# On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance

Tetsuya Sakai
Toshiba Corporate R&D Center / NewsWatch, Inc. (current affiliation)
sakai@newswatch.co.jp

## Abstract

*Large-scale information retrieval evaluation efforts such as TREC and NTCIR have tended to adhere to binary-relevance evaluation metrics, even when graded relevance data were available. However, the NTCIR-6 Crosslingual Task has finally started adopting graded-relevance metrics, though only as additional metrics. This paper compares three existing graded-relevance metrics that were mentioned in the Call for Participation of the NTCIR-6 Crosslingual Task in terms of the ability to control how severely "late arrival" of relevant documents should be penalised. We argue and demonstrate that Q-measure is more flexible than normalised Discounted Cumulative Gain and generalised Average Precision. We then suggest a brief guideline for conducting a reliable information retrieval evaluation with graded relevance.*

**Keywords:** *Q-measure, nDCG, generalised average precision, rank correlation, bootstrap sensitivity.*

## 1 Introduction

The Information Retrieval (IR) tasks at NTCIR [6] have tended to adhere to IR evaluation metrics based on *binary* relevance, most notably *Average Precision* (AveP), even though they have test collections with *graded* relevance assessments. In evaluations based on binary relevance, relevant documents with different degrees of relevance are treated as if they are of equal value. This is a curious situation since, by definition, we expect real users to prefer highly relevant documents to only partially/marginally relevant ones. The only exception at NTCIR is the currently discontinued Web track [11], which used the *unnormalised* Discounted Cumlative Gain (DCG) metric proposed by Järvelin and Kekäläinen [4][1]. However, while the unnormalised DCG can utilise graded relevance, it is known that the metric takes arbitrarily large values for topics with many relevant documents, and are not suitable for averaging across topics [17].

The situations are similar outside Asia: For example, the Robust Track and the Genomics track at TREC 2005 [3, 22] used binary AveP and related metrics such as *bpref* [2, 18], thereby failing to exploit the relevance levels that are available. As long as researchers keep evaluating their systems based on binary relevance, it is doubtful that they will ever be able to build a system that retrieves highly relevant documents on top of partially relevant ones.

In 2002, Järvelin and Kekäläinen [5] proposed *normalised DCG* (nDCG), which compares a system output with an *ideal* ranked ouput (See Section 3) and is therefore averageable across topics. At NTCIR-4, Sakai [12] also proposed an averageable graded-relevance metric called *Q-measure* which is very highly correlated with AveP and is at least as *stable* and *discriminative* as AveP [14, 17]. However, neither of these graded-relevance metrics was used officially at NTCIR-5.

At last, the Call for Participation for the NTCIR-6 Crosslingual Task (as of April 2006) announced that graded-relevance metrics will be used for ranking the participating systems, though only as *additional* metrics. The metrics mentioned in the CFP were nDCG, Q-measure and *generalised Average Precision* (genAveP) recently proposed by Kishida [10].

The objective of this paper[2] is to clarify the advantages of Q-measure over nDCG and genAveP from the viewpoint of *flexibility*, by which we mean the ability to control how severely "late arrival" of relevant documents should be penalised. IR metrics based on graded-relevance are required to:

(a) Prefer systems that return highly relevant documents to those that return partially relevant documents;

(b) Prefer systems that have relevant documents near the top of the ranked list to those that have relevant documents near the bottom.

[1] The NTCIR Web track also proposed a graded-relevance metric called Weighted Reciprocal Rank (WRR), but what the track actually used was the traditional *binary* Reciprocal Rank. See [16] for details.

[2] An early version of this paper was published as an unrefereed technical report [15].

All of the above three metrics can control the impact of (a), i.e., that of relevance levels, by means of a set of parameters called the *gain values* [4, 5]. However, we show that only Q-measure can properly control the impact of (b), i.e., that of ranks of relevant documents, by adjusting one of its parameters called $\beta$ (See Section 3). Our experiments using the Chinese/Japanese test collections and submitted runs from NTCIR-5 show that Q-measure can maintain reliable system ranking and high discriminative power for different choices of $\beta$. Finally, we suggest a brief guideline for conducting a reliable information retrieval evaluation with graded relevance.

Section 2 provides an overview of related studies. Section 3 defines and discusses the characteristics of AveP, nDCG, Q-measure and genAveP. Section 4 describes our experimental methods. Section 5 presents our results and provides discussions. Finally, Section 6 concludes this paper.

## 2 Related Work

All graded-relevance metrics considered in this paper are based on *Cumulative Gain* proposed at ACM SIGIR 2000 by Järvelin and Kekäläinen [4]. The same authors proposed normalised Cumulative Gain (nCG) and normalised Discounted Cumulative Gain (nDCG) in 2002 [5]. However, it is known that nCG has a defect, namely, that it cannot penalise late arrival of relevant documents properly. For example, for a topic with $R = 10$ relevant documents, a system that has only one relevant document at Rank 10 and a system that has only one relevant document at Rank 1000 are equally effective according to nCG at document cut-off 1000 [17]. Whereas, nDCG alleviates this problem by discounting each gain by the logarithm of the document rank. The *logarithm base* ($a$) for discounting has been interpreted as a parameter for controlling how severely late arrival of relevant documents should be penalised. Section 3 will provide more details.

Unlike nDCG which uses the *document rank* as the basis for comparison across topics, Sakai's Q-measure [12] is akin to AveP in that it uses *recall* as the basis. As we shall see in Section 3, Q-measure has its own late arrival parameter called $\beta$. Sakai [17] compared the stability and discriminative power of graded-relevance metrics such as Q-measure and n(D)CG using the *stability* method proposed at SIGIR 2000 [1] and the *swap* method proposed at SIGIR 2002 [21], as well as Kendall's rank correlation. He showed that both Q-measure and nDCG are stable and discriminative, but that a large document cut-off should be used with nDCG. However, his study did not explictly investigate the effect of the late arrival parameters (Q-measure's $\beta$ and nDCG's $a$). Moreover, it did not consider genAveP at all.

Kekäläinen [9] reported that shed tried both $a = 2$ and $a = 10$ with nDCG for some TREC data with their own graded relevance assessments, but that "the results regarding system performance differences did not differ notably." However, the present study demonstrates that if a large value of $a$ is used, nDCG becomes counterintuitive and insensitive, due to the fact that such a metric inherits the aforementioned defect of nCG for handling late arrivals.

At SIGIR 2006, Sakai [14] proposed the *Bootstrap Sensitivity Method* for comparing IR metrics in terms of discriminative power that is less *ad hoc* than the stability and the swap methods. This method relies on time-honoured *Bootstrap Hypothesis Tests* and yet yields results that are very similar to those based on the *ad hoc* methods. This paper therefore adopts this method for comparing the discriminative power of Q-measure, nDCG and genAveP with different parameter settings. Again, Sakai's SIGIR paper studied neither the effect of the late arrival parameters (Q-measure's $\beta$ and nDCG's $a$) nor the properties of genAveP.

