamma = 1$ (u)	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.5$
$rb, \beta = 1$	29/45=64.4 (Q) 0.08	26/45=57.8 0.09	17/45=37.8 0.11	14/45=31.1 0.13	78/120=65.0 (Q) 0.08	68/120=56.7 0.09	45/120=37.5 0.10	38/120=31.7 0.12
$rb, \beta = 10000$	30/45=66.7 (Q) 0.07	27/45=60.0 0.10	16/45=35.6 0.10	12/45=26.7 0.13	57/120=47.5 (Q) 0.08	56/120=46.7 0.08	40/120=33.3 0.10	36/120=30.0 0.12
$gu, \beta = 0$	27/45=60.0 0.08	-	-	-	72/120=60.0 0.09	-	-	-
$gu, \beta = 1$	31/45=68.9 0.08	-	-	-	78/120=65.0 0.08	-	-	-
$gu, \beta = 10000$	29/45=64.4 0.07	-	-	-	68/120=56.7 0.09	-	-	-

with $\gamma = 0.5$ is only 41.7% for TREC03.

(ii) Because the heavily rank-biased NCU metrics lack discriminative power, they require a relatively large overall performance difference for achieving statistical significance. For example, Table 6(b) shows that, while a performance difference of 0.07 in Mean AP is usually statistically significant, a performance difference in Mean $NCU_{rb,\beta=0}$ with $\gamma = 0.5$ reaches statistical significance only when it is around 0.14.

(iii) Utilising graded relevance in the form of gain values and/or stopping weights can result in higher discriminative power. For example, while the discriminative power of AP at $\alpha = 0.05$ for NTCIR-6J is 57.8% (Table 6(a)), that of $NCU_{gu,\beta=1}$ with 1:5:10 is 68.9% (Table 6(c)). For NTCIR-6J, $NCU_{gu,\beta=1}$ with 1:5:10 is the most discriminative among our NCU metrics, while for TREC03, the same metric with 1:2:3 is the most discriminative (Table 6(b)).

(iv) Most of the results are consistent across NTCIR and TREC: Even the overall performance differences required are similar. However, the $\beta = 10000$ results are exceptions: $NCU_{rb,\beta=1000}$ with $\gamma = 1$ (i.e., $NCU_{u,\beta=1000}$) and $NCU_{gu,\beta=1000}$ show high discriminative power for NTCIR, but relatively low discriminative power for TREC.

Observations (i) and (ii) suggest that it is a good idea to look beyond the stopping point of an ordinary user for obtaining reliable conclusions from experiments. Even if users tend to stop examining the ranked list near the top of the list, it does not follow that researchers should follow exactly the same strategy. Observation (iii) generalises previous findings by Sakai, who demonstrated the high discriminative power of graded-relevance metrics such as Q [15, 17]. Note that $NCU_{gu,\beta=1}$ is even more discriminative than Q in some of our experiments. As for Observation (iv), the estimated overall performance differences for achieving statistical significance are similar across NTCIR-6J and TREC03 not only because the two data sets both use 50 topics, but also because the performance distributions of the runs involved are reasonably similar. For example, if we used a set of runs that are extremely easy to distinguish from one another, then the required performance differences would be very small.

The above discussions of discriminative power used $\alpha = 0.05$ for the statistical significance tests, but the choice of this threshold is arbitrary. We therefore provide an overview across different significance levels below.

Figures 2 and 3 show the *achieved significance level (ASL) curves* [15] of $NCU_{rb,\beta=0}$ with different values of γ for NTCIR-6J and TREC03, respectively. For example, the vertical axis of Figure 3 represents ASL, and the horizontal axis represents the 120 run pairs sorted by the ASL values. Note that low ASL values yield high discriminative power, since a run pair is statistically significant when $ASL < \alpha$. Figure 3 clearly shows that smaller values of γ gradually hurt discriminative power for TREC03. The NTCIR-6J results in Figure 2 are less clear, possibly because of the smaller number of run pairs.

