Results of Evaluation Runs and Data Analysis in the LibRank project

Christiane Behnert

Kim Plassmeier

10.08.2016

Table of contents

List of tables

1. Introduction

Within the research project LibRank[1] we identified factors for relevance ranking in library information systems based on ranking in web search engines (Behnert & Lewandowski, 2015). To empirically evaluate the identified ranking factors, we developed a research design involving several evaluation runs with human assessors using the Relevance Assessment Tool (RAT) (see report "Working Paper on Evaluation Methods within the LibRank Project"). After implementing the ranking factors based on (1) underlying data (see internal technical report "Requirements specification for the test system") from heterogeneous sources and (2) the relevance model (see internal technical report "Relevance model"), we conducted a pretest and started the first of three evaluation runs in August 2015. Table 1 provides an overview of the evaluation runs.

	Evaluation run #1	**Evaluation run #2**	**Evaluation run #3**
Number of assessors	4	8	a) Students: 45
Group of assessors	Subject specialists, German National Library of Economics	Subject specialists, German National Library of Economics	Diverse a) Students with economic background b) Doctoral students in economics c) Researchers holding a Doctoral degree in economics
Number of tasks/Access codes per assessor	30	15	Per contingent: a) 5 b) 5 c) 3
Total number of tasks	120	120	600 a) 450 b) 100 c) 50
Total number of completed tasks	83	109	a) 363 b) n.n.* c) n.n.*
Cut-off value	20	20	20
Rankings	10	10	10
Assessment period	6.8. - 28.8.2015	15. - 28. 10.2015	16.12.2015-8.2.2016

Table 1: Overview on evaluation runs. *The analysis of assessment data by these two groups cannot be completed within the project due to time delay in recruiting the critical number of assessors.

The rest of the paper is structured as follows: we present the results of the data analysis to every evaluation run and discuss possible reasons for the results in separate sections before we summarize our findings.

[1] http://www.librank.info/

2. Results

2.1. Run #1

In run #1 we asked subject indexers from the German National Library of Economics (ZBW – Zentralbibliothek Wirtschaftswissenschaften) to assess the relevance of documents. The documents had been pooled according to the ranking schemes given in Table 13. In fact, the documents from Ranking 1 were not included in the pool due to an input error. However, most documents produced by Ranking 1 were also produced by other rankings, so that the effect on the pool is negligible.

Four subject indexers attended in this run. In total, they judged 83 search tasks, resulting in 3,470 binary and graded relevance judgements.

We observed an unexpected effect in the relevance judgements for several tasks. In these tasks some documents marked as non-relevant (binary relevance) received high gradual judgements. This may be because assessors judged the degree of *non-relevance* in these cases. These judgments might also be considered as noise.

Nevertheless, we opted to remove these suspicious tasks from the evaluation by applying a high sensitive filter to be confident that no "inverted" judgements are included in the pool. Therefore we removed all tasks for which the difference of the maximal non-relevant judgement and the minimal relevant judgement exceeded a threshold of ten points.

This leaves us with a cleaned pool, which consists of 68 tasks and 2,778 binary and graded relevance judgements. On average a search task had 41 ± 17 documents that indicates a high overlap between the documents returned by the individual rankings.

About $64 \pm 30\%$ of the documents of a task have been judged relevant (binary relevance), i.e., the pool consists of very many relevant documents. This is probably an artifact of the simplified pooling method.

For each ranking we calculate precision at 10 (Prec@10), normalized discounted cumulative gain at 10 (NDCG@10), and normalized expected reciprocal rank (nERR). The results are given in Table 2. We note that the mean performance only varies slightly for the different rankings while the variance over the tasks is large.

Ranking	Prec@10	NDCG@10	nERR
Ranking 1	0.69+/-0.33	0.58+/-0.27	0.66+/-0.30
Ranking 2	0.73+/-0.28	0.66+/-0.21	0.75+/-0.24
Ranking 3	0.71+/-0.32	0.60+/-0.26	0.68+/-0.29
Ranking 4	0.69+/-0.34	0.58+/-0.28	0.66+/-0.29
Ranking 5	0.72+/-0.32	0.64+/-0.24	0.72+/-0.27
Ranking 6	0.72+/-0.31	0.60+/-0.24	0.67+/-0.27
Ranking 7	0.72+/-0.31	0.60+/-0.26	0.67+/-0.30
Ranking 8	0.72+/-0.32	0.63+/-0.26	0.72+/-0.27
Ranking 9	0.71+/-0.33	0.60+/-0.27	0.67+/-0.29
Ranking 10	0.72+/-0.32	0.62+/-0.26	0.71+/-0.27

Table 2: Mean performance and standard deviation for the rankings in run #1.

We use the methodology proposed by Sakai (2014) to test for statistically significant differences. We only report the differences with respect to nDCG@10. First, we apply a two-way ANOVA test in order

to identify statistically significant differences between the test rankings. The results of the ANOVA test are given in Table 3. They show that the effect is significant. Hence, we use Tukey's HSD test to assess pairs of systems for significant differences. The significant pairs are listed in Table 4. The differences between the EconBiz baseline (Ranking 2) and all other test rankings, except Ranking 5, Ranking 8 and Ranking 10, are statistically significant ($p < 0.05$). The difference between the EconBiz baseline and the "text statistics only" baseline (Ranking 1) is in particular statistically significant. As we will see, this is not the case in the other runs. Hence, no system achieved a better performance than the EconBiz baseline and only the EconBiz baseline and Ranking 5 performed better than the "text statistics only" baseline. Figure 1 shows the mean performances of the systems with the simultaneous 95% confidence intervals.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Test rankings	9	0.39	0.04	5.2	0
SearchTask	67	38.65	0.58	69.8	0
Residuals	603	4.98	0.01	NA	NA

Table 3: Result of the two-way ANOVA test. Given are the degrees of freedom (Df), the sum of squares (Sum Sq), the mean sum of squares (Mean Sq), the F value, and the p-value (Pr(>F)).

	diff	lwr	upr	p adj
Ranking 2-Ranking 1	0.08	0.03	0.13	0.00
Ranking 5-Ranking 1	0.06	0.01	0.11	0.01
Ranking 3-Ranking 2	-0.06	-0.11	-0.01	0.00
Ranking 4-Ranking 2	-0.08	-0.13	-0.03	0.00
Ranking 6-Ranking 2	-0.06	-0.11	-0.01	0.01
Ranking 7-Ranking 2	-0.06	-0.11	-0.01	0.01
Ranking 9-Ranking 2	-0.06	-0.11	-0.01	0.01
Ranking 5-Ranking 4	0.06	0.01	0.11	0.01

Table 4: Pairs of test rankings with statistically significant differences. Given are the mean difference (diff), the boundaries of the confidence interval (lwr, upr) and the adjusted p-value (p adj).

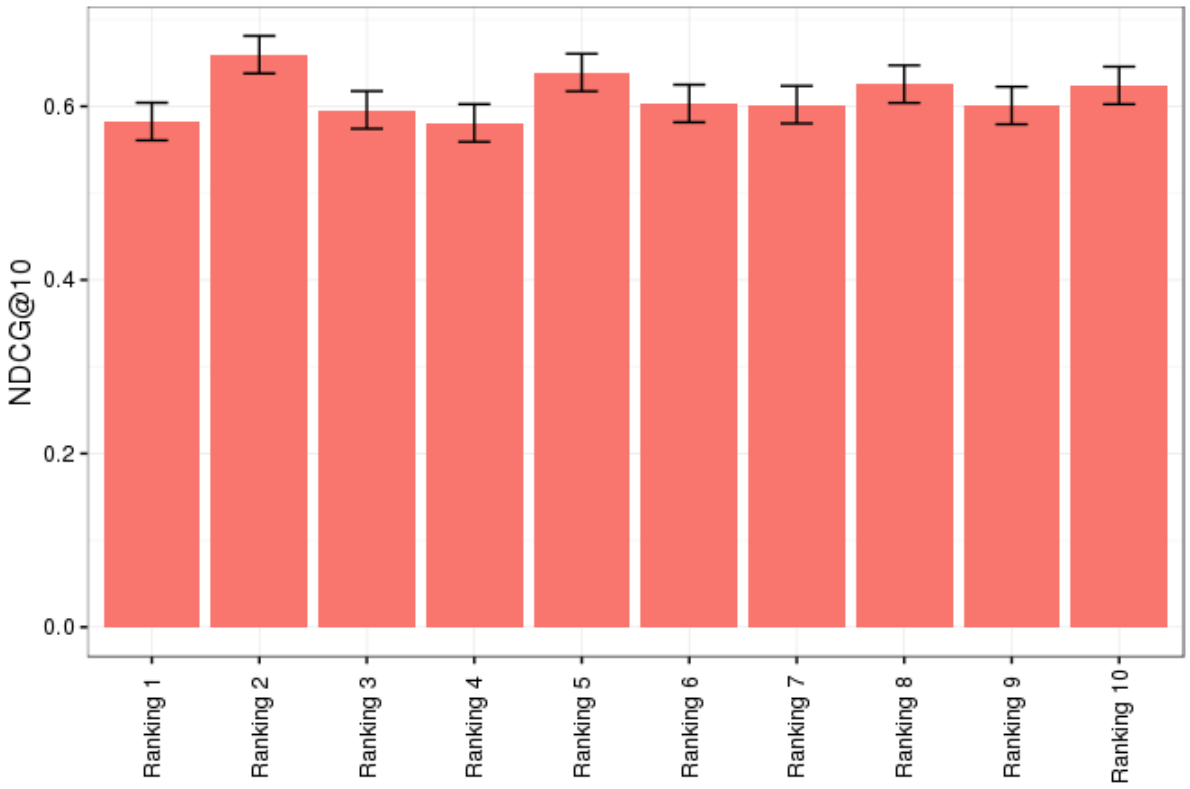


Figure 1: Comparison of the mean performances of the systems in run #1. Error bars indicate the simultaneous 95% confidence intervals obtained from Tukey's HSD test.

