社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (9)
2006－ＮＬ－175 (9)
2006／9／12

Controlling the Penalty on Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance

Tetsuya Sakai

Knowledge Media Laboratory, Toshiba Corporate R&D Center
tetsuya.sakai@toshiba.co.jp

Abstract

Large-scale information retrieval evaluation efforts such as TREC and NTCIR have always used binary-relevance evaluation metrics, even when graded relevance data were available. However, the NTCIR-6 crosslingual task has finally announced that it will use graded-relevance metrics, though only as additional metrics. This paper compares graded-relevance metrics in terms of the ability to control the balance between retrieving highly relevant documents and retrieving any relevant documents early in the ranked list. 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.

1 Introduction

The Information Retrieval (IR) tasks at NTCIR [7] have always used IR evaluation metrics based on *binary* relevance, most notably *Average Precision* (AveP), even though they have test collections with *graded* relevance assessments. The only exception is the NTCIR Web track [12] which is now discontinued but used the *unnormalised* Discounted Cumlative Gain (DCG) metric proposed by Järvelin and Kekäläinen [5][1]. However, it is known that unnormalised DCG takes arbitrarily large values for topics with many relevant documents, and are not suitable for averaging across topics [14, 20]. The situations are similar outside Asia: For example, the Robust Track and the Genomics track at TREC 2005 [4, 23] used binary AveP and their variants, thereby failing to exploit the relevance levels they have. 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 [6] 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 [13] also proposed an averageable graded-relevance metric called *Q-measure* which is very highly correlated with AveP, and claimed that it deserves to be used as an official metric. He also showed that Q-measure is at least as stable and sensitive as AveP [17, 20]. 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 are nDCG, Q-measure and *generalised Average Precision* (genAveP) recently proposed by Kishida [10].

The objective of this paper is to discuss and demonstrate the advantages of Q-measure over nDCG and genAveP from the viewpoint of *flexibility*, by which we mean the ability to control the balance between retrieving highly relevant documents and retrieving any relevant documents early in the ranked list. Through experiments using the Chinese/Japanese test collections and submitted runs from NTCIR-5, we show that Q-measure's parameter β can determine how severely late arrival of relevant documents should be penalised, while maintaining reliable system ranking and evaluation sensitivity. We then 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.

[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 [14, 16, 19] for details.

2 Related Work

All graded-relevance metrics considered in this paper are based on *Cumulative Gain* proposed at ACM SIGIR 2000 [5]. Järvelin and Kekäläinen proposed normalised Cumulative Gain (nCG) and normalised Discounted Cumulative Gain (nDCG) in 2002 [6]. However, Sakai [14, 20] pointed out that nCG, which in fact is almost identical to *sliding ratio* proposed back in 1960s [11], has a defect: It cannot penalise late arrival of relevant documents properly. In fact, even nDCG inherits this defect under some circumstances, as we shall explain in Section 3. They also proposed *Average* n(D)CG by taking the average across document ranks, but straight nDCG appears to be the most popular among their metrics: Kekäläinen [9] compared nDCG with Precision in terms of Kendall's rank correlation; Asakawa and Selberg [1] reported that Microsoft uses nDCG for tuning their Web search engine.

Sakai [20] compared the stability and sensitivity of graded-relevance metrics such as Q-measure and nDCG using the *stability* method proposed at SIGIR 2000 [2] and the *swap* method proposed at SIGIR 2002 [22], as well as Kendall's rank correlation. He showed that both Q-measure and nDCG are stable and sensitive, but that a large document cut-off should be used with nDCG. He also discussed why Average Distance Measure [3], proposed for evaluation with *continuous* (as op-

posed to graded) relevance, is not suitable for traditional document ranking tasks. At SIGIR 2006, Sakai [17] proposed a sensitivity comparison method that is less *ad hoc* than the stability and the swap methods. This method relies on Bootstrap Hypothesis Tests and yields results that are similar to those based on the *ad hoc* methods. This paper therefore uses this method for comparing the sensitivity of Q-measure, nDCG and genAveP with different parameter settings.