Kishida [10] proposed genAveP and compared it with Q-measure and Average nDCG using a small-scale, artificial data set from the viewpoint of system ranking. He did not use real data, and did not discuss the stability and discriminative power of genAveP. Vu and Gallinari [23] also proposed a generalised version of AveP and compared it with Q-measure for an INEX XML retrieval task, but their metric does not average well (See Section 3). They tried a few values for Q-measure's $\beta (= 0.1, 1, 10)$, and reported that the choice affects both system ranking and discriminative power for the XML task. As we shall see later, our own results suggest the contrary, at least for traditional document retrieval tasks. It should also be noted that Vu and Gallinari used only one data set (namely, "INEX'03") in their experiments. Kazai and Lalmas [7] have also used Q-measure along with their own metrics designed for the INEX XML retrieval task, but the effect of $\beta$ was out of their scope.

Finally, note that the metrics studied in this paper (Q-measure, nDCG, genAveP and AveP) are those designed for the traditional task of finding *as many relevant documents as possible*, and that there are other kinds of IR tasks: Sakai [13, 16] have examined the resemblance, stability and discriminative power of IR metrics for the task of finding *one highly relevant document only*, namely, *$P^+$-measure, P-measure, O-measure* and *Weighted Reciprocal Rank*.

The contributions of this paper can be summarised as follows:

1. This is the first study to compare the effect of late arrival parameters (Q-measure's $\beta$ and nDCG's $a$) on system ranking and discriminative power, and to demonstrate that while nDCG quickly becomes counterintuitive and insensitive as $a$ gets large, Q-measure remains quite stable;

2. This is the first study to demonstrate that genAveP is a reasonable metric through experiments using real data, but that it cannot control how severely late arrival of relevant documents should be penalised.

## 3 Metrics

### 3.1 Definitions

Let $R$ denote the number of relevant documents for a topic, and let $L$ ($\leq L' = 1000$) denote the size of a ranked output. For each rank $r$ ($\leq L$), let $isrel(r)$ be 1 if the document at Rank $r$ is relevant and 0 otherwise, and let $count(r) = \sum_{1 \leq i \leq r} isrel(i)$. Clearly, precision at Rank $r$ is given by $P(r) = count(r)/r$. Hence:

$$AveP = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r)P(r) \ . \tag{1}$$

Let $R(\mathcal{L})$ denote the number of $\mathcal{L}$-relevant documents so that $\sum_{\mathcal{L}} R(\mathcal{L}) = R$, and let $gain(\mathcal{L})$ denote the *gain value* (i.e., reward) for retrieving an $\mathcal{L}$-relevant document [5]. For example, for an NTCIR crosslingual test collection which has S-, A- and B-relevant (highly relevant, relevant and partially relevant) documents, we let $gain(S) = 3, gain(A) = 2, gain(B) = 1$ by default. Let $cg(r) = \sum_{1 \leq i \leq r} g(i)$ denote the *cumulative gain* [4, 5] at Rank $r$ of the system's output, where $g(i) = gain(\mathcal{L})$ if the document at Rank $i$ is $\mathcal{L}$-relevant and $g(i) = 0$ otherwise. In particular, consider an *ideal* ranked output, such that $isrel(r) = 1$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$: An ideal output for an NTCIR topic simply has all S-, A- and B-relevant documents listed up in this order. We denote the cumulative gain at Rank $r$ for this ideal case by $cg_I(r)$.

We can then evaluate a given system output by comparing it with the ideal one, thereby normalising the evaluation statistic across topics. However, it is known that metrics based on *weighted precision* $WP(r) = cg(r)/cg_I(r)$ cannot properly penalise late arrival of relevant documents below Rank $R$, because the ideal ranked output runs out of relevant documents at Rank $R$ and $cg_I(r)$ becomes a constant after this rank. An example is *Normalised Cumulative Gain* at document cut-off $l$, $nCG_l = WP(l)$: As we mentioned in Section 2, for a topic with $R = 10$ relevant documents, a system that has only one relevant document at Rank 10 and a system that has only one relevant document at Rank 1000 are equally effective according to $nCG_{1000}$ [17].

Normalised Discounted Cumulative Gain (nDCG) partially overcomes this problem by *discounting* the gains according to the ranks of relevant documents. That is, by using $dg(i) = g(i)/\log_a(i)$ instead of $g(i)$ for $i > a$, we obtain the (ideal) *discounted* cumulative gain $dcg(r)$ and $dcg_I(r)$, and compute:

$$nDCG_l = dcg(l)/dcg_I(l) \tag{2}$$

for a given document cut-off $l$. Since Sakai [17] showed that $l$ should be large to ensure high stability and discriminative power, we let $l = L' = 1000$ throughout this paper. The logarithm base $a$ can be interpreted as nDCG's late arrival parameter, as we shall discuss in Section 3.2.

Whereas, Q-measure solves the late arrival problem by replacing the Precision $P(r)$ in Eq. 1 by the *Blended Ratio* $BR(r)$:

$$BR(r) = \frac{cg(r) + count(r)}{cg_I(r) + r} \tag{3}$$

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r)BR(r) \ . \tag{4}$$

Just like $P(r)$, $BR(r)$ has an $r$ in the denominator and is therefore guaranteed to decrease as $r$ increases, i.e., to penalise late arrivals [12, 17].

Using our notations, genAveP, recently proposed by Kishida [10], can be expressed as:

$$genAveP = \frac{\sum_{1 \leq r \leq L} isrel(r)cg(r)/r}{\sum_{1 \leq r \leq R} cg_I(r)/r} \tag{5}$$

where $P'(r) = cg(r)/r$ is known as the *generalised precision* proposed by Kekäläinen [8]. Note that $P'(r)$ also has an $r$ in the denominator and therefore does not share the late arrival problem with nCG. Vu and Gallinari [23] defined a similar metric, but they used $R$ instead of the denominator of the above formula, which causes a normalisation problem.

### 3.2 Advantages of Q-measure over nDCG and genAveP

As mentioned earlier, IR metrics based on graded relevance are required to reward:

(a) Systems that return *highly* relevant documents; and

(b) Systems that return relevant documents *early* in the ranked list.

and to maintain the balance between (a) and (b).

Q-measure, nDCG and genAveP can control the effect of (a) using the *gain value ratio*, $gain(S) : gain(A) : gain(B)$. For example, a "steep" setting such as 10:5:1 rewards retrieval of highly relevant documents heavily. However, the three metrics essentially differ from the viewpoint of (b), as we shall discuss below.

nDCG penalises late arrival of relevant documents by means of discounting: A large logarithm base $a$

respresents a patient user who is quite forgiving for late arrival of relevant documents [5, 9]. However, as Sakai [17] pointed out, a large $a$ makes nDCG inherit the aforementioned defect of nCG, because discounting cannot be applied for Ranks 1 through $a$. For example, if $R = 3$ and $a = 10$, it makes no difference whether a relevant document is at Rank 3 or at Rank 10. Thus nDCG with a large $a$ is a counterintuitive metric.