2.2. Run #2

Since the rankings in run #1 performed very similar and the pool was smaller than expected, in this run we wanted to maximize the differences between the rankings. Therefore, we increased the weight for the query independent features until the mean pool size started to saturate (w_{qi}=1,000,000). By increasing w_{qi} we reduced the effect of the text statistics on the rankings. Since the text statistics still acted as a multiplicative factor to the query independent features, the top20 results of each of the test rankings were supposed to not only include non-relevant documents.

The rankings used to create the pool in run #2 are given in Table 13. Eight subject indexers took part in this run (three of them also attended in run #1). In total, they judged 113 search tasks, resulting in 8,114 binary and graded relevance judgements.

As in run #1, we again observed large overlaps in the graded relevance judgments for relevant and non-relevant documents for some tasks. In this run we manually inspected the suspicious tasks and kept the tasks for which the overlap was only related to a small fraction of marginally documents for further analysis. Removing the other tasks left us with a cleaned pool, which consisted of 108 tasks and 7,728 binary and graded relevance judgements. We note that removing all suspicious tasks, had only a minimal effect on the performance results.

On average a search task included 72 ± 29 documents, which was considerably larger than the mean task size in run 1. This indicates that the top 20 results returned by each ranking were indeed more diverse than in run #1.

About $53 \pm 31\%$ of the documents of a task have been judged relevant (binary scale). The difference to run #1 is statistically significant, but the magnitude is still large.

The performance results for the rankings of run #2 are given in Table 5. While the adjusted rankings had a clear effect on the pool size, the mean performances for most rankings are still very similar. The magnitude of the performance values is now lower than in run #3. We will discuss this effect in section 3.3.

Ranking	**Prec@10**	**NDCG@10**	**nERR**
Ranking 11	0.64+/-0.34	0.56+/-0.28	0.66+/-0.31
Ranking 12	0.62+/-0.35	0.55+/-0.30	0.66+/-0.34
Ranking 13	0.62+/-0.35	0.55+/-0.29	0.66+/-0.33
Ranking 14	0.61+/-0.34	0.52+/-0.27	0.66+/-0.31
Ranking 15	0.64+/-0.34	0.57+/-0.29	0.66+/-0.32
Ranking 16	0.50+/-0.35	0.42+/-0.27	0.56+/-0.33
Ranking 17	0.61+/-0.35	0.51+/-0.29	0.62+/-0.33
Ranking 18	0.64+/-0.34	0.58+/-0.29	0.67+/-0.32
Ranking 19	0.57+/-0.36	0.47+/-0.30	0.60+/-0.35
Ranking 20	0.63+/-0.34	0.56+/-0.29	0.65+/-0.33

Table 5: Mean performance and standard deviation for the rankings in run #2.

The results of the two-way ANOVA test are given in Table 6. Again, the system effect is significant. The significant pairs according to Tukey's HSD test are listed in Table 7. The differences between Ranking 16 and all other test rankings, except Ranking 19, and the differences of Ranking 19 and all other rankings, except Ranking 14, Ranking 16 and Ranking 17, are significant. The best performing ranking in this run is Ranking 18 (Figure 2). However, the differences to the two baseline systems are not statistically significant. Hence, no ranking achieved a significantly better performance than the two baseline systems. In contrast to the first run, the difference between the two baselines is again not statistically significant.

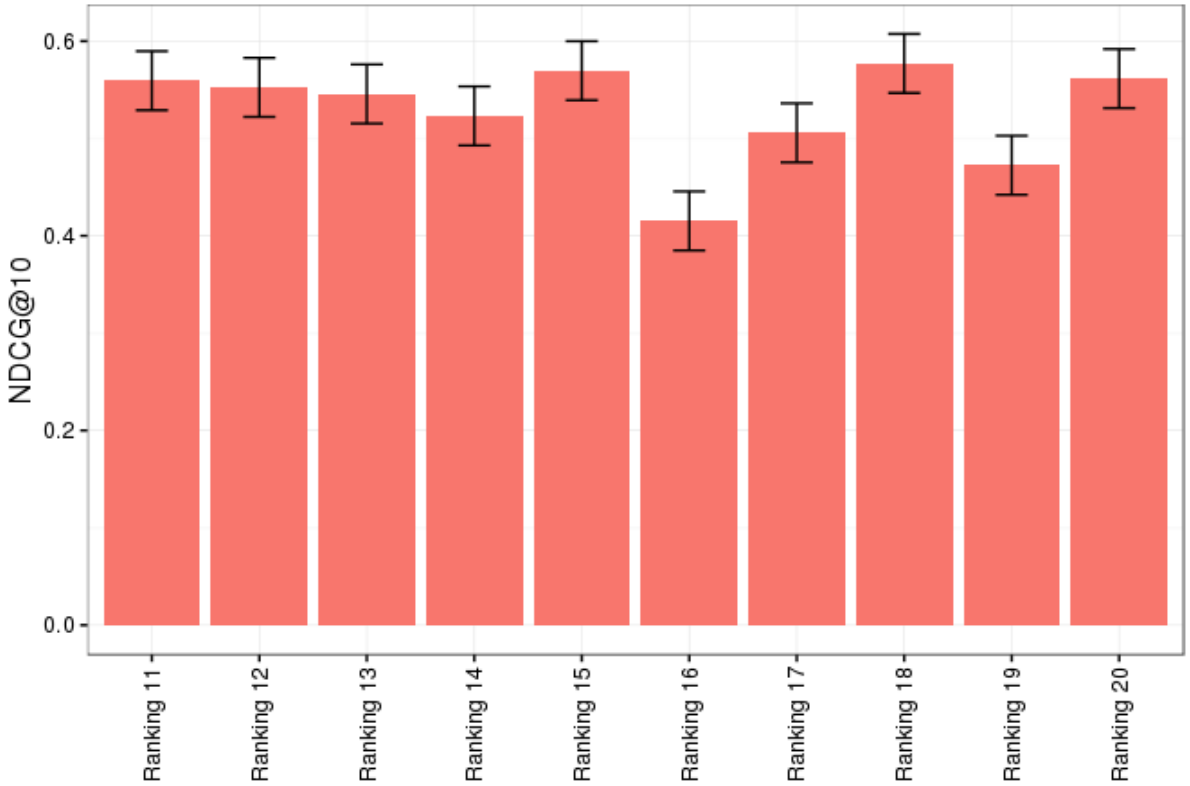
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Test ranking	9	2.5	0.28	11	0
SearchTask	107	64.2	0.60	23	0
Residuals	963	24.8	0.03	NA	NA

Table 6: Result of the two-way ANOVA test. Given are the degrees of freedom (Df), the sum of squares (Sum Sq), the mean sum of squares (Mean Sq), the F value, and the p-value (Pr(>F)).

	diff	lwr	upr	p adj
Ranking 16-Ranking 11	-0.14	-0.21	-0.07	0.00
Ranking 19-Ranking 11	-0.09	-0.16	-0.02	0.00
Ranking 16-Ranking 12	-0.14	-0.21	-0.07	0.00
Ranking 19-Ranking 12	-0.08	-0.15	-0.01	0.01
Ranking 16-Ranking 13	-0.13	-0.20	-0.06	0.00
Ranking 19-Ranking 13	-0.07	-0.14	0.00	0.03
Ranking 16-Ranking 14	-0.11	-0.18	-0.04	0.00
Ranking 16-Ranking 15	-0.15	-0.22	-0.08	0.00
Ranking 19-Ranking 15	-0.10	-0.17	-0.03	0.00
Ranking 17-Ranking 16	0.09	0.02	0.16	0.00
Ranking 18-Ranking 16	0.16	0.09	0.23	0.00
Ranking 20-Ranking 16	0.15	0.08	0.22	0.00
Ranking 18-Ranking 17	0.07	0.00	0.14	0.04

Ranking 19-Ranking 18	-0.10	-0.17	-0.04	0.00
Ranking 20-Ranking 19	0.09	0.02	0.16	0.00

Table 7: Pairs of systems with statistically significant differences. Given are the mean difference (diff), the boundaries of the confidence interval (lwr, upr) and the adjusted p-value (p adj).



2.3. Figure 2: Comparison of the mean performances of the systems in run #2. Error bars indicate the simultaneous 95% confidence intervals obtained from Tukey's HSD test.Run #3

Run #3 is different from the previous runs primarily with respect to the group and number of assessors. In this run we recruited the assessors via crowdsourcing. We intended to recruit assessors for different user groups (undergraduate students, graduate students, and graduated researchers), however, we only achieved to recruit a sufficient number of assessors for the first group (undergraduate students).

We contacted undergraduate students of economic fields of studies from different German universities via their institutional mailing lists or by putting up a notice on their Faculty's bulletin board. As a financial incentive for participation we offered Amazon gift cards with a value of 20 EUR for every student completing a contingent of five tasks. The more contingents would be completed, the more gift card codes one assessor could earn, i.e., if a student completed ten tasks, he or she got 40 EUR and so forth (the max. amount of money one student received was 100 EUR).

When relevance judgements are gathered via crowdsourcing, in general, quality control mechanism, such as qualification tests or honey pots, should be implemented (Alonso & Lease, 2011; Alonso, 2013). Due to our recruitment procedure, we assumed that our assessors were sufficiently qualified to judge the relevance of the documents (with regard to their user group). As the only additional

quality control, we filtered out tasks, for which more than 20% of the documents had been assessed in less than 6 seconds (plausibility check). We considered the corresponding assessors to show a lack of effort.

In the third run the characteristics of the pool composition were adjusted again. The pools in the previous runs contained a substantial amount of duplicate records, which was criticized by some assessors. Duplicate records exist in EconBiz, since it aggregates multiple databases but does not apply any means to merge records from different source databases, yet. Because of the feedback of the assessors, we identified and removed duplicate records in the pool of run #3 semi-automatically.

