ANOTHER APPENDIX TO
New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering

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

This appendix shows the reliability of Q-measure and R-measure using the actual submitted runs from the NTCIR-3 CLIR task. The following files were used for the analyses:

- `ntc3clir-allCruns.20040511.zip` (45 Runs for retrieving Chinese documents)
- `ntc3clir-allJruns.20040511.zip` (33 Runs for retrieving Japanese documents)
- `ntc3clir-allEruns.20040511.zip` (24 Runs for retrieving English documents)
- `ntc3clir-allKruns.20040511.zip` (14 Runs for retrieving Korean documents)

Prior to empirical analyses, we provide some theoretical analyses that will help interpret the experimental results.

By definition of the *cumulative bonused gain* (See Section 3.1),

$$cbg(r) = cg(r) + count(r) \tag{14}$$

holds for $r \geq 1$. Therefore, Q-measure and R-measure can alternatively be expressed as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{cg(r) + count(r)}{cig(r) + r} \tag{15}$$

$$R\text{-}measure = \frac{cg(R) + count(R)}{cig(R) + R} \tag{16}$$

Comparing the above with Equations (1), (2), (3) and (4), it can be observed that Q-measure and R-measure are "blended" metrics: Q-measure inherits the properties of both AWP and Average Precision, and R-measure inherits the properties of both R-WP and R-Precision. Moreover, it is clear from the above that using large gain values would emphasise the AWP aspect of Q-measure, while using small gain values would emphasise its Average Precision aspect. Similarly, using large gain values would emphasize the R-WP aspect of R-measure, while using small gain values would emphasise its R-Precision aspect. For example, letting $gain(S) = 30$, $gain(A) = 20$, and $gain(B) = 10$ (or conversely $gain(S) = 0.3$, $gain(A) = 0.2$, and $gain(B) = 0.1$) instead of $gain(S) = 3$, $gain(A) = 2$, and $gain(B) = 1$ is equivalent to using the following generalised equations and letting $\beta = 10$ (or conversely $\beta = 0.1$):

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{\beta cg(r) + count(r)}{\beta cig(r) + r} \tag{17}$$

$$R\text{-}measure = \frac{\beta cg(R) + count(R)}{\beta cig(R) + R} \tag{18}$$

If the relevance assessents are binary, then both

$$cg(r) = count(r) \tag{19}$$

$$cig(r) = r \tag{20}$$

hold for $r \leq R$. Thus, as have been mentioned in Section 2.3, with binary relevance,

$$cg(r)/cig(r) = count(r)/r \tag{21}$$

holds for $r \leq R$. Therefore, with binary relevance, AWP is equal to Average Precision if the system output does not have any relevant documents below Rank R. Moreover, Equation (21) implies that, with binary relevance, R-WP is always equal to R-Precision.

A similar theoretical analysis is possible for Q-measure and R-measure as well. If the relevance assessments are binary, then, from Equations (19) and (20),

$$\frac{cg(r) + count(r)}{cig(r) + r} = \frac{2count(r)}{2r} = \frac{count(r)}{r} \tag{22}$$

holds for $r \leq R$. Therefore, for binary relevance, Q-measure is equal to Average Precision (and to AWP) if the system output does not have any relevant documents below Rank R. Similarly, with binary relevance, R-measure is always equal to R-Precision (and to R-WP).

Furthermore, as $count(r) \leq r$ holds for $r \geq 1$,

$$Q\text{-}measure \leq AWP \quad (23)$$

and

$$R\text{-}measure \leq R\text{-}WP \quad (24)$$

hold.

Tables 3-6 show the Spearman and Kendall Rank Correlations for Q-measure and its related metrics based on the NTCIR-4 CLIR C-runs, J-runs, E-runs, and K-runs, respectively. The correlation coefficients are equal to 1 when two rankings are identical, and are equal to -1 when two rankings are completely reversed. (It is known that the Spearman's coefficient is usually higher than the Kendall's.) Values higher than 0.99 (i.e. extremely high correlations) are indicated in **boldface**. "Relaxed" represents Relaxed Average Precision, "Rigid" represents Rigid Average Precision, and "Q-measure" and "AWP" use the *default* gain values: $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$. Moreover, the columns in Part (b) of each table represent Q-measure with different gain values: For example, "Q30:20:10" means Q-measure using $gain(S) = 30$, $gain(A) = 20$ and $gain(B) = 10$ (Recall Equation 17). Thus, "Q1:1:1" implies binary relevance, and "Q10:5:1" implies stronger emphasis on highly relevant documents.