Whereas, Q-measure controls how severely late arrival of relevant documents should be penalised by using *large* or *small* gain values. To describe this feature more explicitly, we hereafter use an alternative formalisation of the blended ratio [12]:

$$BR(r) = \frac{\beta cg(r) + count(r)}{\beta cg_I(r) + r} \qquad (6)$$

where $\beta$ is the parameter that controls how severely late arrivals should be penalised. Using a large $\beta$ makes $BR(r)$ resemble weighted precision $WP(r)$ and diminishes the effect of $r$ in the denominator, thereby making Q-measure more “forgiving” for late arrivals. Whereas, using a small $\beta$ makes $BR(r)$ resemble precision $P(r)$, and therefore makes Q-measure resemble AveP. Thus, Q-measure’ $\beta$ can control the balance between retrieving a highly relevant document and retrieving any relevant document early in the ranked list, *and* is free from the defect of n(D)CG. Perhaps the downside is that $\beta$ is more difficult to interpret intuitively than the gain values, and must be set empirically.

Unlike Q-measure and nDCG, Kishida’s genAveP lacks a parameter for controlling the penalty on late arrival of relevant documents: Since genAveP relies on $P'(r) = cg(r)/r$, it *assumes* that if a relevant document is retrieved at Rank $r$ instead of Rank 1, the reward should always be reduced to $1/r$ of the original value. This is a property akin to that of Reciprocal Rank.

In summary, only Q-measure and nDCG have a parameter for controlling how severely late arrival of relevant documents should be penalised, but adjusting the parameter $a$ for nDCG makes it a counterintuitive metric. Below, we describe experiments to demonstrate the advantages of Q-measure over nDCG and genAveP from this viewpoint, and to suggest a practical guideline for conducting information retrieval evaluation with Q-measure as the primary metric.

# 4 Data and Methods for Comparing Metrics

## 4.1 NTCIR-5 Data

Our experiments used two different data sets: the Chinese and Japanese test collections with submitted runs from the NTCIR-5 crosslingual task [6]. The statistics of the data are shown in Table 1, where $|Q|$ denotes the number of topics. Recall also that $\sum_{\mathcal{L}} R(\mathcal{L}) = R$ for each topic, where $\mathcal{L}$ is a relevance level. With each data set, we used 30 runs for computing Kendall’s rank correlation values, and 15 runs for conducting the Bootstrap Sensitivity Experiments, as described below.

**Table 1. Statistics of the NTCIR-5 data.**

| | $\|Q\|$ | $R$ | $R(S)$ | $R(A)$ | $R(B)$ | #runs |
|---|---|---|---|---|---|---|
| | | per topic | | | | used |
| Chinese | 50 | 61.0 | 7.0 | 30.7 | 23.3 | 30 (15) |
| Japanese | 47 | 89.1 | 3.2 | 41.8 | 44.2 | 30 (15) |

## 4.2 Kendall’s Rank Correlation

Our first set of experiments computed Kendall’s rank correlation [9, 12, 17, 14, 20] among the system rankings produced by different metrics (with different parameters), to discuss the *resemblance* among metrics. Kendall’s rank correlation computes the minimum number of adjacent swaps to turn one ranking into another: It lies between 1 (identical rankings) and $-1$ (completely reversed rankings), and its expected value is zero for two rankings that are in fact not correlated with each other. For this purpose, we used top 30 runs as measured by AveP from each data set. Given 30 runs, Kendall’s rank correlation is statistically significant at $\alpha = 0.01$ if it is over 0.34 (two-sided normal test) [14].

## 4.3 Bootstrap Sensitivity Method

Our second set of experiments used Sakai’s Bootstrap Sensitivity Method [14, 16] for comparing the *discriminative power* of metrics, that is, for how many pairs of runs a statistically significant difference can be detected given a set of runs submitted to a task. This method can also estimate the overall performance difference required to achieve a statistically significant difference for a given topic set size $|Q|$. For these experiments, we selected the best run in terms of AveP from every participating team for each of our two data sets, which, by coincidence, resulted in 15 unique-team runs for both data sets. We chose to use unique-team runs because we are more interested in detecting a significant difference between two teams than that between a pair of runs submitted by a single team, which could be extremely similar. This also reduces computational cost: with 15 teams, we only have 15*14/=105 run pairs. We generated $B = 1000$ bootstrap samples $Q^{*b}$ by sampling with replacement from the original topic set $Q$ to conduct paired Bootstrap Hypothesis Tests: Due to lack of space, we refer the reader to [14, 16] for the exact algorithm of the Bootstrap Sensitivy Method.

# 5 Results and Discussions

## 5.1 Rank Correlation Results

Figures 1 and 2 visualise Kendall's rank correlations among the system rankings produced by AveP, Q-measure, nDCG and genAveP, for the NTCIR-5 Chinese and Japanese data, respectively. The gain value ratio used is $gain(S) : gain(A) : gain(B) =$ 3:2:1, and the late arrival parameter values used for Q-measure and nDCG are the *default* ones, namely, $\beta = 1$ and $a = 2$. Figures 3 and 4 show similar graphs when the gain value ratio is 10:5:1. Recall that rank correlations lie between $-1$ and 1, and that all the correlation values reported in this paper are over 0.5 and are statistically highly significant. From the four tables, we can observe that:

- Q-measure and genAveP are consistently highly correlated with each other, and are both highly correlated with AveP. But Q-measure is slightly more highly correlated with AveP than genAveP is. This reflects the fact that both AveP and Q-measure use $R$ as the denominator and therefore emphasises recall.

- nDCG is not as highly correlated with AveP as Q-measure and genAveP are. This reflects the fact that nDCG is a *rank-based* (as opposed to *recall-based*) metric: It is more forgiving for low-recall topics [14, 17].

Figures 5-8 show, for each of the aforementioned four cases, the effect of varying Q-measure's $\beta$ on the rank correlation with AveP and with the *default* Q-measure ($\beta = 1$). Similarly, Figures 9-12 show the effect of varying nDCG's $a$ on the rank correlation with AveP and with the *default* nDCG ($a = 2$). It can be observed that:

- The system ranking by nDCG changes dramatically as $a$ increases. When $a = 100$, for example, the correlation with AveP is only around 0.5. This reflects the fact that nDCG with a large $a$ is a counterintuitive metric as we have explained earlier. Thus nDCG with a large $a$ is probably not suitable for practical use.

- In contrast, the system ranking by Q-measure is relatively robust to the change in $\beta$. For example, Q-measure with $\beta = 100$ and that with $\beta = 1000$ are in fact very similar metrics because, as $\beta$ becomes large, $BR(r)$ approaches weighted precision $WP(r)$. Whereas, it can be observed that as $\beta$ approaches zero, Q-measure approaches AveP since the blended ratio $BR(r)$ approaches precision $P(r)$. The results also suggest that $\beta = 0.1, 1, 10$ are reasonable choices for practical use.