Finally, the assessment procedure was also adjusted. In the third run we swapped the order of the assessment of binary and graded relevance, i.e., graded relevance was assessed prior to binary relevance. This was motivated by the occurrence of the "inverted" judgements as described in run #1. We speculated that judging a document as non-relevant (binary relevance) first favored the interpretation to judge the degree of *non-relevance* on the gradual scale.

As in the previous run, the rankings used to create the pool for run #3were adjusted. The "text only" baseline (Ranking 21) and the EconBiz baseline (Ranking 22) were the same as in the previous two runs, as well as Rankings 24, 25, 26, and 27 used in run #2. Rankings 28, 29, and 30 were based on hypothesized profiles for the relative importance of the relevance criteria for the different user groups. The hypotheses are given in Table 8. Finally, we used a learning-to-rank approach based on the relevance assessments of run #2 to determine Ranking 23: we used a Coordinate Ascent (CA) optimization approach to directly optimize nDCG@10. Despite its simplicity, Busa-Fekete, Szarvas, Élteto, & Kégl (2012) found CA being competitive with other state-of-the-art learning-to-rank algorithms. The different rankings are listed in Table 13.

Priority with regard to…	**a) Students**	**b) Doctoral students**	**c) Having a doctorate**	**Affects ranking factor…**
Immediate, free access	+++ (free of charge, prefer electronic access, prepare papers at home)	+ (licensed via library)	+ (licensed via library)	Group Locality & Availability / Availability
Freshness	+ (general overview, subordinated)	++ (recent information, general overview)	+++	Group Freshness / Publication Date
Popularity: Usage	+++ (possibly makes up for freshness, e.g., stronger demand for recent textbooks)	+	+	Group Popularity / Usage
Popularity: Citations, Impact	+++ (presumed to be unknowingly prioritized, as implied in other	+++	+	Group Popularity / Reference

	factors like usage)			
RANKING WEIGHTS MAX.	10 P. = 100% 1 => 10 % 3 => 30 %	7 P. = 100% 1 => 14,23% 2 => 28,57 3 => 42,86	6 P. = 100% 1 => 16,67 2 => 33,33 3 => 50	

Table 8: Hypotheses regarding assessor groups for the choice of ranking factors and the occurrence of their weights in evaluation run #3.

In this final run 45 undergraduate students participated. In total, they judged 363 search tasks, resulting in 23,574 binary and graded relevance judgements.

We still observed tasks with large overlaps between relevant and non-relevant gradual judgements. Thus, swapping the order of the binary and gradual assessment did not resolve this issue. In fact, in this run the proportion of affected tasks is much larger than in the previous runs (129 out of 363 tasks). However, we speculate that this is more related to the assessors than to the altered assessment procedure.

105 tasks did not pass the plausibility filter, while 47 of these tasks also show overlapping relevance judgements. Hence, we additionally removed 58 tasks. Finally, we removed another 13 tasks for which the results lists contained less than 30 documents, and another one with scraping errors. This leaves us with a cleaned poolincluding 162 tasks and 10,396 binary and graded relevance judgements. On average a search task has 64 ± 24 documents, which is similar to the task size in run #2.

About $58 \pm 27\%$ of the documents of a task have been judged relevant (binary scale). The performance results for the rankings of run #3 are given in Table 9.

Ranking	Prec@10	NDCG@10	nERR
Ranking 21	0.72+/-0.30	0.69+/-0.21	0.79+/-0.22
Ranking 22	0.68+/-0.31	0.66+/-0.23	0.77+/-0.24
Ranking 23	0.71+/-0.3	0.70+/-0.2	0.82+/-0.2
Ranking 24	0.66+/-0.30	0.63+/-0.21	0.75+/-0.25
Ranking 25	0.70+/-0.30	0.67+/-0.22	0.78+/-0.23
Ranking 26	0.57+/-0.32	0.58+/-0.22	0.75+/-0.24
Ranking 27	0.65+/-0.30	0.64+/-0.20	0.78+/-0.22
Ranking 28	0.68+/-0.30	0.66+/-0.22	0.79+/-0.23
Ranking 29	0.67+/-0.31	0.66+/-0.22	0.80+/-0.22
Ranking 30	0.63+/-0.32	0.63+/-0.23	0.78+/-0.22

Table 9: Mean performance and standard deviation for the rankings in run #3.

The results of the two-way ANOVA test are given in Table 10. Again, the system effect is significant. The significant pairs according to Tukey's HSD test are listed in Table 11. The difference of Ranking 26 and all other systems is significant. The two best performing rankings are the "text statistics only" ranking (Ranking 21) and the learned model (Ranking 23), which is shown in Figure 3. For both

systems the differences to Ranking 24, Ranking 26, Ranking 27 and Ranking 30 are statistically significant. The only significant difference for the EconBiz baseline is achieved with respect to the worst performing system (Ranking 26). The difference between the two baselines is not statistically significant. Hence, again none of the other test rankings achieved a significantly better performance than the two baseline systems.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IR_System	9	1.8	0.20	16	0
SearchTask	161	56.5	0.35	27	0
Residuals	1449	18.7	0.01	NA	NA

Table 10: Result of the two-way ANOVA test. Given are the degrees of freedom (Df), the sum of squares (Sum Sq), the mean sum of squares (Mean Sq), the F value, and the p-value (Pr(>F)).

	diff	lwr	upr	p adj
Ranking 24-Ranking 21	-0.07	-0.11	-0.03	0.00
Ranking 26-Ranking 21	-0.12	-0.16	-0.08	0.00
Ranking 27-Ranking 21	-0.05	-0.09	-0.01	0.00
Ranking 30-Ranking 21	-0.06	-0.10	-0.02	0.00
Ranking 26-Ranking 22	-0.08	-0.12	-0.04	0.00
Ranking 24-Ranking 23	-0.07	-0.11	-0.03	0.00
Ranking 26-Ranking 23	-0.12	-0.16	-0.08	0.00
Ranking 27-Ranking 23	-0.05	-0.09	-0.01	0.00
Ranking 30-Ranking 23	-0.06	-0.10	-0.02	0.00
Ranking 25-Ranking 24	0.05	0.01	0.09	0.01
Ranking 26-Ranking 24	-0.05	-0.09	-0.01	0.00
Ranking 26-Ranking 25	-0.10	-0.14	-0.06	0.00
Ranking 30-Ranking 25	-0.04	-0.08	0.00	0.02
Ranking 27-Ranking 26	0.07	0.03	0.11	0.00
Ranking 28-Ranking 26	0.09	0.05	0.13	0.00
Ranking 29-Ranking 26	0.08	0.04	0.12	0.00
Ranking 30-Ranking 26	0.05	0.01	0.09	0.00

Table 11: Pairs of systems with statistically significant differences. Given are the mean difference (diff), the boundaries of the confidence interval (lwr, upr) and the adjusted p-value (p adj).

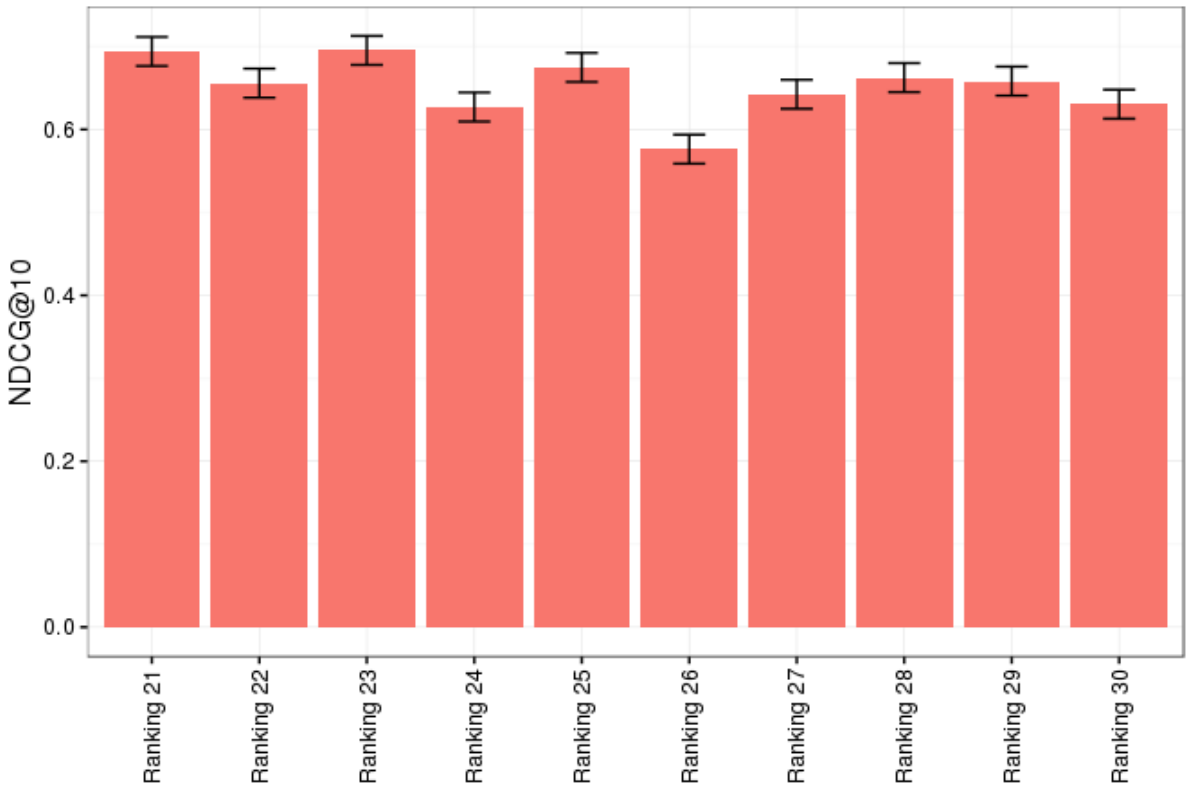


Figure 3: Comparison of the mean performances of the systems in run #3. Error bars indicate the simultaneous 95% confidence intervals obtained from Tukey's HSD test.