Figures 4-7 visualise the above tables, respectively, by sorting systems in decreasing order of *Relaxed Average Precision* and then renaming each system as System No. 1, System No. 2, and so on. Thus, the Relaxed Average Precision curves are guaranteed to decrease monotonically, and the other curves (representing system rankings based on other metrics) would also decrease monotonically only if their rankings agree perfectly with that of Relaxed Average Precision. That is, an increase in a curve represents a *swop*.

The above tables and figures are shown in order of decreasing reliability: Table 3/Figure 4 are based on 45 systems, while Table 6/Figure 7 are based on only 14 systems. Furthermore, Table 7 condenses Tables 3-6 into one by taking averages over the four sets of data.

From the above results regarding Q-measure, we can observe the following:

1. While it is theoretically clear that AWP is unreliable when relevant documents are retrieved below Rank R, our experimental results confirm this fact. The AWP curves include many swops, and some of them are represented by a very "steep" increase. This is due to the fact that AWP overestimates a system's performance which rank many relevant documents below Rank R.

2. Compared to AWP, the Q-measure curves are clearly more stable. Moreover, from Part (a) of each table, Q-measure is more highly correlated with Relaxed Average Precision than AWP is, *and* is more highly correlated with Rigid Average Precision than AWP is. Thus, Q-measure nicely combines the advantages of Average Precision and AWP.

3. From Part (a) of each table, it can be observed that Q-measure is more highly correlated with *Relaxed* Average Precision than with *Rigid* Average Precision. (The same is true for AWP as well.) This is natural, as Rigid Average Precision ignores the B-relevant documents completely.

4. It can be observed that the behaviour of Q-measure is relatively stable with respect to the choice of the gain values. Moreover, by comparing "Q30:20:10", "Q-measure" (i.e. Q3:2:1) and "Q0.3:0.2:0.1" in terms of correlations with "Relaxed", it can be observed that using smaller gain values means more resemblance with Relaxed Average Precision (Recall Equation (17)). For example, in Table 3, the Spearman's correlation is 0.9909 for "Q30:20:10" and "Relaxed", 0.9982 for "Q-measure" and "Relaxed", and 0.9997 for "Q0.3:0.2:0.1" and "Relaxed". This property is also visible in the graphs: while each "Q30:20:10" curve resembles the corresponding AWP curve, each "Q0.3:0.2:0.1" curve is almost indistisguishable from the "Relaxed" curve.

5. From Part (b) of each table, it can observed that "Q1:1:1" (i.e. Q-measure with binary relevance) is very highly correlated with Relaxed Average Precision. (Recall that "Q1:1:1" would equal Relaxed Average Precision if a system output does not have any relevant documents below Rank R.)

Tables 8-11 show the Spearman and Kendall Rank Correlations for R-measure and its related metrics based on the NTCIR-4 CLIR C-runs, J-runs, E-runs, and K-runs, respectively. Table 12 condenses Tables 8-11 into one by taking averages over the four sets of data. Again, "Q-measure", "R-measure" and "R-WP" use the default gain values, "R30:20:10" represents R-measure using $gain(S) = 30$, $gain(A) = 20$ and $gain(B) = 10$, and so on. As "R1:1:1" (R-measure with binary relevance) is identical to R-Precision (and R-WP), it is not included in the tables.

From the above results regarding R-measure, we can observe the following:

1. From Part (a) of each table, it can be observed that R-measure, R-WP and R-Precision are very highly correlated with one another. Moreover, R-measure is slightly more highly correlated with R-Precision than R-WP is: Compare Equations (2), (4) and (16).

2. From the tables, it can be observed that R-measure is relatively stable with respect to the

choice of the gain values. By comparing "R30:20:10", "R-measure" (i.e. R3:2:1) and "R0.3:0.2:0.1" in terms of correlations with R-Precision, it can be observed that using smaller gain values means more resemblance with R-Precision (Recall Equation (18)). For example, in Table 8, the Spearman's correlation is 0.9939 for "R30:20:10" and "Relaxed", 0.9960 for "R-measure" and "Relaxed", and 0.9982 for "R0.3:0.2:0.1" and "Relaxed".

Thus, our experiments show that Q-measure and R-measure are reliable IR performance metrics for evaluations based on multigrade relevance.