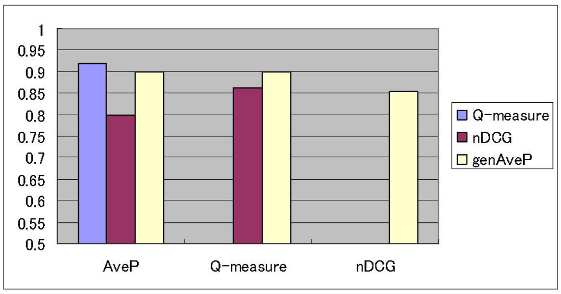

**Figure 1. Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Chinese runs).**

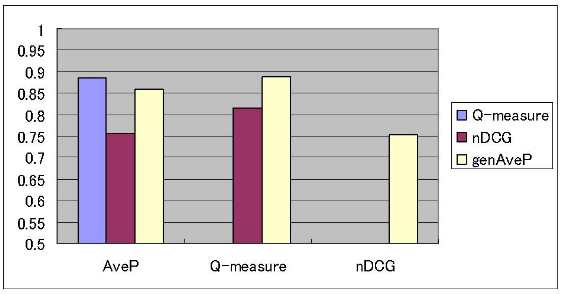

**Figure 2. Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Japanese runs).**

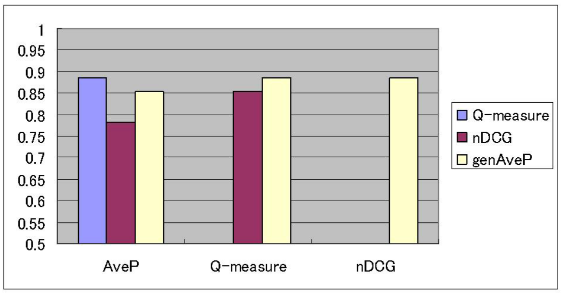

**Figure 3. Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Chinese runs).**

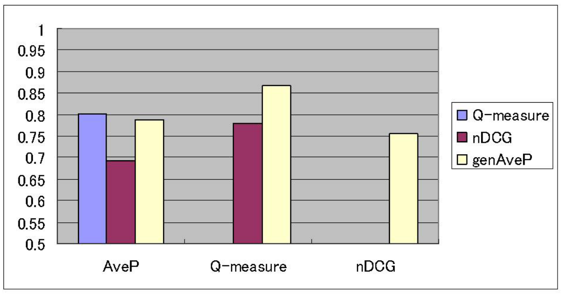

**Figure 4. Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Japanese runs).**

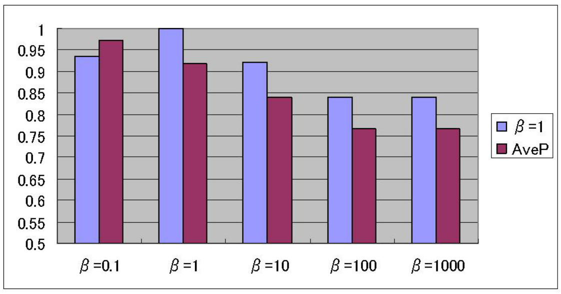

Figure 5. The effect of Q-measure's $\beta$ on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Chinese runs).

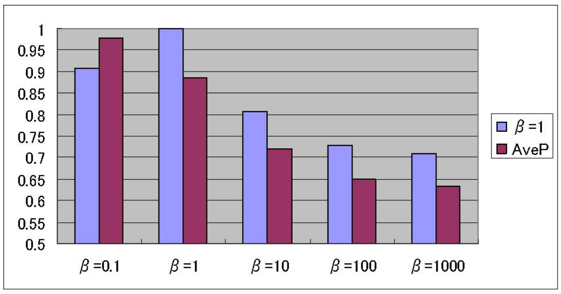

Figure 6. The effect of Q-measure's $\beta$ on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Japanese runs).

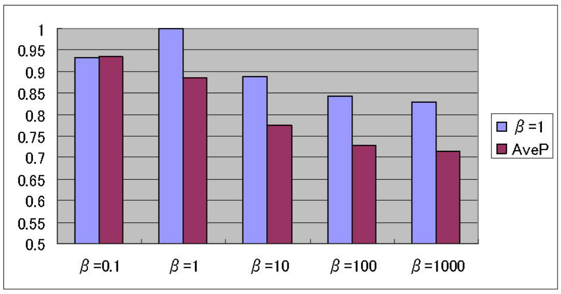

Figure 7. The effect of Q-measure's $\beta$ on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Chinese runs).

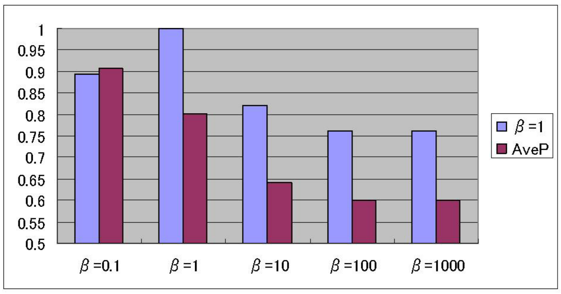

Figure 8. The effect of Q-measure's $\beta$ on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Japanese runs).

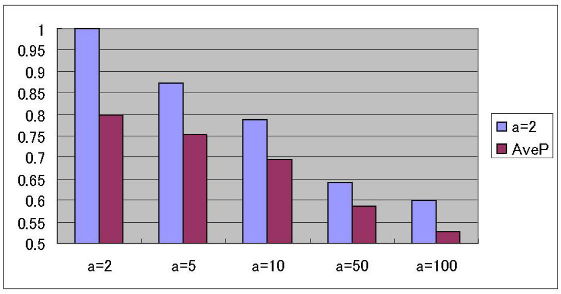

Figure 9. The effect of nDCG's $a$ on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Chinese runs).

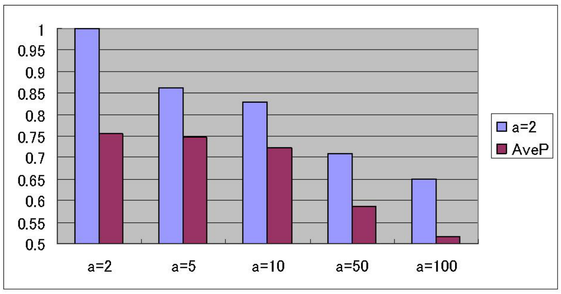

Figure 10. The effect of nDCG's $a$ on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Japanese runs).

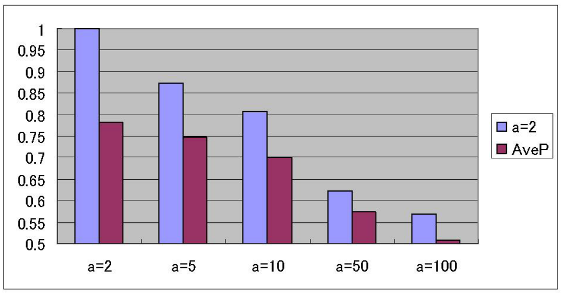

Figure 11. The effect of nDCG's $a$ on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Chinese runs).

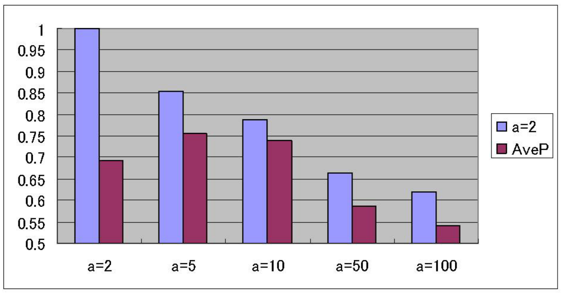

Figure 12. The effect of nDCG's $a$ on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Japanese runs).