3. Discussion

The performances of the tested systems are all almost the same. Only very few significant differences could be found. However, this might be due to several reasons, for example, the results produced by the different rankings could be similar to each other, which would of course imply similar performance; missing data could either lead to similar results, since only a few records are affected by the new ranking factors. In the following we discuss five issues that may have influenced the data analysis to a certain degree.

3.1. Intra-assessor consistency

Intra-assessor consistency refers to the reliability of the judgements of one assessor. Scholer, Turpin, & Sanderson (2011) used the judgments on duplicate records in various TREC test collections to assess the self-consistency of the TREC assessors. Since numerous duplicates have been present in the document pools of run #1 and run #2, in this section we will evaluate the self-consistency of the subject indexers.

In the following we will focus on the relevance judgements from run #2. Further, for simplicity we will only consider pairs of duplicate records, i.e., we neglect duplicates with records from three or more source databases.

In total, there were 469 pairs of duplicate records to be assessed. The assessors were not given any instructions how to judge duplicate records. Since the second record of a pair does not provide any new information, the assessors might have judged the second record consistently worse than the

first record. To check this, we determined the difference between the two judgments for each pair with respect to the order in which the records have been judged. The mean difference is 1.5 ± 18. Hence, the data does not provide evidence to conclude that the second record has been judged systematically different than the first one.

Figure 4 shows the distribution over the absolute differences. Many duplicates have been judged very similar. However, for a few duplicates the difference between the two judgements is rather large. Hence, one could state that these records have been judged are rather diametrically. The mean absolute difference is 9.6 ± 15.

	ja	nein
Ja	282	16
Nein	14	157

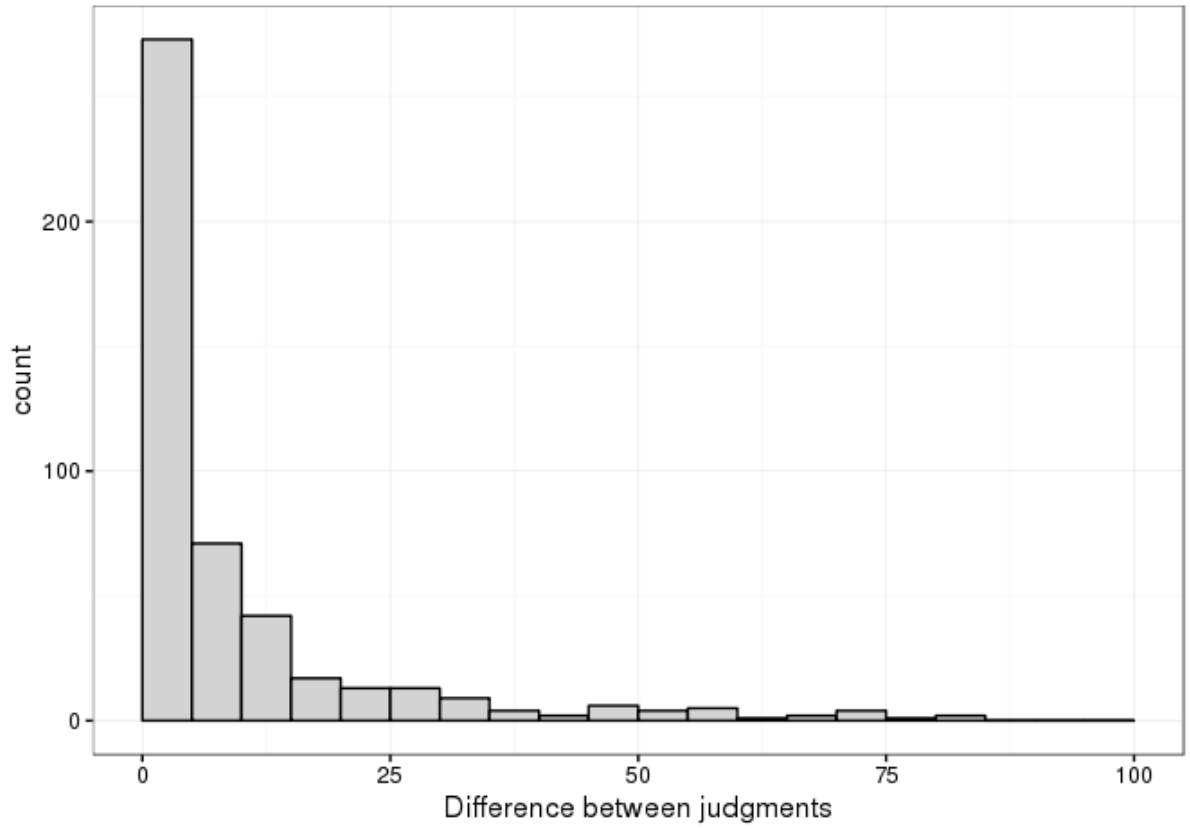


Figure 4: Histogram over the absolute judgement differences between duplicates.

If the difference is larger than some threshold, we consider the pair as inconsistently judged. If we set the threshold to 10, the percentage of inconsistent pairs is with about 25% rather high. Figure 5 shows the degree of inconsistency in dependence of the threshold. With respect to the binary relevance judgments the fraction of inconsistent pairs is only 6.4%. When only considering duplicates for that at least one document was judged relevant, the fraction is 9.6%. Scholer et al. (2011) reported that the self-consistency of the binary relevance judgments across various TREC collections is between 15% and 19%. Hence, compared to the TREC assessors we can consider the self-consistency of the subject indexers as relatively high (with respect to binary judgments). However, consistency for gradual judgments on the slider scale seems to be more difficult to achieve.

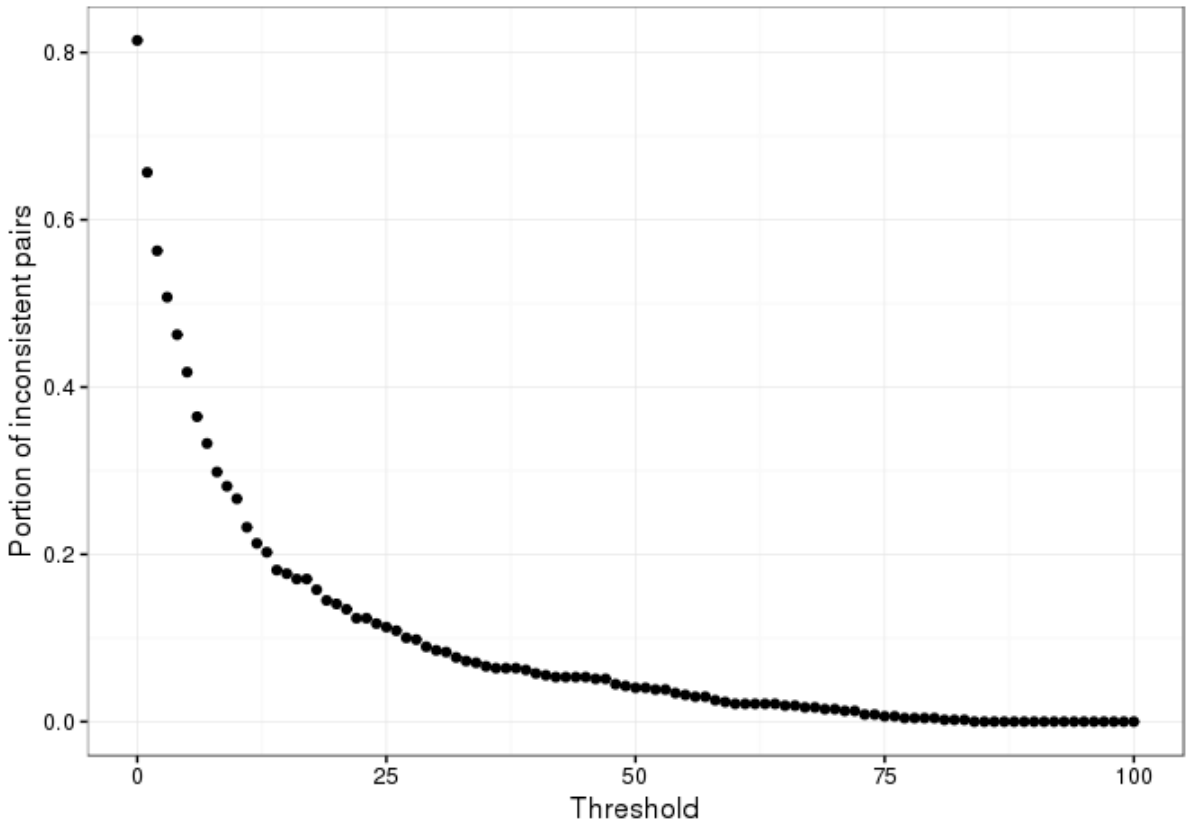


Figure 5: Fraction of pairs inconsistent in dependence of the threshold.

We assume that the gradual judgment of every document is similarly inaccurate. In principle, this uncertainty will give rise to an uncertainty in the per-task performance scores and hence to an additional uncertainty in the systems' mean performance scores. This uncertainty will further lower the statistical significance of the differences between systems. However, we did not determine this uncertainty formally.

3.2. Inter-assessor agreement

The tasks used in run #3 were partly used in the previous runs.Thus, for several tasks we have two sets of judgments:- one set from the subject indexers and another set from the undergraduate students. The two groups agree in their binary relevance judgements for about $66 \pm 19\%$ of the documents per task (see also Table 12). Compared to agreement rates reported in the literature, this is a relatively high rate (Saracevic, 2007). However, the mean expected agreement rate by chance is $60 \pm 17\%$.

	relevant	non-relevant
relevant	747	388
non-relevant	301	545

Table 12: Agreement between subject indexers and undergraduate students with respect to binary relevance (over all tasks).

Hence, we also use Cohen's κ to evaluate the agreement between the two assessor groups. Figure 6 shows the cumulative distribution of Cohen's κ over the tasks. The mean is 0.19 ± 0.25. For most tasks the agreement is statistically significantly different from agreement by chance ($\alpha = 0.05$).