Figures 4 and 5 show the ASL curves of $NCU_{u,*}$ and $NCU_{gu,*}$ with 1:2:3 for NTCIR-6J and TREC03, respectively. It can be observed, for example, that $NCU_{u,\beta=0}$ (i.e., AP) and $NCU_{gu,\beta=0}$ are less discriminative than other metrics for NTCIR-6J, while

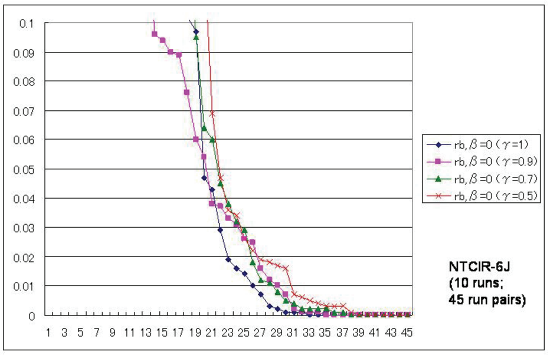

Figure 2. ASL curves of $NCU_{rb,\beta=0}$ for different values of γ (NTCIR-6J).

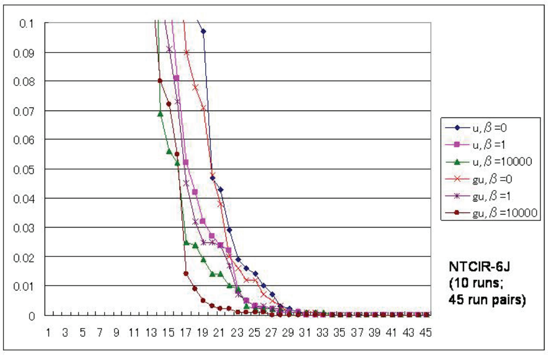

Figure 4. ASL curves of $NCU_{u,*}$ and $NCU_{gu,*}$ with 1:2:3 (NTCIR-6J).

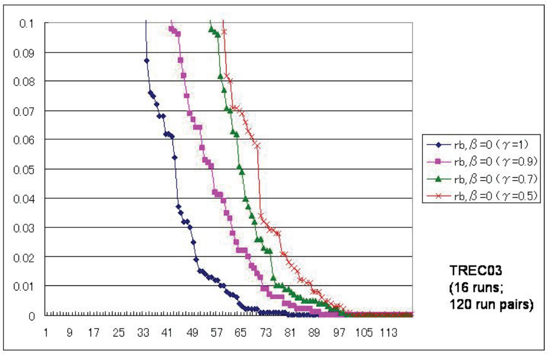

Figure 3. ASL curves of $NCU_{rb,\beta=0}$ for different values of γ (TREC03).

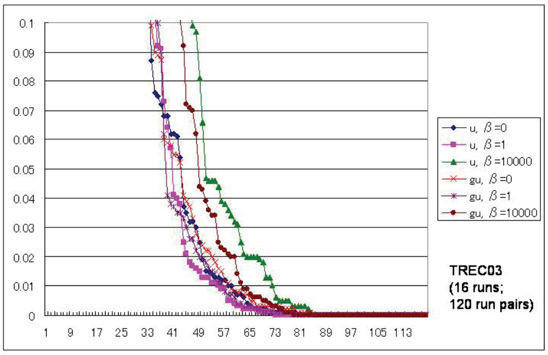

Figure 5. ASL curves of $NCU_{u,*}$ and $NCU_{gu,*}$ with 1:2:3 (TREC03).

$NCU_{u,\beta=10000}$ and $NCU_{gu,\beta=1000}$ are less discriminative than other metrics for TREC03. We discussed this inconsistency in Observation (iv) above. Whereas, $NCU_{u,\beta=1}$ (i.e., Q) and $NCU_{gu,\beta=1}$ do well for both NTCIR and TREC.