### 5.2 Bootstrap Sensitivity Results

Figures 13 and 14 show the *Achieved Significance Level (ASL) curves* for AveP, Q-measure ($\beta = 1$), nDCG ($a = 2$) and genAveP with the gain value ratio 3:2:1 for the 15 unique-team runs from the Chinese and the Japanese data, respectively. Thus, for each of the 15*14/2=105 run pairs, a paired Bootstrap Hypothesis Test using $B = 1000$ bootstrap topic samples was conducted, and the run pairs were sorted by the estimated ASL value [14, 16]. The horizontal axis represents the sorted run pairs, and the vertical axis represents the ASL: Note that low ASL means high significance. For example, Figure 13 shows that, if a significance level of $\alpha = 0.01$ is required, Q-measure is the most discriminative metric: it fails to detect a significant difference for only 37 run pairs out of 105 (35%). Whereas, in the same graph, nDCG fails to detect a significant difference for as many as 46 run pairs (44%).

Based on graphs such as those shown in Figures 13 and 14, Tables 2 and 3 summarise the discriminative power of AveP, Q-measure ($\beta = 0.1, 1, 10, 100, 1000$), nDCG ($a = 2, 5, 10, 50, 100$) and genAveP for $\alpha = 0.05, 0.01$ and the gain value ratios 3:2:1 and 10:5:1. For example, Table 2(b) shows that, when Q-measure with $\beta = 1$ (denoted by Q$\beta = 1$ for short) and gain value ratio 3:2:1 is used for comparing the 15 unique-team Chinese runs, it manages to detect a significant difference at $\alpha = 0.01$ for 68 out of the 105 run pairs (65%). (This is actually the data we discussed at the end of the last paragraph.) The metrics have been sorted by this measure of discriminative power. The same row in the table also shows that, if $|Q| = 50$ topics are used for comparing runs, an overall difference of approximately 0.10 is required in order to detect statistical significance (which is quite large). In each table, metrics that are more discriminative than AveP for all four combinations of $\alpha$ and the gain value ratio are shown in bold. Thus, in Table 2, only Q-measure with $\beta = 0.1, 1, 10$ are consistently more discriminative than AveP; in Table 3, only Q-measure with $\beta = 10$ is consistently more discriminative than AveP.

Figures 15 and 16 visualise the sensitivity columns of Table 2(a)(c) and Table 2(b)(d), respectively. Figures 17 and 18 visualise the sensitivity columns of Table 3(a)(c) and Table 3(b)(d), respectively.

From these tables and figures, we can observe that:

- nDCG loses its discriminative power rather quickly as $a$ becomes large. For example, in Figure 16 and Table 2(d), although the Bootstrap Sensitivity of nDCG$a = 5$ (10:5:1) is 64%, that of nDCG$a = 100$ (10:5:1) is as low as 45%. Thus, nDCG with a large $a$ is not only counter-intuitive, but also insensitive.
- Q-measure does consistently well, even with an extremely large $\beta$. For example, in Figure 16 and Table 2(d), the Bootstrap sensitivity of Q-measure (10:5:1) is over 64% for $\beta = 1, 10, 100, 1000$.
- genAveP also does quite well in terms of discriminative power. However, it does not consistently outperform binary AveP: genAveP is listed below AveP in Table 2(d) and Table 3(a)(c).

### 5.3 Discussions

Our experiments have shown that, even though both Q-measure and nDCG have a parameter for controlling the impact of late arrival of relevant documents (which genAveP lacks), increasing the logarithm base $a$ with nDCG changes the system ranking considerably and seriously hurts discriminative power. To make nDCG practical, one should make sure that $a \leq R$ holds for any topic from the test collection that is being used, where $R$ is the number of relevant documents. For example, the minimum $R$ for the NTCIR-5 Chinese test collection is 4, and that for the Japanese collection is 7. So $a$ should not be larger than these values. This leaves us very little choice in practice.

In contrast, Q-measure's $\beta$ can control the impact of late arrival of relevant documents, while maintaining relatively stable system ranking and high discriminative power. Both the rank correlation and the Bootstrap Sensitivity experiments suggest that $\beta = 0.1, 1, 10$ are reasonable choices for Q-measure. However, Q-measure with an extremely small value of $\beta$ would not be informative if it is to be used along with AveP, since $\beta = 0$ reduces Q-measure to AveP. Thus, in practice, it may be wise to try $\beta = 0, 1, 10$, so that Q-measure subsumes AveP.

We have also demonstrated, for the first time to our knowledge, that genAveP is a reliable metric. However, as we have argued earlier, it is less flexible than Q-measure and nDCG in that it lacks a parameter for controlling the impact of late arrivals in the same way it can control the impact of the relevance levels.

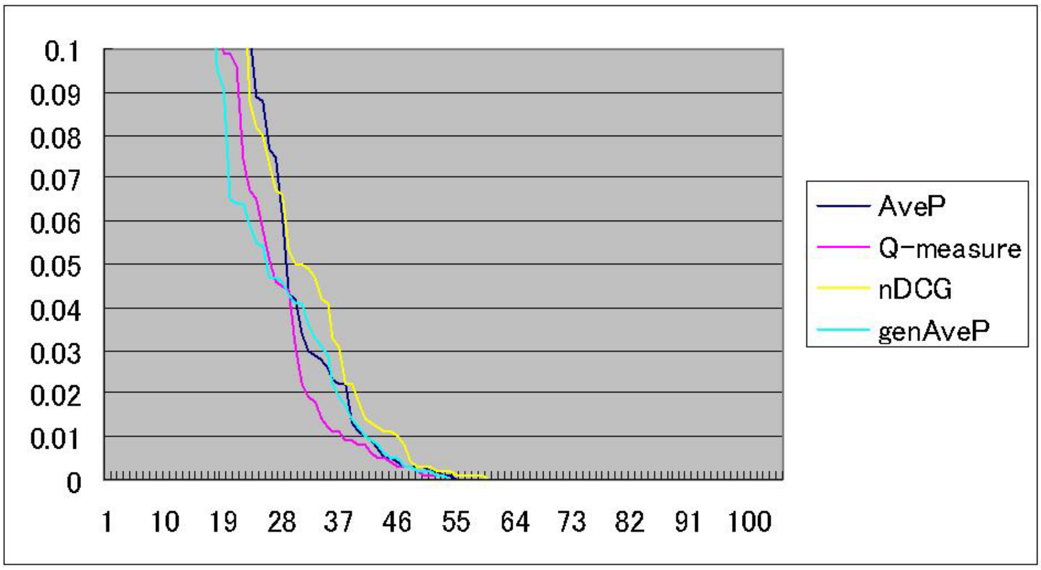

Figure 13. Bootstrap ASL curves: gain value ratio = 3:2:1 (15 unique-team Chinese runs).