While there is no commonly agreed interpretation of the values of Cohen's κ, we consider the agreement for most tasks as relatively low. Bailey, Craswell, Soboroff, Vries, & Yilmaz (2008) reported very similar agreement levels.

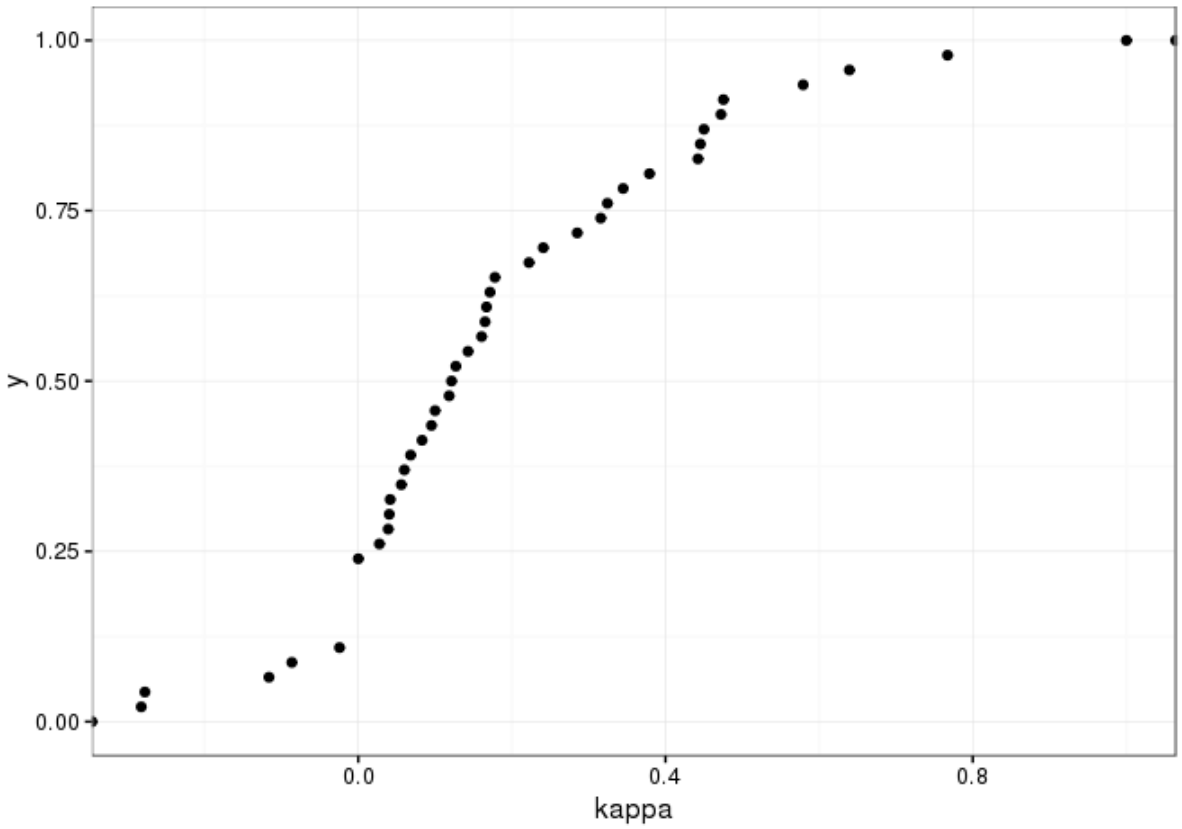


Figure 6: Cumulative distribution of Cohen's κ.

The evaluation of inter-assessor agreement with respect to the gradual judgements is not readily possible, though. When using relevance judgements based on an ordinal scale, it seems to be common to assume that these are comparable between assessors (probably due to explicit judging guidelines) (e.g. Bailey et al., 2008). This assumption is even more stressed if the relevance judgments are elicited via a continuous scale, e.g., because assessors might differ in the range of values they effectively use for judging (Kazai, Craswell, Yilmaz, & Tahaghoghi, 2012).

Despite these potential individual differences, one might consider the relevance judgments as comparable [@...]. However, our data exhibits another artifact with respect to run #3 compared to the other two runs. Figure 7 shows the distributions of the gradual judgements for each run and for relevant and non-relevant (binary relevance) documents (over all tasks). Clearly, the range used for relevant and non-relevant documents in run #3 differs from the ranges used in the other two runs. While this could be in principle due to differences in the assessor groups, we speculate, that this is due to the altered order of the assessment of binary and gradual relevance in the third run. Apparently, the swap had the effect that the undergraduate students used another threshold when differentiating between relevant and non-relevant documents.

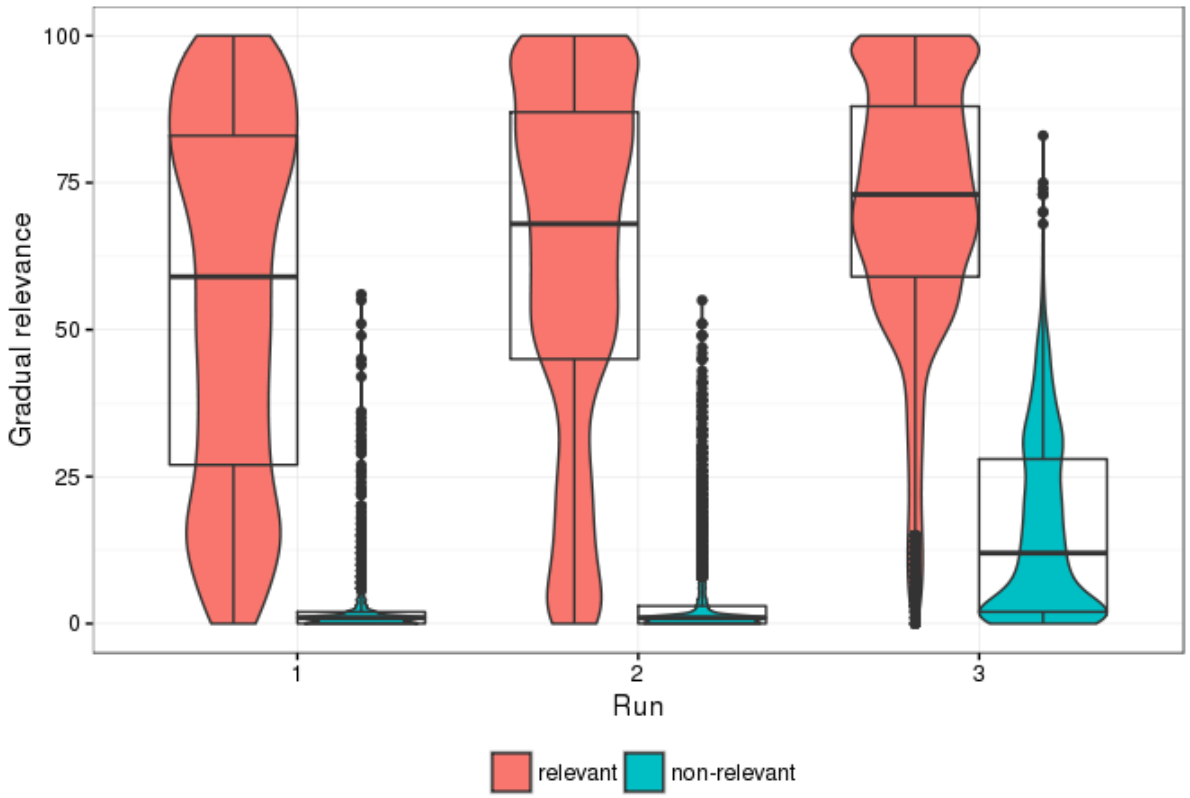


Figure 7: Distribution of gradual judgments for relevant and non-relevant documents.

In effect, the range of values used to judge non-relevant in the third run is much larger than in the previous runs, and, accordingly, the range used to judge relevant documents is more compressed. Therefore, especially the comparison of the gradual relevance of non-relevant documents is problematic, since in the first two runs almost all judgements are conflated into one value, i.e., the various degrees present in the judgments in run #3 are evened out in the other two runs.

Hence, as a very rough measure for inter-assessor agreement we only determined the intersection between the top 10 documents with regard to each judgement set. The mean intersection is 5.9 ± 1.7 and 4.8 ± 2.4 for run #1 and run #2, respectively. Thus, we conclude that both groups do only agree in part about what the most relevant documents are. Since we realized this issue in a rather late project phase, we encourage future analyses to apply more sophisticated approaches to evaluate the inter-assessor agreement between both groups.

3.3. Task ease

From the relevance judgments for a specific task we could determine the probability distribution of the performance for any possible ranking (with regard to the pool). In order to determine this probability distribution we built a random sample of 10,000 permutations from the relevance judgments and calculated nDCG@10 for each permutation. Figure 8 shows the probability distributions for two exemplary tasks. The distributions are quite different ranging from low mean and positive skewness to high mean and negative skewness.

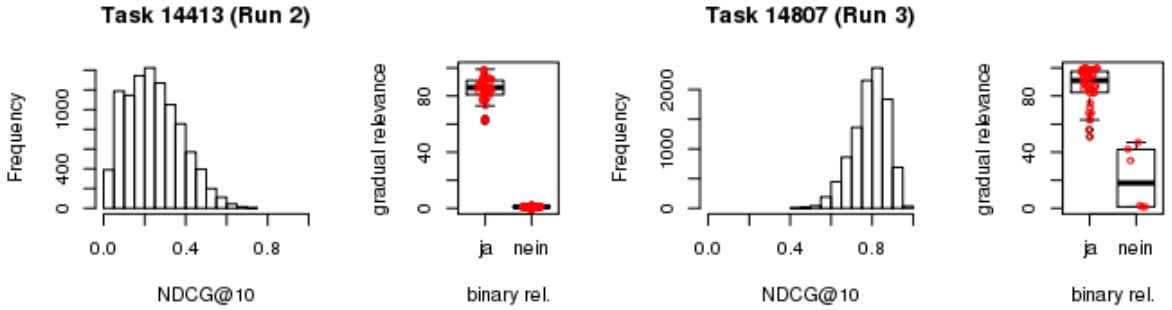


Figure 8: Probability distribution of nDCG@10 and distribution of the relevance judgments for two exemplary tasks.