Kishida [10] proposed genAveP and compared it with Q-measure and Average nDCG using a small-scale, artificial data set. His work did not use real data, and did not discuss the stability and sensitivity of genAveP. Vu and Gallinari [24] also proposed a generalised version of AveP and compared it with Q-measure for an XML retrieval task, but their metric does not average well (See Section 3). They tried a few values for Q-measure's $\beta(= 0.1, 1, 10)$, the late arrival parameter which we will explain in Section 3, and reported that the choice affects both system ranking and sensitivity for the XML task.

This paper discusses Q-measure, nDCG, genAveP and AveP, all of which are IR metrics for the task of finding *as many relevant documents as possible*. However, there are other kinds of IR tasks: Sakai [15, 16, 18] examined the resemblance, stability and sensitivity of IR metrics for the task of finding *one highly relevant document only*, namely, *P-measure*, *O-measure* and *Weighted Reciprocal Rank*.

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. Let $count(r)$ denote the number of relevant documents within top r ($\leq L$), and let $isrel(r)$ be 1 if the document at Rank r is relevant and 0 otherwise. Clearly, precision at Rank r is given by $P(r) = count(r)/r$. Hence:

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

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 [6]. For example, for an NTCIR test collection which has S-, A- and B-relevant (highly relevant, relevant and partially relevant) documents, we can let $gain(S) = 3, gain(A) = 2, gain(B) = 1$. Let $cg(r) = \sum_{1 \leq i \leq r} g(i)$ denote the *cumulative gain* [6] 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 $r > 1$: For an NTCIR topic, it 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 $cg_I(r)$. Then, by using $dg(i) = g(i)/\log_a(i)$ instead of $g(i)$ for $i > a$, we can obtain the (ideal) *discounted* cumulative gain $dcg(r)$ and $dcg_I(r)$. nDCG is defined as:

$$nDCG_l = dcg(l)/dcg_I(l) \quad (2)$$

where $l (\leq L')$ is a document cut-off. Since Sakai [20] showed that l should be large to ensure stability and sensitivity, we let $l = L' = 1000$ throughout this paper.

Q-measure is defined as:

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

where $BR(r)$ is the *blended ratio* given by:

$$BR(r) = \frac{cg(r) + count(r)}{cg_I(r) + r} . \quad (4)$$

It is known that metrics based on *weighted precision* $WP(r) = cg(r)/cg_I(r)$, such as $nCG_l = WP(l)$, 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. nDCG partially overcomes this problem by discounting the gains, while Q-measure overcomes it by using the blended ratio $BR(r)$ instead, which includes r in the denominator and therefore decreases as r increases [13, 14, 20].

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

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

where $P'(r) = cg(r)/r$ is known as the *generalised precision* proposed by Kekäläinen [8]. Vu and Gallinari [24] defined a similar metric, but they used R for the denominator, which causes a normalisation problem.

3.2 Advantages of Q-measure over nDCG and genAveP

Graded-relevance IR metrics 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" ratio 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 [6, 9]. However, as Sakai [20] 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 [13]:

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

where β is the parameter that controls how severely late arrivals should be penalised. Using a large β makes $BR(r)$

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)

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 β makes $BR(r)$ resemble precision $P(r)$, and therefore makes Q-measure resemble AveP. Thus, Q-measure' β 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 β is more difficult to interpret intuitively than the gain value ratio, 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 generalised precision, 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.

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 α 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 Experimental Methods

Our experiments used two data sets: the Chinese/Japanese test collections and submitted runs from the NTCIR-5 crosslingual task [7]. The statistics of the data are shown in Table 1, where $|Q|$ denotes the number of topics.

Our first set of experiments computed Kendall's rank correlation [9, 13, 20, 17, 21] among the system rankings produced by different metrics (with different parameters), to discuss the *resemblance* among metrics. 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 test) [17].