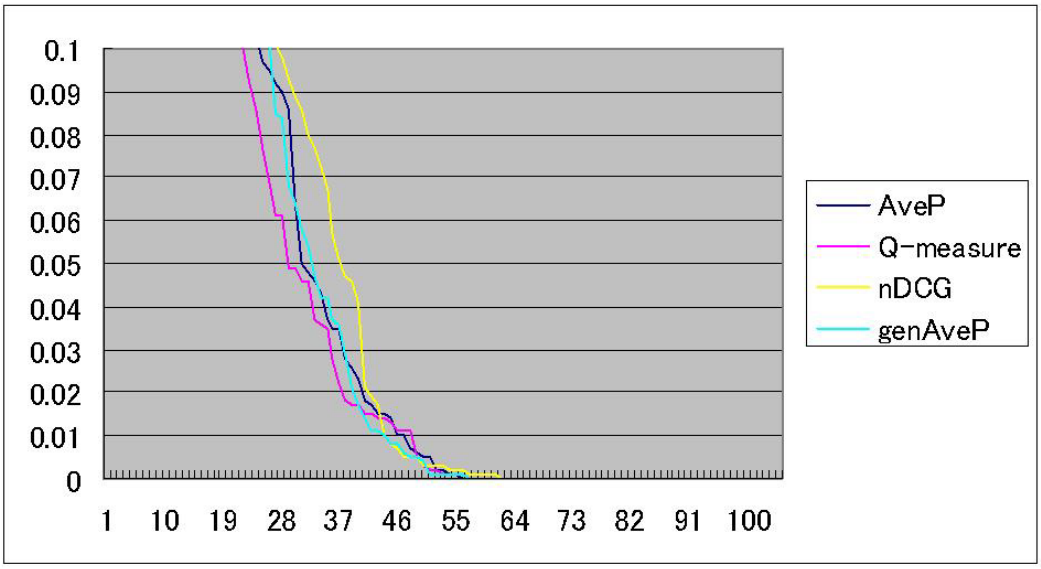

Figure 14. Bootstrap ASL curves: gain value ratio = 3:2:1 (15 unique-team Japanese runs).

**Table 2. Bootstrap sensitivity based on 15 unique-team Chinese runs. Metrics more discriminative than AveP under all four conditions are indicated in bold.**

| metric | sensitivity | estimated diff. |
|---|---|---|
| (a) gain value ratio = 3:2:1, $ASL < \alpha = 0.05$ | | |
| genAveP | 80/105=76% | 0.08 |
| **Q**$\beta = 1$ | **79/105=75%** | **0.10** |
| **Q**$\beta = 10$ | **78/105=74%** | **0.08** |
| **Q**$\beta = 0.1$ | **78/105=74%** | **0.08** |
| nDCG$a = 5$ | 78/105=74% | 0.10 |
| Q$\beta = 100$ | 77/105=73% | 0.08 |
| Q$\beta = 1000$ | 77/105=73% | 0.08 |
| *AveP* | *77/105=73%* | *0.09* |
| nDCG$a = 10$ | 74/105=70% | 0.09 |
| nDCG$a = 2$ | 74/105=70% | 0.09 |
| nDCG$a = 50$ | 73/105=70% | 0.10 |
| nDCG$a = 100$ | 71/105=68% | 0.11 |
| (b) gain value ratio = 3:2:1, $ASL < \alpha = 0.01$ | | |
| **Q**$\beta = 1$ | **68/105=65%** | **0.10** |
| **Q**$\beta = 10$ | **67/105=64%** | **0.10** |
| Q$\beta = 100$ | 67/105=64% | 0.12 |
| Q$\beta = 1000$ | 67/105=64% | 0.12 |
| **Q**$\beta = 0.1$ | **65/105=62%** | **0.11** |
| genAveP | 64/105=61% | 0.10 |
| *AveP* | *64/105=61%* | *0.11* |
| nDCG$a = 5$ | 62/105=59% | 0.12 |
| nDCG$a = 10$ | 60/105=57% | 0.12 |
| nDCG$a = 2$ | 59/105=56% | 0.11 |
| nDCG$a = 50$ | 54/105=51% | 0.12 |
| nDCG$a = 100$ | 49/105=47% | 0.13 |
| (c) gain value ratio = 10:5:1, $ASL < \alpha = 0.05$ | | |
| genAveP | 81/105=77% | 0.07 |
| **Q**$\beta = 1$ | **80/105=76%** | **0.08** |
| **Q**$\beta = 10$ | **79/105=75%** | **0.09** |
| **Q**$\beta = 0.1$ | **79/105=75%** | **0.09** |
| Q$\beta = 100$ | 78/105=74% | 0.08 |
| nDCG$a = 5$ | 78/105=74% | 0.09 |
| nDCG$a = 10$ | 78/105=74% | 0.11 |
| nDCG$a = 2$ | 77/105=73% | 0.09 |
| *AveP* | *77/105=73%* | *0.09* |
| Q$\beta = 1000$ | 76/105=72% | 0.09 |
| nDCG$a = 50$ | 72/105=69% | 0.10 |
| nDCG$a = 100$ | 72/105=69% | 0.12 |
| (d) gain value ratio = 10:5:1, $ASL < \alpha = 0.01$ | | |
| **Q**$\beta = 10$ | **69/105=66%** | **0.11** |
| **Q**$\beta = 1$ | **68/105=65%** | **0.12** |
| Q$\beta = 100$ | 67/105=64% | 0.10 |
| Q$\beta = 1000$ | 67/105=64% | 0.12 |
| nDCG$a = 5$ | 67/105=64% | 0.12 |
| **Q**$\beta = 0.1$ | **65/105=62%** | **0.10** |
| *AveP* | *64/105=61%* | *0.11* |
| nDCG$a = 2$ | 63/105=60% | 0.13 |
| nDCG$a = 10$ | 63/105=60% | 0.12 |
| genAveP | 63/105=60% | 0.10 |
| nDCG$a = 50$ | 55/105=52% | 0.15 |
| nDCG$a = 100$ | 47/105=45% | 0.16 |

**Table 3. Bootstrap sensitivity based on 15 unique-team Japanese runs. Metrics more discriminative than AveP under all four conditions are indicated in bold.**