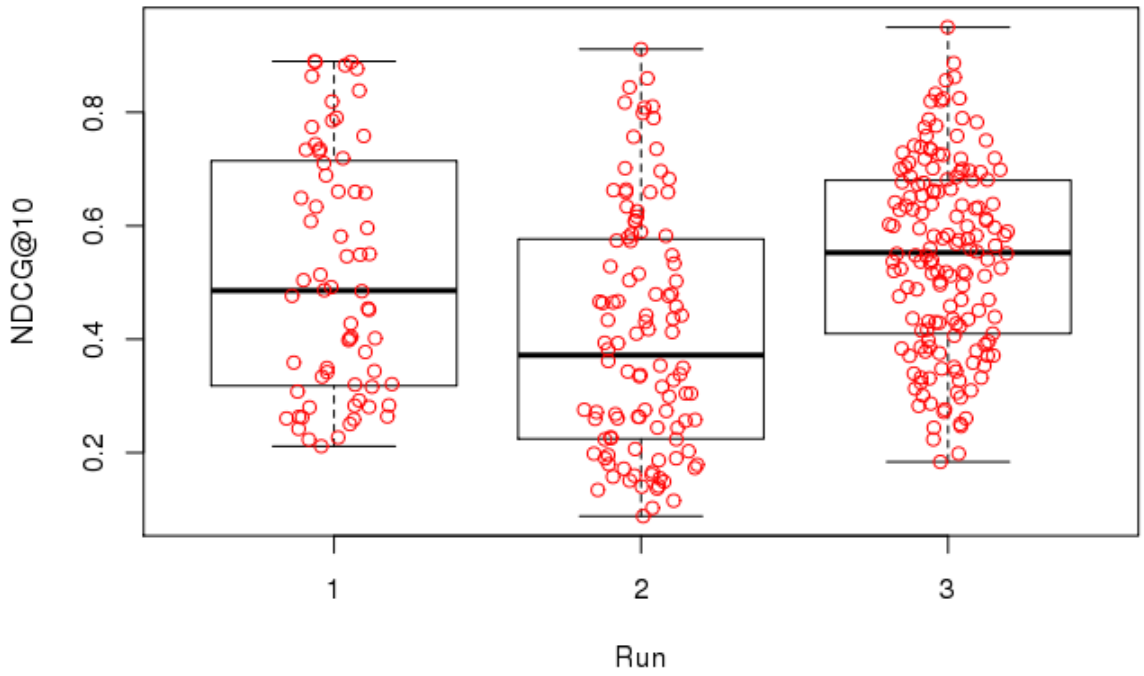


Figure 9: Distribution of the expectation values for random sorting over the tasks.

Figure 10 shows the per-task mean scores over the test rankings from run #2 over the expected per-task scores. Roughly stated, rankings achieved a high actual score for tasks with a high expected score. However, the rankings can also achieve high actual scores for tasks with a low expected score. While the per-task mean performance is a common measure for the *actual task difficulty* (with respect to a set of reference systems) [@...], we refer to the expected nDCG scores as the *a-priori task ease*.

The correlation of the per-task mean scores and the a priori task ease is $r = 0.74$ ($r = 0.7$ for run #1 and $r = 0.77$ for run #3). The pattern for the individual systems is similar and the correlation ranges from 0.59 to 0.76, while Ranking 11 shows a somewhat different behavior having a correlation of only 0.48. Hence, the variability in the performance over tasks observed in section 2 is to a large extent due to the variability of the tasks in the test collection with respect to task ease. This also

illustrates that absolute performance scores are only meaningful in the context of a test collection since the magnitude of the absolute scores is strongly dependent of the characteristics of the test collection.

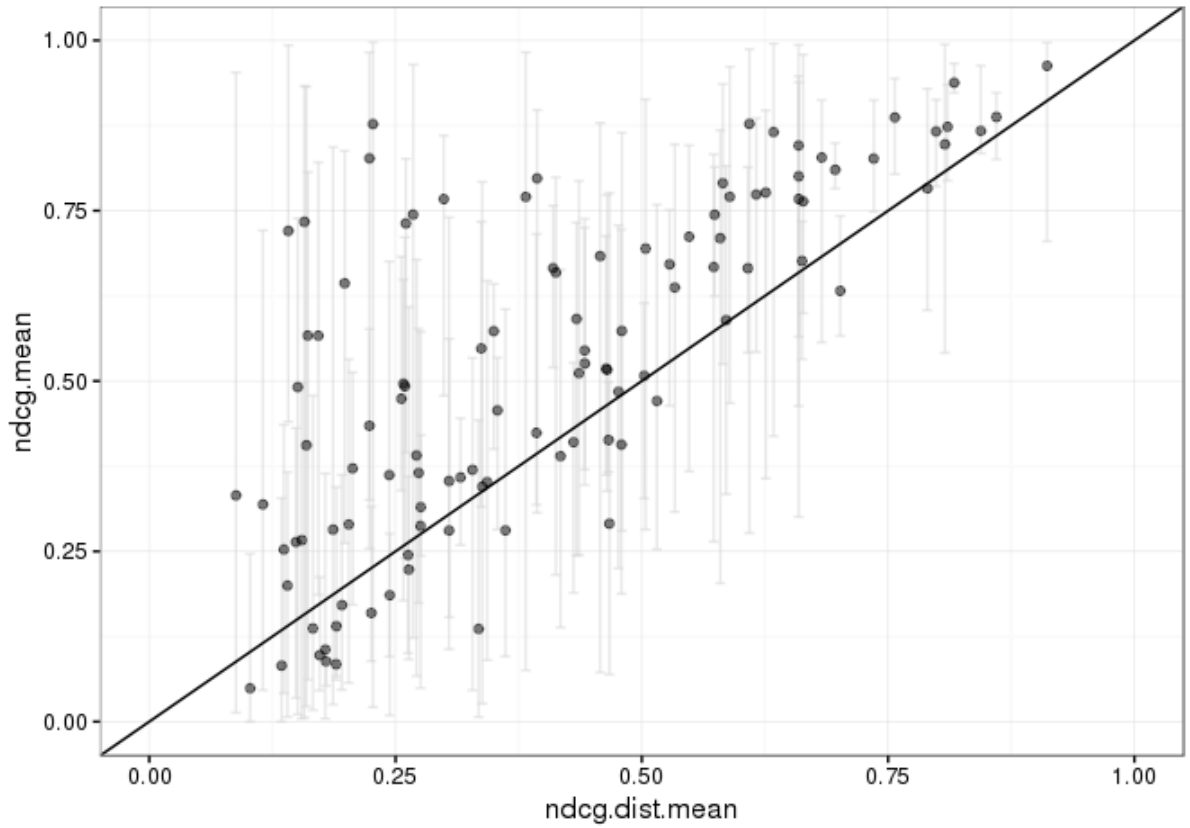


Figure 10: Relation between per-task mean scores and a priori task ease. Error bars indicate the minimum and maximum score achieved for a task.

Another issue related with task ease is the variability of the dispersion of system scores for the tasks. Note that the variance of the per-task probability distributions is negatively correlated with task ease. Hence, the performance of the systems tends to be more similar for easy tasks. This is also visible for the actual dispersion of the system scores in Figure 10. However, low dispersion can also occur for non-easy tasks. Still, the presence of easy tasks can therefore potentially lead to a reduction of the differences between systems. However, a preliminary effort to evaluate this effect indicated that the effect is probably small with respect to the large confidence intervals.

Finally, we note that *a-priori task ease* is related to a simple statistic of the relevance judgement distribution. Figure 11 shows the high correlation between the expected nDCG@10 scores and the ratio of the mean of the relevance judgments and the mean of the top-10 judgments ($r = 0.99$). A high ratio indicates that the judgments of the top-10 documents are not very different from the other judgements, i.e., the relevance judgments only contain very little discriminating information with respect to the most relevant documents.

The *a-priori task ease* might be influenced by two aspects of the experimental design: (1) the bounded assessment scale might limit the discrimination between the most relevant and non-relevant documents, for example, if the threshold between relevant and non-relevant documents is rather high as it is in run #3, then the most relevant document might only be judged a factor of two or three better than a marginally relevant document. Thus, eliciting relevance judgments via

magnitude estimation (Maddalena, Mizzaro, Scholer, & Turpin, 2015) or via preference judgments (Carterette, Bennett, & Dumais, 2008) might decrease the effect of *a-priori task ease*; (2) in general, a larger pool will contain a higher fraction of non-relevant documents. Hence, increasing the pool size might decrease the mean of the relevance judgments and therefore also decrease the effect of *a-priori task ease*.

Moreover, the absolute magnitude of the *a-priori task ease* might be decreased by these two adjustments. Due to this dependence however the absolute magnitude of the *a-priori task ease* might be considered as an artifact of the experimental design. Nevertheless, the relative differences between the tasks might be intrinsic to the tasks. Note that a preliminary analysis indicated that differences in *a-priori task ease* might also be present in the relevance judgments of the TREC 2014 Web Ad-hoc track.