Acknowledgement

The author is indebted to the NTCIR-3 CLIR Organisers, most of all Noriko Kando, for making the NTCIR-3 CLIR data available to us for research purposes. I would also like to thank the NTCIR-3 CLIR participants who have agreed to the release of their submission files.

Table 3. Spearman/Kendall Rank Correlations for the 45 C-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9874/.9273	**.9982**/.9798	.9802/.8990
Rigid	-	.9858/.9192	.9648/.8667
Q-measure	-	-	.9851/.9152
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	**.9909**/.9374	**.9997**/**.9960**	**.9989**/.9879	**.9947**/.9556
Rigid	.9788/.8970	.9874/.9273	.9851/.9192	.9829/.9111
Q-measure	**.9901**/.9333	**.9978**/.9798	**.9984**/.9798	**.9955**/.9636

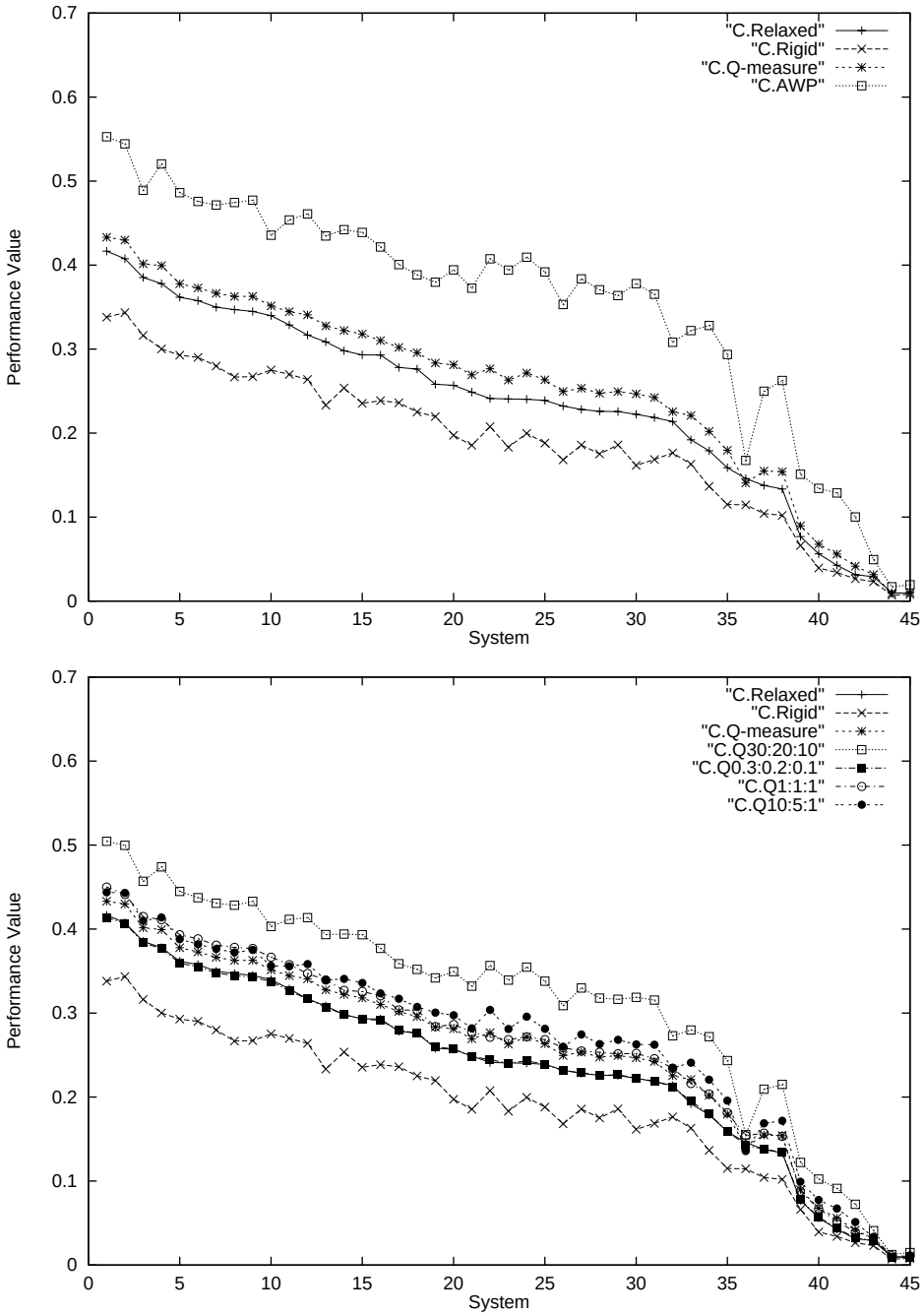

Figure 4. System ranking comparisons with Relaxed Average Precision (C-runs).