| metric | sensitivity | estimated diff. |
|---|---|---|
| (a) gain value ratio = 3:2:1, $ASL < \alpha = 0.05$ | | |
| **Q**$\beta = 10$ | **78/105=74%** | **0.09** |
| Q$\beta = 1$ | 77/105=73% | 0.08 |
| Q$\beta = 0.1$ | 74/105=70% | 0.08 |
| nDCG$a = 5$ | 74/105=70% | 0.10 |
| *AveP* | *74/105=70%* | *0.08* |
| Q$\beta = 100$ | 73/105=70% | 0.09 |
| Q$\beta = 1000$ | 73/105=70% | 0.09 |
| genAveP | 73/105=70% | 0.08 |
| nDCG$a = 10$ | 72/105=69% | 0.09 |
| nDCG$a = 2$ | 68/105=65% | 0.09 |
| nDCG$a = 50$ | 68/105=65% | 0.11 |
| nDCG$a = 100$ | 65/105=62% | 0.11 |
| (b) gain value ratio = 3:2:1, $ASL < \alpha = 0.01$ | | |
| Q$\beta = 100$ | 61/105=58% | 0.11 |
| Q$\beta = 1000$ | 61/105=58% | 0.14 |
| nDCG$a = 2$ | 61/105=58% | 0.12 |
| genAveP | 61/105=58% | 0.09 |
| **Q**$\beta = 10$ | **60/105=57%** | **0.11** |
| nDCG$a = 10$ | 59/105=56% | 0.14 |
| Q$\beta = 0.1$ | 58/105=55% | 0.12 |
| nDCG$a = 5$ | 58/105=55% | 0.12 |
| *AveP* | *58/105=55%* | *0.12* |
| Q$\beta = 1$ | 57/105=54% | 0.11 |
| nDCG$a = 50$ | 56/105=53% | 0.14 |
| nDCG$a = 100$ | 53/105=50% | 0.14 |
| (c) gain value ratio = 10:5:1, $ASL < \alpha = 0.05$ | | |
| Q$\beta = 1$ | 77/105=73% | 0.08 |
| **Q**$\beta = 10$ | **74/105=70%** | **0.09** |
| Q$\beta = 100$ | 74/105=70% | 0.09 |
| Q$\beta = 1000$ | 74/105=70% | 0.09 |
| nDCG$a = 10$ | 74/105=70% | 0.09 |
| *AveP* | *74/105=70%* | *0.08* |
| Q$\beta = 0.1$ | 73/105=70% | 0.09 |
| nDCG$a = 2$ | 73/105=70% | 0.09 |
| genAveP | 72/105=69% | 0.08 |
| nDCG$a = 5$ | 70/105=67% | 0.10 |
| nDCG$a = 50$ | 66/105=63% | 0.11 |
| nDCG$a = 100$ | 66/105=63% | 0.12 |
| (d) gain value ratio = 10:5:1, $ASL < \alpha = 0.01$ | | |
| nDCG$a = 2$ | 65/105=62% | 0.11 |
| **Q**$\beta = 10$ | **64/105=61%** | **0.12** |
| Q$\beta = 1000$ | 64/105=61% | 0.12 |
| genAveP | 64/105=61% | 0.10 |
| Q$\beta = 100$ | 63/105=60% | 0.12 |
| Q$\beta = 1$ | 60/105=57% | 0.12 |
| Q$\beta = 0.1$ | 58/105=55% | 0.12 |
| nDCG$a = 5$ | 58/105=55% | 0.12 |
| nDCG$a = 10$ | 58/105=55% | 0.13 |
| *AveP* | *58/105=55%* | *0.12* |
| nDCG$a = 50$ | 55/105=52% | 0.14 |
| nDCG$a = 100$ | 53/105=50% | 0.13 |

## 6 Conclusions

This paper discussed and demonstrated the advantages of Q-measure over nDCG and genAveP in terms of the ability to control how severely late arrival of relevant documents should be penalised in information retrieval evaluation. Our discussions and experimental findings can be summarised as follows:

- Both Q-measure and nDCG have a parameter for controlling how severely late arrival of relevant documents should be penalised. genAveP lacks this capability: if a relevant document is retrieved at Rank $r$ instead of Rank 1, the reward is always reduced to $1/r$ of the original value.
- Although nDCG can control how to penalise late arrival by adjusting the logarithm base $a$, using a large $a$ makes it inherit the defect of nCG and become a counterintuitive metric. Moreover, if $a$ is increased, the system ranking is affected substantially, and the metric loses its discriminative power quickly.
- Q-measure is free from the defect of n(D)CG. Moreover, Q-measure is relatively robust to the choice of the "late arrival" parameter $\beta$, both in terms of system ranking and in terms of discriminative power. $\beta = 1, 10$ are probably good choices.

In short, although Q-measure, nDCG and genAveP are highly correlated with one another and are all generally reliable, Q-measure is probably the most flexible graded-relevance metric.

So how should one conduct IR experiments using graded relevance? This paper, along with previous studies [14, 17], provides grounds for us to claim that Q-measure deserves to be the primary metric. The gain value ratio may be set intuitively, say 3:2:1 or 10:5:1, since Q-measure is known to be fairly robust to the choice. (Sakai [17] discusses how the gain value ratio can optionally be adjusted *per topic*.) Alternatively, if the relevance levels are defined based on the *amount* or *proportion* of relevant content in each document, it may be possible to set the ratio so that it approximates these actual statistics. Then, a few values of $\beta$ could be tried, say, $\beta = 1, 10$. As mentioned earlier, conservative researchers may also want to try $\beta = 0$, to reduce Q-measure to binary AveP. Moreover, since Q-measure is recall-based, one may additionally use the rank-based nDCG with a small logarithm base $a$: We recommend $a = 2$.

The above practice yields several summary statistics for a single system, which is good: It is always useful to evaluate systems from several different angles. It is useful to observe trends that hold across different metrics, and also to examine phenomena that occur with a particular metric only.

An open question is, how should the parameters such as the gain value ratio and $\beta$ be set so that the metrics correlate well with *user satisfaction*? This is certainly a difficult one to answer, but it should not be used as an excuse for not using graded-relevance metrics: At any rate, it is unlikely that a binary relevance metric does any better in terms of user satisfaction. We believe that *in vitro* experiments using graded relevance are useful for building effective information retrieval systems efficiently, even if they must eventually be "rerun" *in vivo* somehow. Whether criticisms of AveP from the viewpoint of user satisfaction (e.g. [19]) apply to different IR environments and different IR metrics including Q-measure needs to be investigated also.

## References

[1] Buckley, C. and Voorhees, E. M.: Evaluating Evaluation Measure Stability, *ACM SIGIR 2000 Proceedings*, pp. 33-40, 2000.

[2] Buckley, C. and Voorhees. E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp. 25-32, 2006.

[3] Hersh, W. *et al.*: TREC 2005 Genomics Track Overview, *TREC 2005 Proceedings*, 2006.

[4] Järvelin, K. and Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents, *ACM SIGIR Proceedings*, pp. 41-48, 2000.

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

[6] Kando, N.: Overview of the Fifth NTCIR Workshop, *NTCIR-5 Proceedings*, 2005.

[7] Kazai, G., and Lalmas, M.: eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval, *ACM Transactions on Information Systems*, Vol. 24, Issue 4, pp. 503-542, 2006.

[8] Kekäläinen, J. and Järvelin, K.: Using Graded Relevance Assessments in IR Evaluation, *Journal of the American Society for Information Science and Technology*, Vol 53, No. 13, pp. 1120-1129, 2002.

[9] Kekäläinen, J.: Binary and Graded Relevance in IR evaluations - Comparison of the Effects on Ranking of IR Systems, *Information Processing and Management*, Vol. 41, pp. 1019-1033, 2005.

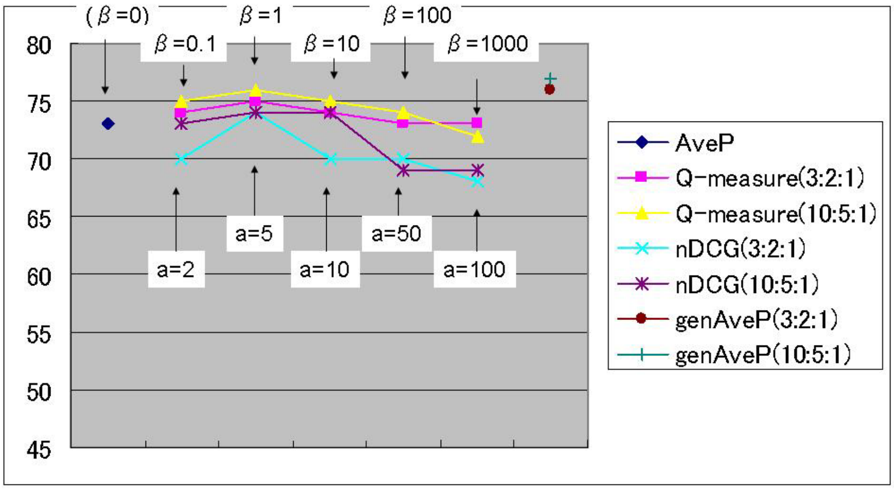

**Figure 15. The effect of Q-measure's $\beta$ and nDCG's $a$ on Bootstrap Sensitivity at $\alpha = 0.05$ (NTCIR-5 Chinese data).**