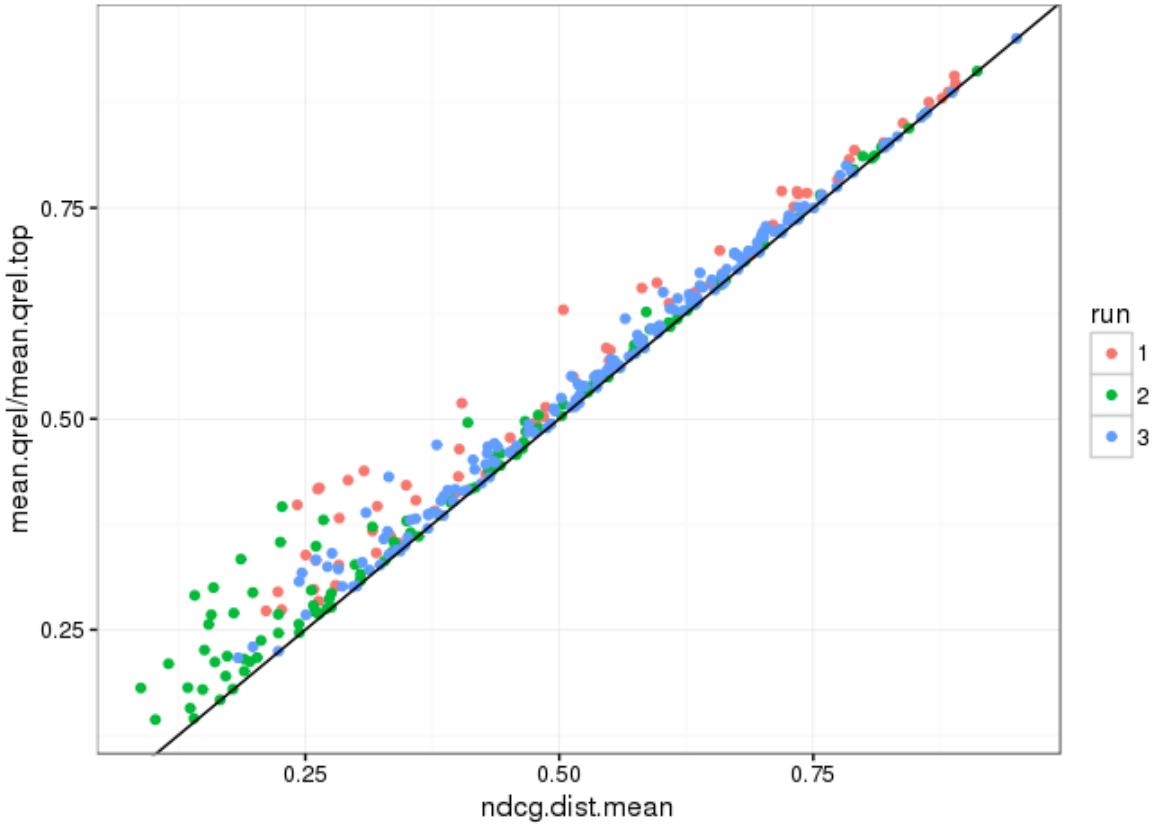


Figure 11: Relationship between task ease and the relevance judgment distribution.

3.4. System-task interaction

In the results of the ANOVA test in section 2, we observed a small system effect compared to the large residual deviations. This indicates that a system is more effective for some tasks, but less effective for other tasks in comparison to other systems. To further illustrate this, Figure 12 shows the cumulative distribution over the per-task differences between Ranking 11 and the other systems of run #2. While for many tasks the absolute difference is quite low (say less than 0.1), for other tasks large positive differences as well as large negative differences can be observed. For most other pairs of test rankings the distributions are quite similar.

Tasks with degraded performance are one limiting factor for the differences in the mean performance. However, this is a commonly observed pattern in retrieval evaluation studies and this is actually the ongoing research objective of the risk-sensitive task in TREC.

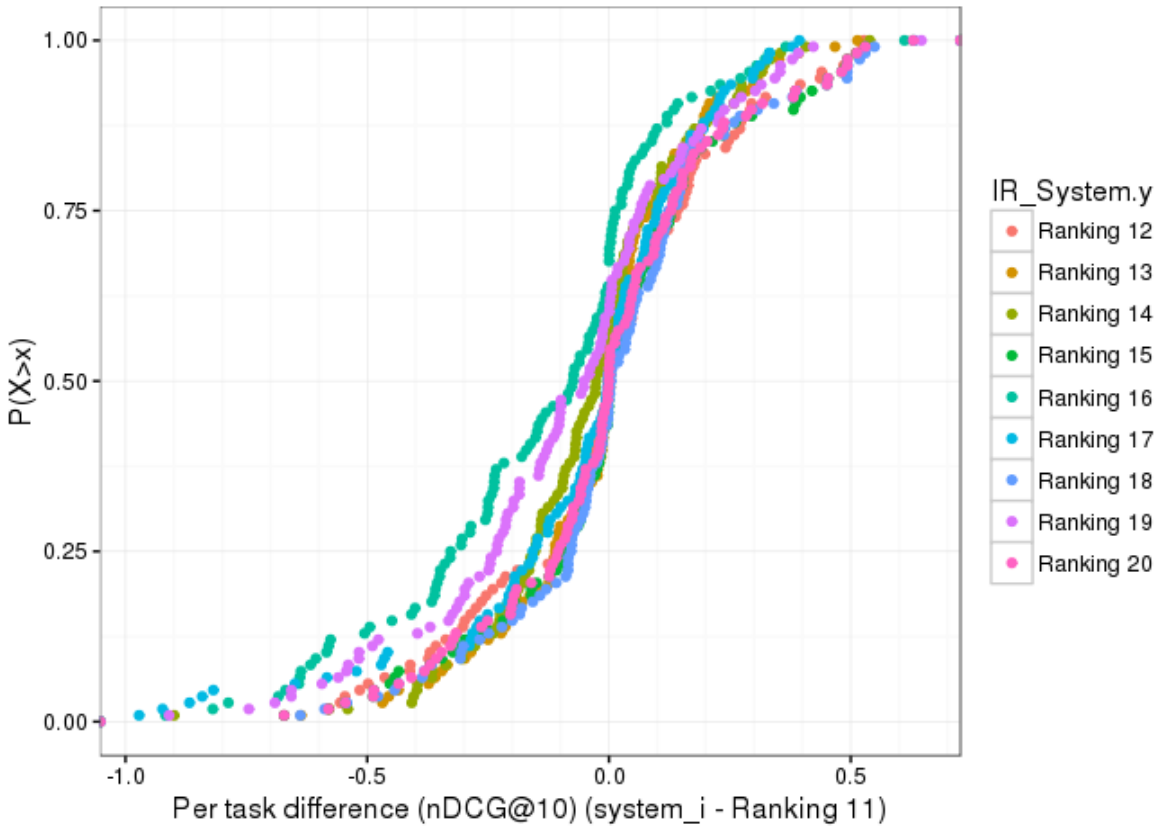


Figure 12: Cumulative distribution over the per-task differences between Ranking 11 and all other systems from run #2.

To further investigate the per-task differences, Figure 13 shows a scatter plot of the per-task differences of Ranking 11 and Ranking 18 over the per-task performance of Ranking 11. There is a slight negative correlation indicating that Ranking 18 tends to improve more on tasks for which Ranking 11 only achieved a low performance ($r = -0.39$). However, this correlation should not be overrated because the correlation is biased, as high performing tasks can only be marginally improved and low performing tasks can only be marginally degraded.

Nevertheless, when aggregating the per-task scores or the per-task differences by the arithmetic mean, all tasks are given equal weight. This has been criticized since one might prefer an improvement of a low performing task over a degradation of a high performing task. Thus, the geometric mean has been suggested as an alternative aggregation scheme, which gives more weight to low performing tasks. Thus, we calculated the ϵ-adjusted geometric mean with $\epsilon = 0.01$ (Ravana & Moffat, 2009; Robertson, 2006). However, results are very similar to the arithmetic mean, probably due to the relative low frequency of very low scores.

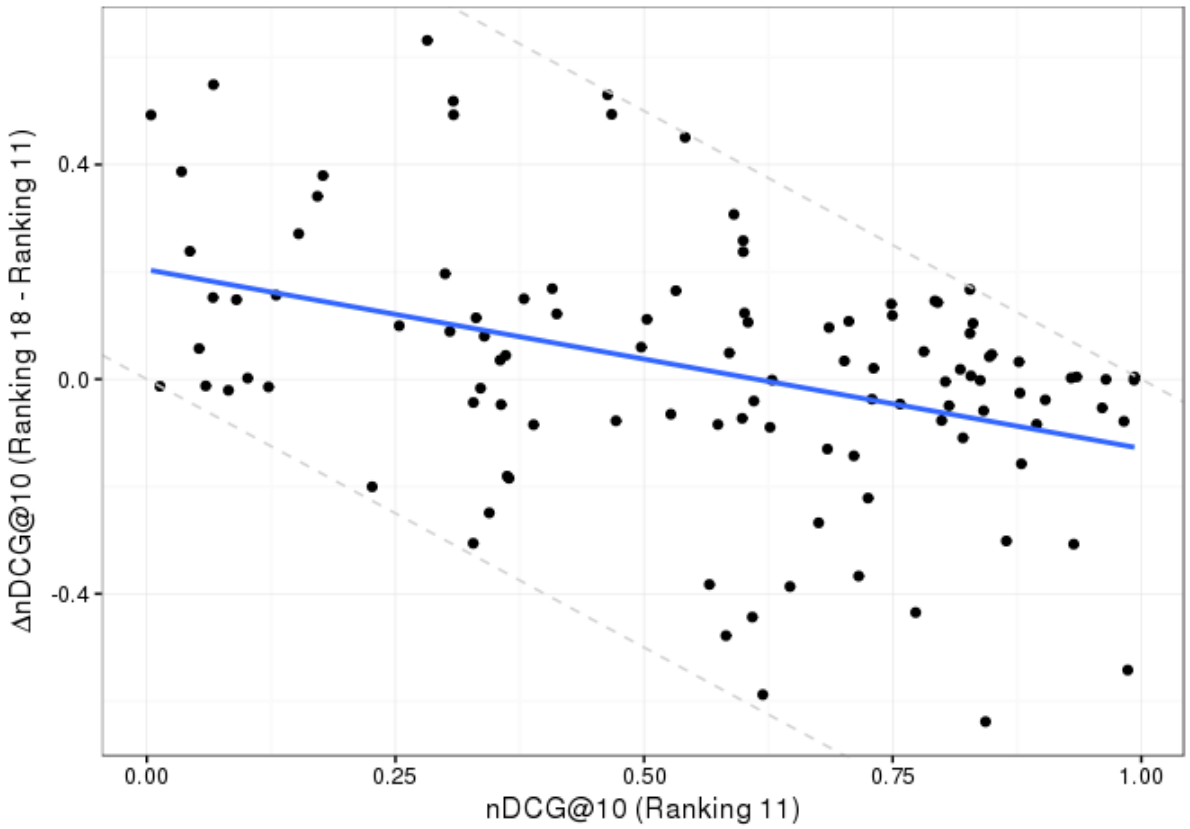


Figure 13: Per-task differences between Ranking 11 and Ranking 18 in dependence of the performance of Ranking 11.