Our second set of experiments used Sakai's method based on paired Bootstrap Hypothesis Tests [17] for comparing the *sensitivity* of metrics, that is, for how many pairs of runs statistically significant difference can be detected. This method can also estimate the overall performance difference required to achieve a statistically significant difference for a given topic set size $c = |Q|$. 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. Due to lack of space, we refer the reader to [17] for details on Sakai's method.

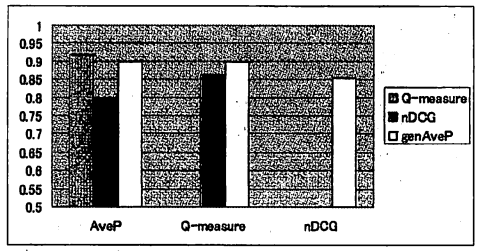

Figure 1. Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Chinese runs).

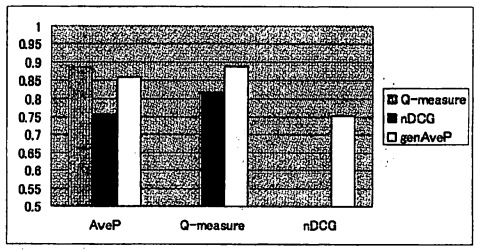

Figure 2. Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Japanese runs).

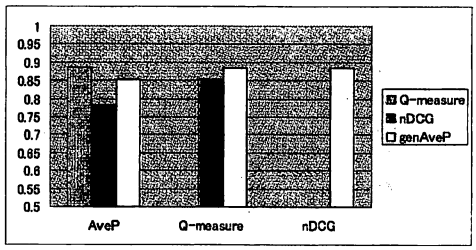

Figure 3. Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Chinese runs).

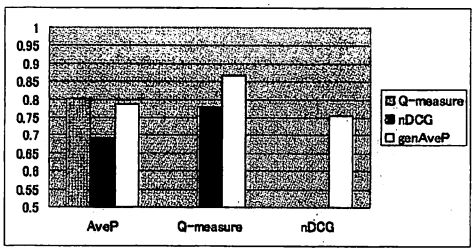

Figure 4. Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Japanese runs).

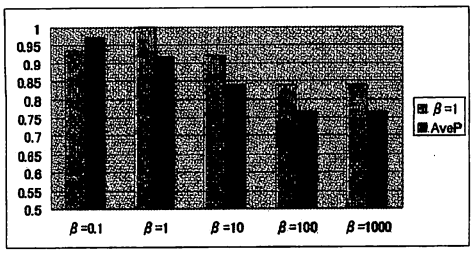

Figure 5. The effect of Q-measure's β on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Chinese runs).

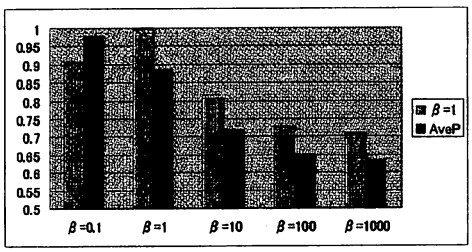

Figure 6. The effect of Q-measure's β on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Japanese runs).

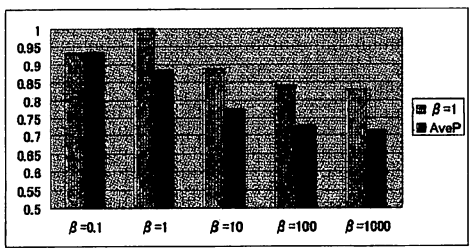

Figure 7. The effect of Q-measure's β on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Chinese runs).

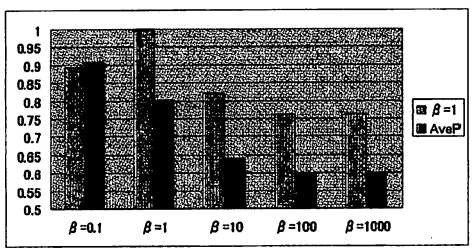

Figure 8. The effect of Q-measure's β on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Japanese runs).