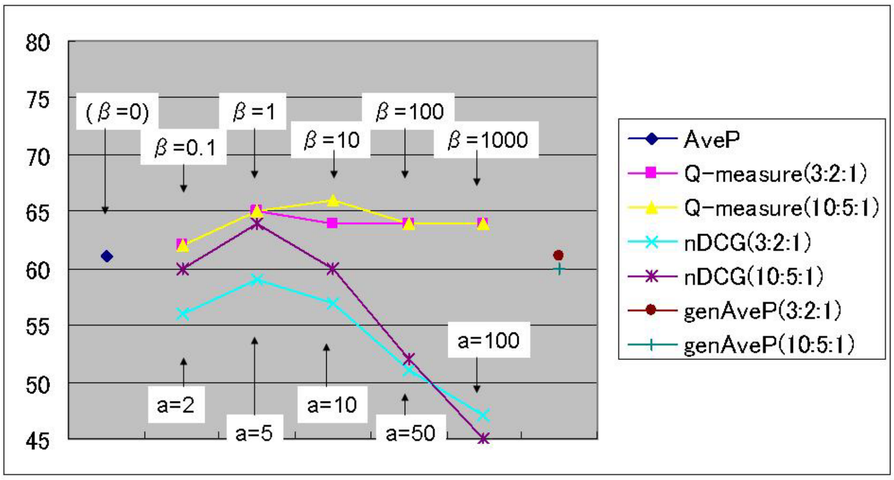

**Figure 16. The effect of Q-measure's $\beta$ and nDCG's $a$ on Bootstrap Sensitivity at $\alpha = 0.01$ (NTCIR-5 Chinese data).**

[10] Kishida, K.: Property of Average Precision and its Generalization: An Examination of Evaluation Indicator for Information Retrieval Experiments, *National Institute of Informatics Technical Report*, NII-2005-014E, 2005.

[11] Oyama, K. *et al.*: Overview of the NTCIR-5 WEB Navigational Retrieval Subtask2 (Navi-2), *NTCIR-5 Proceedings*, 2005.

[12] Sakai, T.: New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering, *NTCIR-4 Proceedings*, 2004.

[13] Sakai, T.: On the Task of Finding One Highly Relevant Document with High Precision, *IPSJ Transactions on Databases*, Vol. 47, No. SIG 4 (TOD29), pp. 13-27, 2006. Also available in *IPSJ Digital Courier*, Vol. 2, pp. 174-188, 2006.

[14] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR 2006 Proceedings*, pp. 525-532, 2006.

[15] Sakai, T.: Controlling the Penalty on Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *IPSJ SIG Technical Reports*, 2006-FI-84/2006-NL-175, pp. 57-64, 2006.

[16] Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, *AIRS 2006 Proceedings*, LNCS 4182, pp. 374-389, 2006.

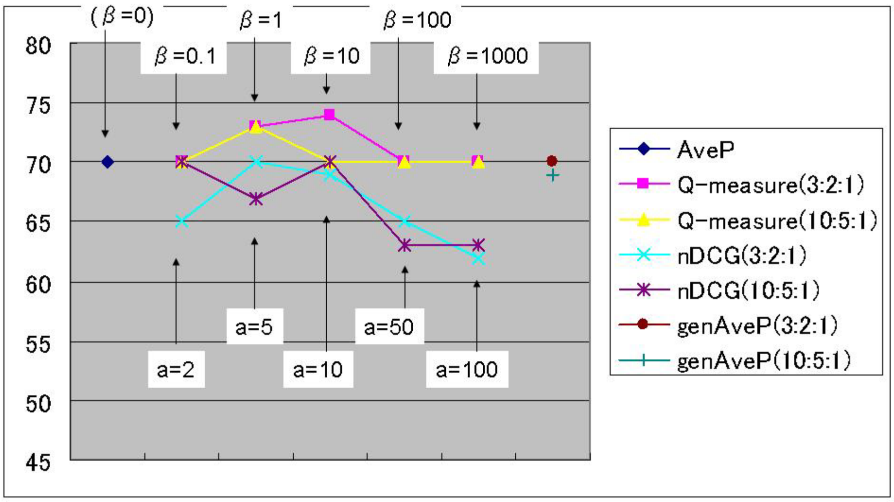

**Figure 17. The effect of Q-measure's $\beta$ and nDCG's $a$ on Bootstrap Sensitivity at $\alpha = 0.05$ (NTCIR-5 Japanese data).**

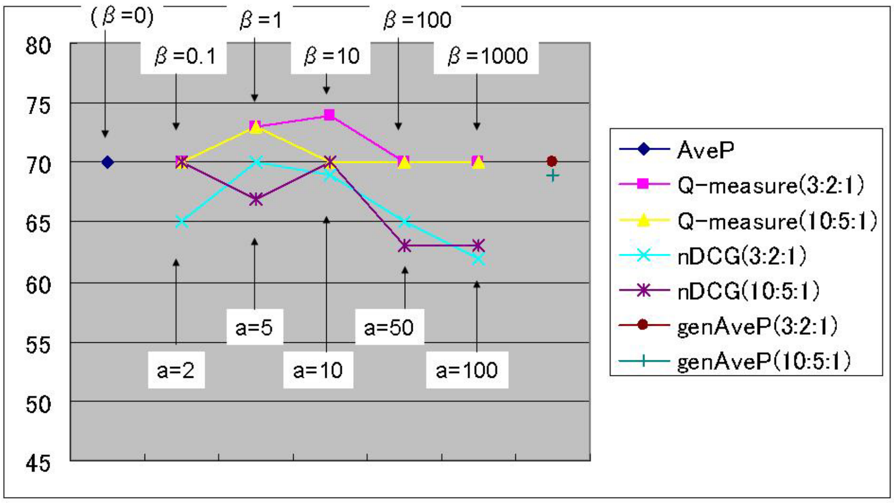

**Figure 18. The effect of Q-measure's $\beta$ and nDCG's $a$ on Bootstrap Sensitivity at $\alpha = 0.01$ (NTCIR-5 Japanese data).**

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

[18] Soboroff, I.: Dynamic Test Collections: Measuring Search Effectiveness on the Live Web, *ACM SIGIR 2006 Proceedings*, pp. 276-283, 2006.

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

[20] Voorhees, E. M.: Evaluation by Highly Relevant Documents, *ACM SIGIR 2001 Proceedings*, pp. 74-82, 2001.

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

[22] Voorhees, E. M.: Overview of the TREC 2005 Robust Retrieval Track, *TREC 2005 Proceedings*, 2006.

[23] Vu, H.-T. and Gallinari, P.: On Effectiveness Measures and Relevance Functions in Ranking INEX Systems, *AIRS 2005 Proceedings*, LNCS 3689, pp. 312-327, 2005.