Table 4. Spearman/Kendall Rank Correlations for the 33 J-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9619/.8561	**.9947**/.9583	.9833/.9242
Rigid	-	.9616/.8447	.9505/.8182
Q-measure	-	-	.9813/.9129
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	.9769/.9015	**.9980**/.9811	**.9990**/.9886	.9759/.8977
Rigid	.9395/.7879	.9592/.8447	.9616/.8523	.9519/.8144
Q-measure	.9729/.8826	**.9943**/.9545	**.9943**/.9545	.9706/.8864

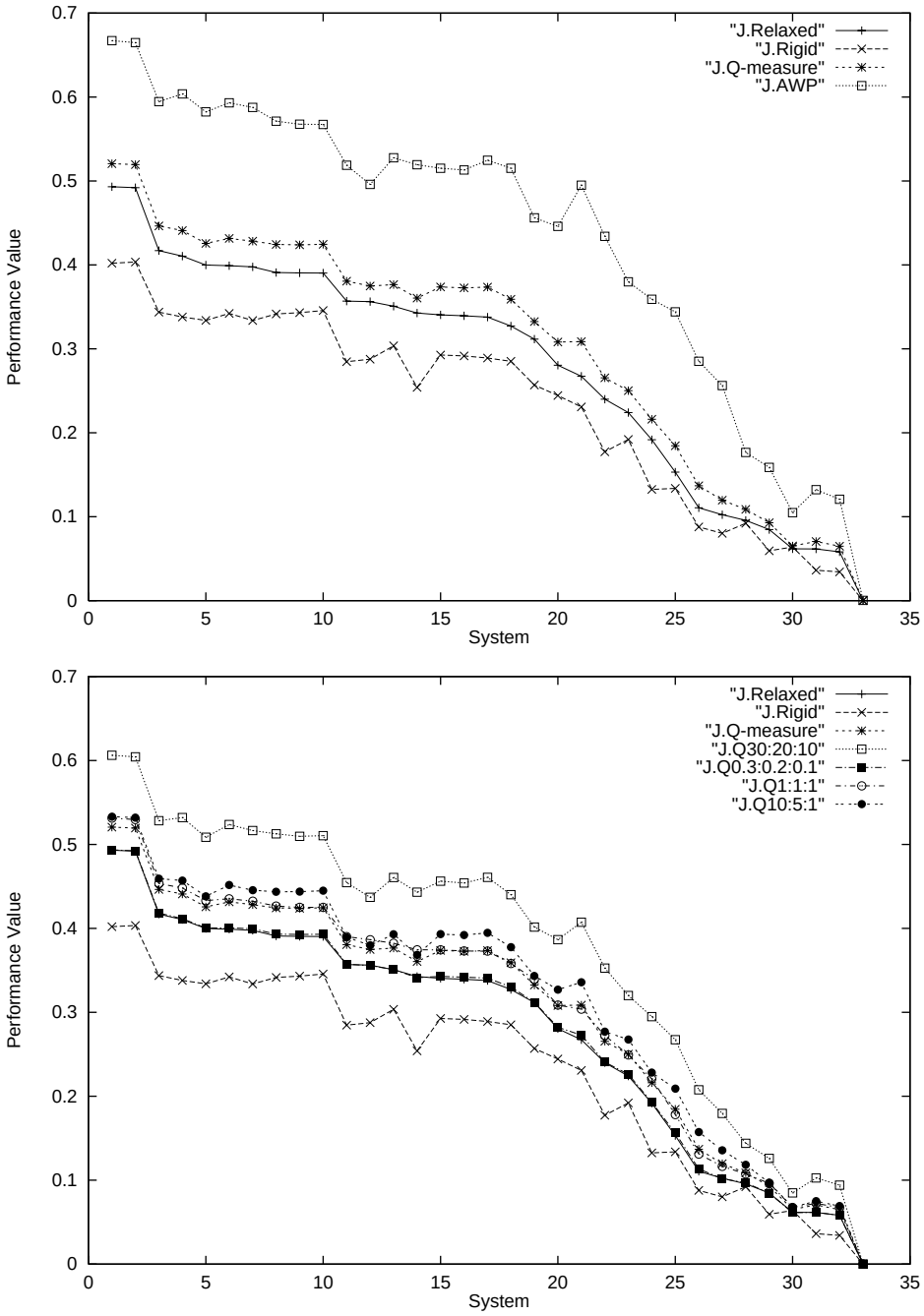

Figure 5. System ranking comparisons with Relaxed Average Precision (J-runs).