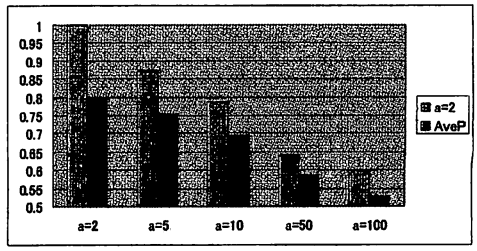

Figure 9. The effect of nDCG's a on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Chinese runs).

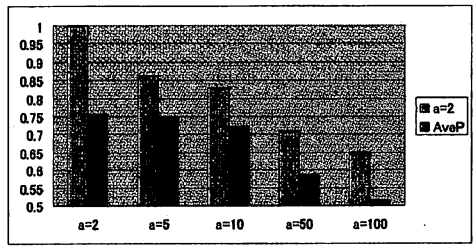

Figure 10. The effect of nDCG's a on Kendall's rank correlation: gain value ratio = 3:2:1 (top 30 Japanese runs).

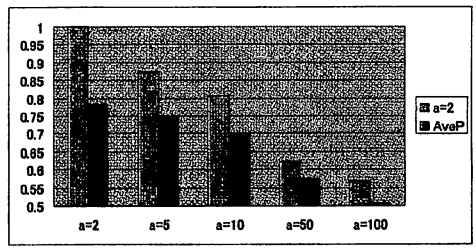

Figure 11. The effect of nDCG's a on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Chinese runs).

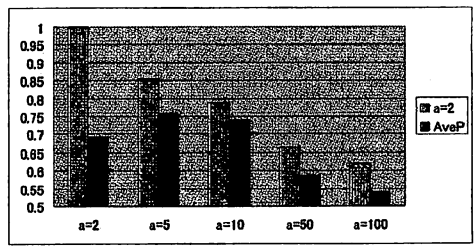

Figure 12. The effect of nDCG's a on Kendall's rank correlation: gain value ratio = 10:5:1 (top 30 Japanese runs).

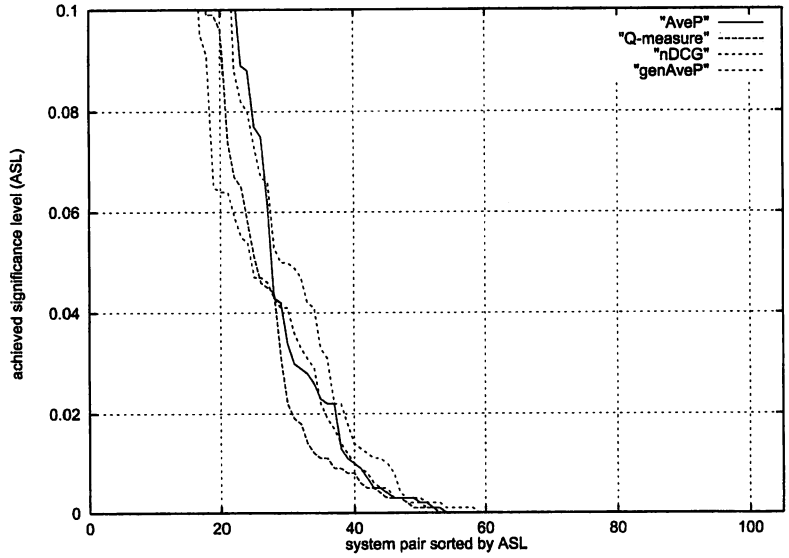

Figure 13. Bootstrap ASL curves: gain value ratio = 3:2:1 (15 unique-team Chinese runs).