To sum up our findings:

- Heavily rank-biased metrics yield system rankings that are very different from that based on AP. Moreover, they lack discriminative power. This suggests that it is a good idea to look beyond the stopping point of an ordinary user for obtaining reliable conclusions from experiments. Hence, metrics such as AP and Q, which rely on a uniform distribution across all relevant documents, may in fact be very reasonable.

- Utilising graded relevance, in the form of gain values and/or stopping weights, can provide both high consistency with AP and higher discriminative power than AP. According to our experiments using both NTCIR and TREC data, $NCU_{gu,\beta=1}$ appears to be a good choice among the family of NCU metrics.

5 Conclusions

In this paper, we generalised Robertson's NCP, which assumes a uniform stopping probability distribution (p_u) over all relevant documents and uses precision (P) as its utility function, in two ways:

1. We considered two new probability distributions over all relevant documents, namely, a rank-biased one (p_{rb}) and a graded-uniform (p_{gu}) one.
2. We considered a generalised utility function that can handle graded relevance, namely, the blended ratio (BR).

Our experiments using data from both NTCIR and TREC sugget that introducing a rank-biased distribution over relevant documents is not necessarily desirable, and that AP and its graded-relevance version Q, which rely on a uniform probability distribution, are in fact reasonable metrics. From a probabilistic perspective, these metrics emphasise long-tail users who tend to dig deep into the ranked list, and thereby achieve high reliability. Moreover, one of our new metrics $NCU_{gu,\beta=1}$ maintains high consistency with AP *and* achieve the highest discriminative power among our

NCU metrics, by utilising graded relevance in two ways: First, as a measure of utility for obtaining an $\mathcal{L}$-relevant document, and second, as the likelihood of the user eventually stopping at an $\mathcal{L}$-relevant document. An implementation of all of the aforementioned metrics is available at `http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en`.

The present study used three criteria for comparing metrics: Kendall's rank correlation, the recently-proposed YAR rank correlation, and discriminative power. The YAR rank correlation was proposed because the widely-used Kendall's rank correlation cannot emphasise change near the top of a system ranking. The *bootstrap sensitivity method* [15] which we used for computing discriminative power was proposed to replace the more ad hoc Voorhees/Buckley *swap method* [22]. Other researchers have tried to directly measure the relationship between effectiveness metrics and "user performance" [21] or "user satisfaction" [1], and reported some negative results for metrics such as AP and nDCG. Hence, currently there is no standard set of criteria for discussing which metric is better than another. In future work, we would like to consider other possible criteria for choosing good effectiveness metrics, including the ability to predict the behaviour of a simple, intuitive metric with an unknown data set, i.e., topics and documents [25].

We also plan to to extend the idea of NCU further. For example, the score standardisation technique introduced by Webber, Moffat and Zobel [24] can easily be incorporated into our framework. Moreover, it may be important to design effectiveness metrics that reflect the construction process of a test collection, for example, how topics are sampled, how documents to be judged for relevance are selected, and how assessors judge graded relevance.

Acknowledgements

We are very grateful to Emine Yilmaz for useful discussions on this paper.

References

[1] Al-Maskari, A., Sanderson, M. and Clough, P.: The Relationship between IR Effectiveness Measures and User Satisfaction, *Proceedings of ACM SIGIR 2007*, pp. 773-774, 2007.

[2] Buckley, C. and Voorhees, E. M.: Retrieval System Evaluation. In Voorhees E. M. and Harman, D. K. (eds.), *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[3] Burges, C. *et al.*: Learning to Rank using Gradient Descent, *Proceeding of ACM ICML 2005*, pp. 89-96, 2005.

[4] Cooper, W. S.: Expected Search Length: A Single Measure of Retrieval Effectiveness based on the Weak Ordering Action of Retrieval Systems, *American Documentation*, 19, pp. 30-41, 1968.