Table 5. Spearman/Kendall Rank Correlations for the 24 E-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	**.9922**/.9565	**.9974**/.9783	.9835/.9058
Rigid	-	**.9948**/.9638	.9748/.8913
Q-measure	-	-	.9843/.9130
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	**.9922**/.9565	**1.000**/**1.000**	**.9965**/.9783	.9887/.9348
Rigid	.9852/.9275	**.9922**/.9565	**.9904**/.9493	.9887/.9348
Q-measure	**.9904**/.9493	**.9974**/.9783	**.9957**/.9710	.9887/.9420

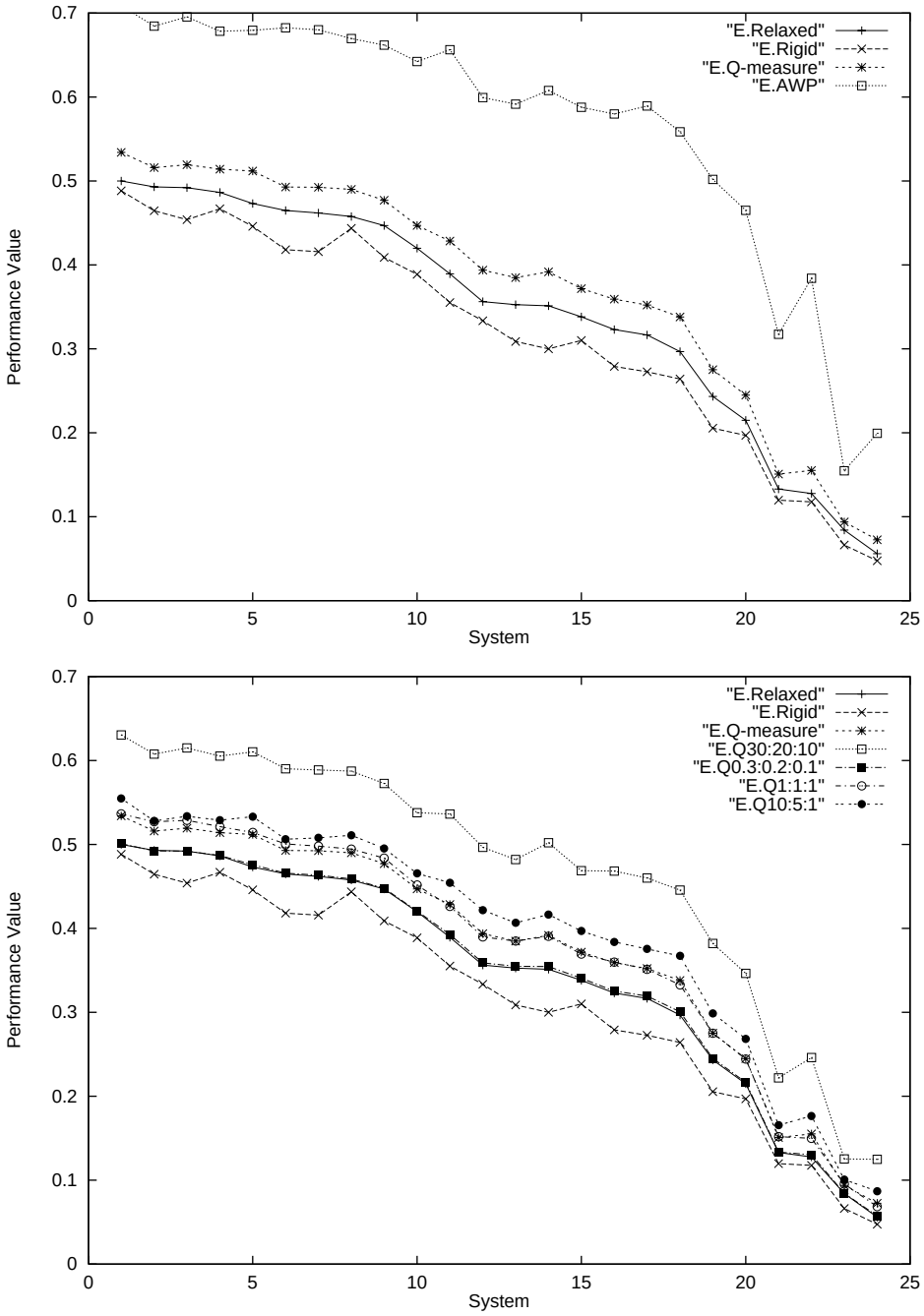

Figure 6. System ranking comparisons with Relaxed Average Precision (E-runs).

Table 6. Spearman/Kendall Rank Correlations for the 14 K-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9560/.8462	**.9912**/.9560	**.9912**/.9560
Rigid	-	.9385/.8022	.9385/.8022
Q-measure	-	-	**1.000**/**1.000**
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	**.9912**/.9560	**.9956**/.9780	**1.000**/**1.000**	**.9912**/.9560
Rigid	.9385/.8022	.9516/.8242	.9560/.8462	.9385/.8022
Q-measure	**1.000**/**1.000**	**.9956**/.9780	**.9912**/.9560	**1.000**/**1.000**