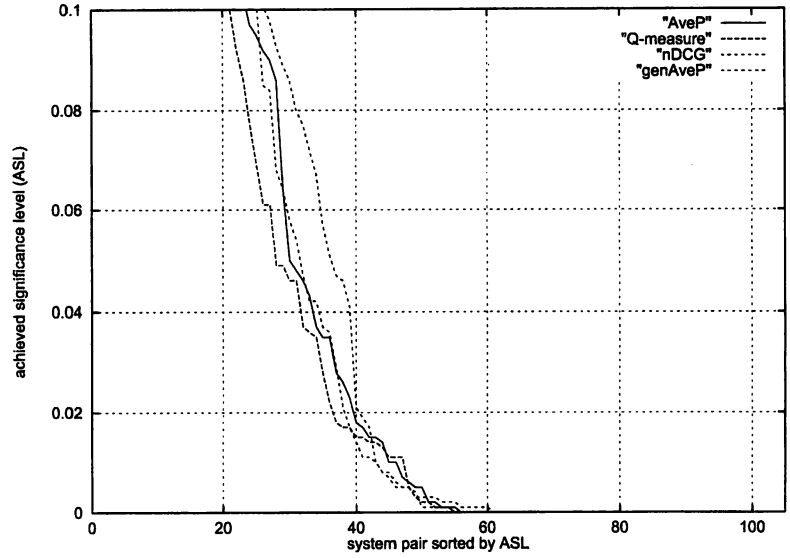

Figure 14. Bootstrap ASL curves: gain value ratio = 3:2:1 (15 unique-team Japanese runs).

Table 2. Bootstrap sensitivity based on 15 unique-team Chinese runs. Metrics more sensitive 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
Qβ = 1	**79/105=75%**	**0.10**
Qβ = 10	**78/105=74%**	**0.08**
Qβ = 0.1	**78/105=74%**	**0.08**
nDCGa = 5	78/105=74%	0.10
Qβ = 100	77/105=73%	0.08
Qβ = 1000	77/105=73%	0.08
AveP	*77/105=73%*	*0.09*
nDCGa = 10	74/105=70%	0.09
nDCGa = 2	74/105=70%	0.09
nDCGa = 50	73/105=70%	0.10
nDCGa = 100	71/105=68%	0.11
(b) gain value ratio = 3:2:1, $ASL < \alpha = 0.01$		
Qβ = 1	**68/105=65%**	**0.10**
Qβ = 10	**67/105=64%**	**0.10**
Qβ = 100	67/105=64%	0.12
Qβ = 1000	67/105=64%	0.12
Qβ = 0.1	**65/105=62%**	**0.11**
genAveP	64/105=61%	0.10
AveP	*64/105=61%*	*0.11*
nDCGa = 5	62/105=59%	0.12
nDCGa = 10	60/105=57%	0.12
nDCGa = 2	59/105=56%	0.11
nDCGa = 50	54/105=51%	0.12
nDCGa = 100	49/105=47%	0.13
(c) gain value ratio = 10:5:1, $ASL < \alpha = 0.05$		
genAveP	81/105=77%	0.07
Qβ = 1	**80/105=76%**	**0.08**
Qβ = 10	**79/105=75%**	**0.09**
Qβ = 0.1	**79/105=75%**	**0.09**
Qβ = 100	78/105=74%	0.08
nDCGa = 5	78/105=74%	0.09
nDCGa = 10	78/105=74%	0.11
nDCGa = 2	77/105=73%	0.09
AveP	*77/105=73%*	*0.09*
Qβ = 1000	76/105=72%	0.09
nDCGa = 50	72/105=69%	0.10
nDCGa = 100	72/105=69%	0.12
(d) gain value ratio = 10:5:1, $ASL < \alpha = 0.01$		
Qβ = 10	**69/105=66%**	**0.11**
Qβ = 1	**68/105=65%**	**0.12**
Qβ = 100	67/105=64%	0.10
Qβ = 1000	67/105=64%	0.12
nDCGa = 5	67/105=64%	0.12
Qβ = 0.1	**65/105=62%**	**0.10**
AveP	*64/105=61%*	*0.11*
nDCGa = 2	63/105=60%	0.13
nDCGa = 10	63/105=60%	0.12
genAveP	63/105=60%	0.10
nDCGa = 50	55/105=52%	0.15
nDCGa = 100	47/105=45%	0.16

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

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

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. 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. Note 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 probably 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 [17, 20].