[5] Dunlop, M. D.: Time, Relevance and Interaction Modelling for Information Retrieval, *Proceedings of ACM SIGIR '97*, pp. 206-213, 1997.

[6] Efron, B. and Tibshirani, R.: *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1993.

[7] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422-446, 2002.

[8] Järvelin, K., Price, S. L., Delcambre, L. M. L. and Nielsen, M. L.: Discounted Cumulative Gain Based Evaluation of Multiple-Query IR Sessions, *Proceedings of ECIR 2008*, LNCS 4956, pp. 4-15, 2008.

[9] Kazai, G., Piwowarski, B. and Robertson, S.: Effort-Precision and Gain-Recall based on a Probabilistic User Navigation Model, *Proceedings of ICTIR 2007*, 2007.

[10] Kishida, K. *et al.*: Overview of CLIR Task at the Sixth NTCIR Workshop, *Proceedings of NTCIR-6*, 2007.

[11] Moffat, A., Webber, W. and Zobel, J.: Strategic System Comparisons via Targeted Relevance Judgments, *Proceedings of ACM SIGIR 2007*, pp. 375-382, 2007.

[12] Robertson, S.: A New Interpretation of Average Precision, *Proceedings of ACM SIGIR 2008*, pp. 689-690, 2008.

[13] Sakai, T. and Sparck Jones, K.: Generic Summaries for Indexing in Information Retrieval, *Proceedings of ACM SIGIR 2001*, pp.190-198, 2001.

[14] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, 43(2), pp. 531-548, 2007.

[15] Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, *Information Processing Society of Japan Transactions on Databases*, Vol.48, No.SIG 9 (TOD35), pp.11-28, 2007. Available at: `http://www.jstage.jst.go.jp/article/ipsjdc/3/0/625/_pdf`

[16] Sakai, T.: On the Properties of Evaluation Metrics for Finding One Highly Relevant Document, *Information Processing Society of Japan Transactions on Databases*, Vol.48,

No.SIG 9 (TOD35), pp.29-46, 2007. Available at: http://www.jstage.jst.go.jp/article/ipsjdc/3/0/643/_pdf

[17] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pp.32-43, 2007. Available at: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf

[18] Sakai, T. and Kando, N.: On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, *Information Retrieval*, Vol.11, No.5, pp.447-470, Springer, 2008. Available at: http://www.springerlink.com/content/k41j115214032614/fulltext.pdf

[19] Sakai, T. *et al.*: Overview of the NTCIR-7 ACLIA IR4QA Task, *Proceedings of NTCIR-7*, to appear, 2008.

[20] Sormunen, E.: Liberal Relevance Criteria of TREC - Counting on Negligible Documents? *Proceedings of ACM SIGIR 2002*, pp. 324-330, 2002.

[21] Turpin, A. and Scholer, F.: User Performance versus Precision Measures for Simple Search Tasks, *Proceedings of ACM SIGIR 2006*, pp. 11-18, 2006.

[22] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *Proceedings of ACM SIGIR 2002*, pp. 316-323, 2002.

[23] Voorhees, E. M.: Overview of the TREC 2003 Robust Retrieval Track, *Proceedings of TREC 2003*, 2004.

[24] Webber, W., Moffat, A., Zobel, J.: Score Standardization for Inter-Collection Comparison of Retrieval Systems, *Proceedings of ACM SIGIR 2008*, pp. 51-58, 2008.

[25] Webber, W., Moffat, A., Zobel, J. and Sakai, T.: Precision-At-Ten Considered Redundant, *Proceedings of ACM SIGIR 2008*, pp. 695-696, 2008.

[26] Yilmaz, E., Aslam, J. and Robertson, S.: A New Rank Correlation Coefficient for Information Retrieval, *Proceedings of ACM SIGIR 2008*, pp. 587-594, 2008.