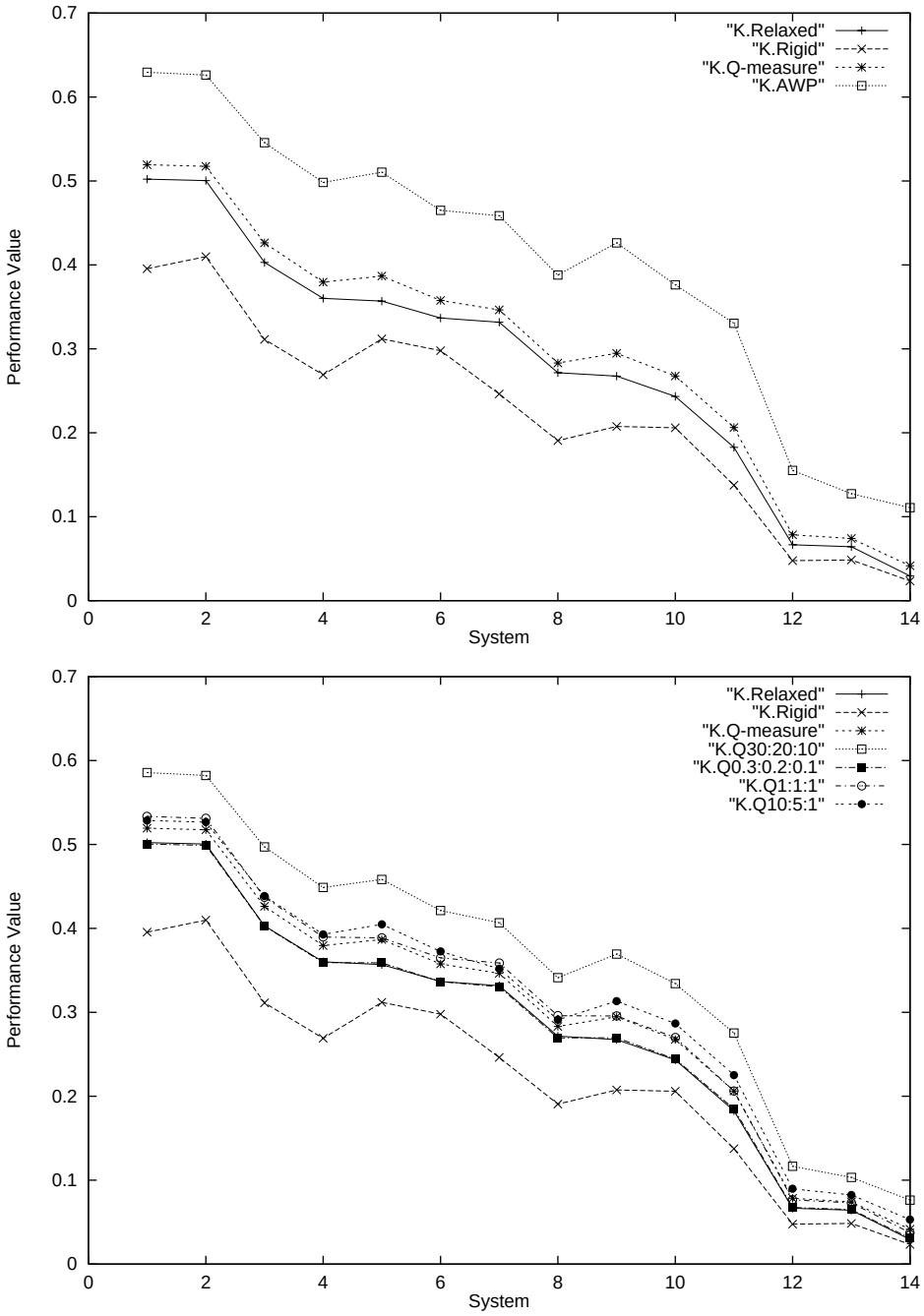

Figure 7. System ranking comparisons with Relaxed Average Precision (K-runs).

Table 7. Spearman/Kendall Rank Correlations: Averages over C, J, E and K (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9744/.8965	**.9954**/.9681	.9846/.9213
Rigid	-	.9702/.8825	.9571/.8446
Q-measure	-	-	.9877/.9353
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	.9878/.9378	**.9983**/.9888	**.9986**/.9887	.9876/.9360
Rigid	.9605/.8537	.9726/.8882	.9733/.8918	.9655/.8656
Q-measure	.9884/.9413	**.9963**/.9727	**.9949**/.9653	.9887/.9480

Table 8. Spearman/Kendall Rank Correlations for the 45 C runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9864/.9313	.9867/.9293	.9863/.9293
Q-measure	.9867/.9232	.9871/.9253	.9883/.9333
R-Precision	-	**.9960**/.9616	**.9938**/.9495
R-measure	-	-	**.9971**/.9758
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9862/.9273	.9870/.9333	.9838/.9232
R-Precision	**.9939**/.9515	**.9982**/.9818	.9845/.9152
R-measure	**.9972**/.9778	**.9976**/.9758	.9893/.9333

Table 9. Spearman/Kendall Rank Correlations for the 33 J runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9886/.9356	.9866/.9318	.9843/.9242
Q-measure	**.9913**/.9318	**.9903**/.9356	.9880/.9280
R-Precision	-	**.9923**/.9583	**.9900**/.9356
R-measure	-	-	**.9910**/.9470
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9850/.9280	.9883/.9356	.9830/.9205
R-Precision	**.9920**/.9470	**.9957**/.9697	.9873/.9242