Figures 5-8 show, for each of the aforementioned four cases, the effect of varying Q-measure's β 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 β. For example, Q-measure with $\beta = 100$ and that with $\beta = 1000$ are very similar metrics because, as β becomes large, $BR(r)$ approaches weighted precision $WP(r)$. Whereas, it can be observed that as β 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.

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 1000 bootstrap topic samples was conducted, and the run pairs were sorted by the estimated ASL value [17]. For example, Figure 13 shows that, if a significance level of around $\alpha = 0.01$ is required, Q-measure is clearly the most sensitive metric: it fails to detect a significant difference for only 37 run pairs out of 105 (35%).

Based on graphs such as those shown in Figures 13 and 14, Tables 2 and 3 summarise the sensitivity 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 the data we already discussed in the last paragraph. The metrics have been sorted by this measure of sensitivity. 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 sensitive than AveP for all four combinations of α and the gain value ratio are shown in bold. We can observe that:

- nDCG loses its sensitivity rather quickly as a becomes large. Thus nDCG with a large a is not only counterintuitive, but also insensitive.
- Q-measure does consistently well, even with an extremely large β. In Table 2, Q-measure with $\beta = 1, 10$ appear to be excellent choices. In Table 3, Q-measure with $\beta = 10$ is the overall winner. genAveP also does relatively well in terms of sensitivity.

5.3 Discussions

From our rank correlation and sensitivity results, it is clear that Q-measure with $\beta = 1, 10$ are good choices for information retrieval evaluation with graded relevance. Note also that $\beta = 0$ can be tried to reduce Q-measure to AveP. nDCG with a small a is good, but one should make sure that $a \leq R$ holds for any topic from the test collection that is being used in order to avoid the defect of n(D)CG. 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. We have also shown, probably for the first time, that genAveP is a reliable metric. However, as we have argued earlier, it is less flexible than Q-measure and nDCG.

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 sensitivity rather 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 β, both in terms of system ranking and in terms of sensitivity. For the NTCIR-5 Chinese and Japanese data, $\beta = 1, 10$ are good choices.

Thus, although Q-measure, nDCG and genAveP are highly correlated with one another, Q-measure is probably the most flexible graded-relevance metric.

So how should one conduct IR experiments using graded relevance? This paper 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 [20] 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, the ratio may be set to approximate these actual statistics. Then, a few values of β could be tried, say, $\beta = 1, 10$. 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 two.

The above practice yields several summary statistics for a single system, which is good: Systems should always be evaluated 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 β be set so that the metrics correlate well with *user satisfaction*? This is 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*.

References

[1] Asakawa, S. and Selberg, E.: The New MSN Search Engine Developed by Microsoft (*in Japanese*), *Information Processing Society of Japan Magazine*, Vol 46, No. 9, pp. 1008-1015, 2005.

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

[3] Della Mea, V. & Mizzaro, S.: Measuring Retrieval Effectiveness: A New Proposal and a First Experimental Validation. *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 6, pp. 530-543, 2004.

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

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

[6] 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.

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

[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.

[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] Korfhage, R. R.: *Information Storage and Retrieval*, Wiley Computer Publishing, 1997.

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

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

[14] Sakai, T.: For Building Better Retrieval Systems: Trends in Information Retrieval Evaluation based on Graded Relevance (in Japanese), *IPSJ Magazine*, Vol. 47, No. 2, pp. 147-158, 2006.

[15] 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.

[16] Sakai, T.: A Further Note on Evaluation Metrics for the Task of Finding One Highly Relevant Document, *IPSJ SIG Technical Reports*, 2006-FI-82/2006-DD-54, pp. 69-76, 2006.

[17] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR 2006 Proceedings*, to appear, 2006.

[18] Sakai, T.: Give Me Just One Highly Relevant Document: P-measure, *ACM SIGIR 2006 Proceedings*, to appear, 2006.

[19] Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, *AIRS 2006 Proceedings*, to appear, 2006.

[20] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, to appear, 2006.

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

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

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

[24] 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.