R-measure	**.9930**/.9583	**.9910**/.9583	.9883/.9356

Table 10. Spearman/Kendall Rank Correlations for the 24 E runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9852/.9275	.9870/.9348	.9870/.9348
Q-measure	.9843/.9203	.9835/.9130	.9835/.9130
R-Precision	-	**.9948**/.9638	**.9948**/.9638
R-measure	-	-	**1.000**/**1.000**
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9870/.9348	.9852/.9275	.9713/.8913
R-Precision	**.9948**/.9638	**.9983**/.9855	.9626/.8478
R-measure	**1.000**/**1.000**	**.9965**/.9783	.9591/.8551

Table 11. Spearman/Kendall Rank Correlations for the 14 K runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9868/.9560	.9868/.9560	.9824/.9341
Q-measure	.9780/.9121	.9780/.9121	.9824/.9341
R-Precision	-	**1.000**/**1.000**	**.9956**/.9780
R-measure	-	-	**.9956**/.9780
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9824/.9341	.9868/.9560	.9824/.9341
R-Precision	**.9956**/.9780	**1.000**/**1.000**	**.9956**/.9780
R-measure	**.9956**/.9780	**1.000**/**1.000**	**.9956**/.9780

Table 12. Spearman/Kendall Rank Correlations: Averages over C, J, E and K (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9868/.9376	.9868/.9380	.9850/.9306
Q-measure	.9851/.9219	.9847/.9215	.9856/.9271
R-Precision	-	**.9958**/.9709	**.9936**/.9567
R-measure	-	-	**.9959**/.9752
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9852/.9311	.9868/.9381	.9801/.9173
R-Precision	**.9941**/.9601	**.9980**/.9843	.9825/.9163
R-measure	**.9964**/.9785	**.9963**/.9781	.9831/.9255