3.5. Result list similarity

In section 2 we observed that the mean performance of the different rankings is very similar. While the per-task performance shows a high variance, there are still many tasks with very similar performance. Accordingly, the question arises whether the reason for this is that the different rankings produced similar result lists for those tasks.

We use a generalized version of Spearman's footrule distance to measure the differences between two result lists. Spearman's footrule distance is defined as the L_1 distance between two permutations, i.e., two lists containing the same elements. Fagin, Kumar, & Sivakumar (2003) proposed the *footrule distance with location parameter* l as a generalization to top-k lists, which is obtained "by placing all missing elements in each list of the lists at position l and computing the usual footrule distance between them". In the following we use the footrule distance with location parameter l with $l = k + 1$, normalized on the maximal distance.

Figure 14 shows the distribution of the Spearman's footrule distance for each pair of ranking from run #2 over the tasks. Ranking 12, 15, 18, and 20 are all pairwise similar and thereby form a small cluster. The mean distance for these rankings ranges from 0.07 (Ranking 15 vs. Ranking 20) to 0.28 (Ranking 12 vs. Ranking 18). The other ranking pairs have distances larger than 0.46.

The EconBiz baseline ranking (Ranking 12) and Ranking 15 (having a mean distance of 0.24) are conceptually very similar, since in both models the correlation of text statistics and recency is the main contributing factor. Ranking 18 and Ranking 20 also incorporate recency, but equally weighted with other factors. However, the utility function for recency is constructed differently than the CSS based utlity functions of the other factors. Thereby, the comparison of recency with other CSS based

factors might be biased and, therfore, the weights do not necessarily represent the relative importance,i.e., even if the weights are equal, recency might effectively dominate the other factors.

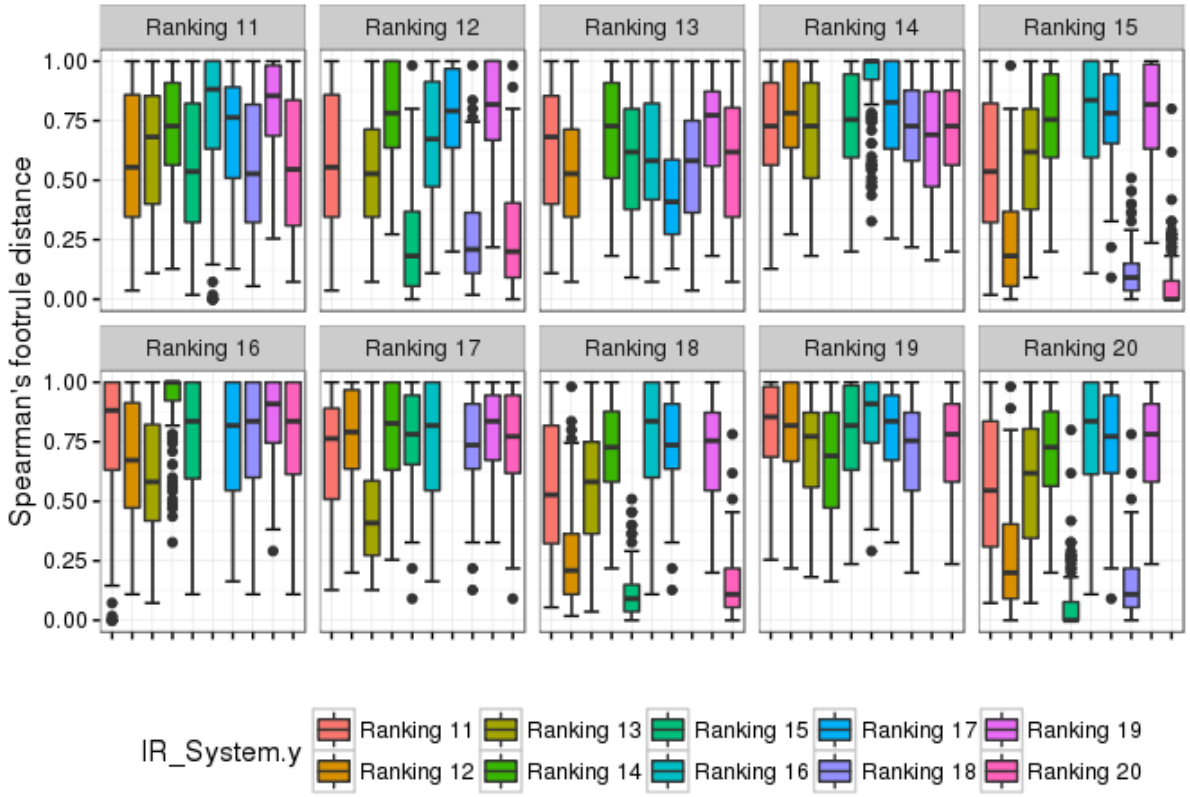


Figure 14: Similarity of result lists for pairs of systems from run #2.

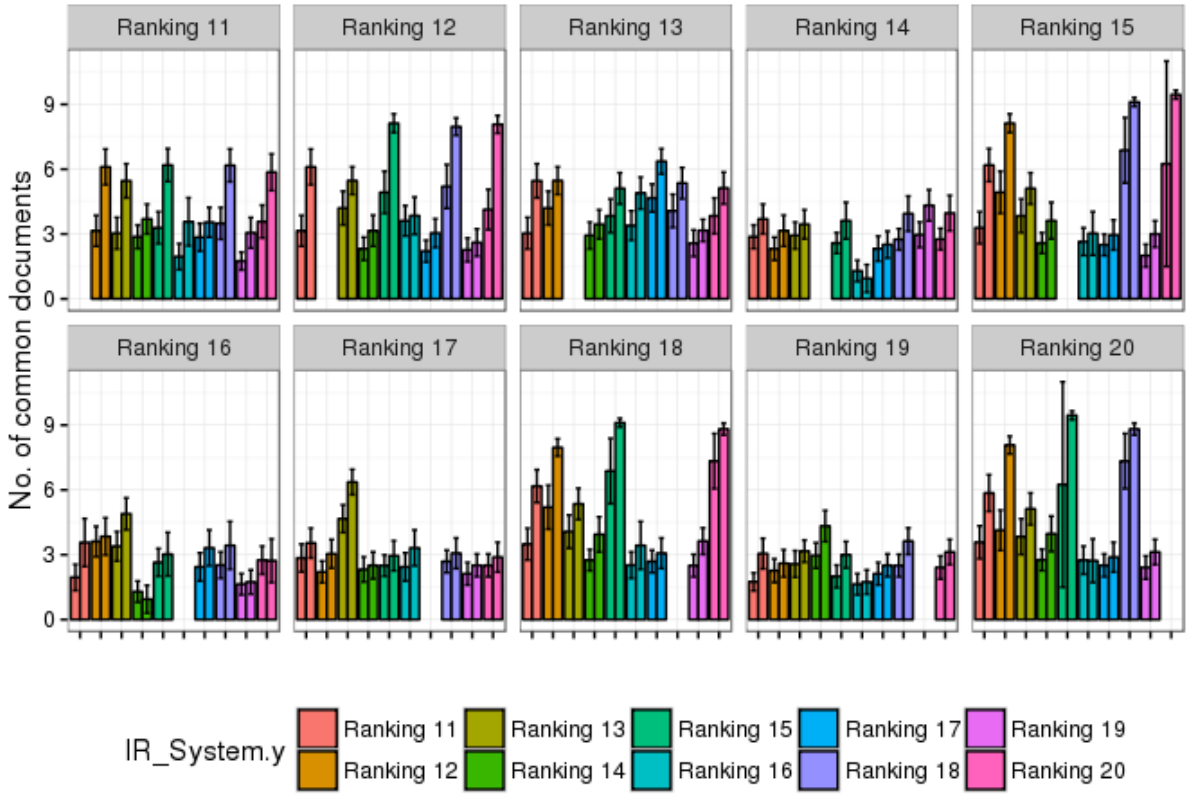


Figure 15: Mean number of common documents in the result lists for pairs of rankings from run #2 for tasks with diverging performance (left bar) and similar performance (right bar). Error bars indicate the 95% confidence interval of the mean.

Since the number of common documents obviously only measures what is common between the result lists, the last question is, whether the small performance differences are due to that only non-relevant documents have been exchanged between the result lists. Figure 16 shows the mean exchanged precision, i.e., the average over the fraction of relevant documents (binary relevance) among the exchanged documents.

For most of the rankings the mean for tasks with similar performance is either equal to or larger than the mean for tasks with diverging performance. However, for most rankings the difference of the means is not statistically significant. Thus, for tasks with similar performance the relevance of exchanged documents tend to be at least not inferior to the relevance of the exchanged documents for tasks with diverging performance (in the mean).

While these observations are somewhat indirect, we are more confident that the reason why several tasks achieve a similar performance by the various rankings is not because these rankings generated near identical result lists. Further, the differences in the result lists are not due to only non-relevant documents. Therefore, we are more confident that tasks with similar performance often contain different relevant documents, but these documents have been judged similarly.

We note that at the time of the analysis we were not aware of more recent suggestions to measure the similarity of top-k lists, e.g., rank-biased-overlap (RBO) (Webber, Moffat, & Zobel, 2010) or information difference (Golbus & Aslam, 2013). These measures might be better suited for the problem at hand, but we leave this for future analyses.

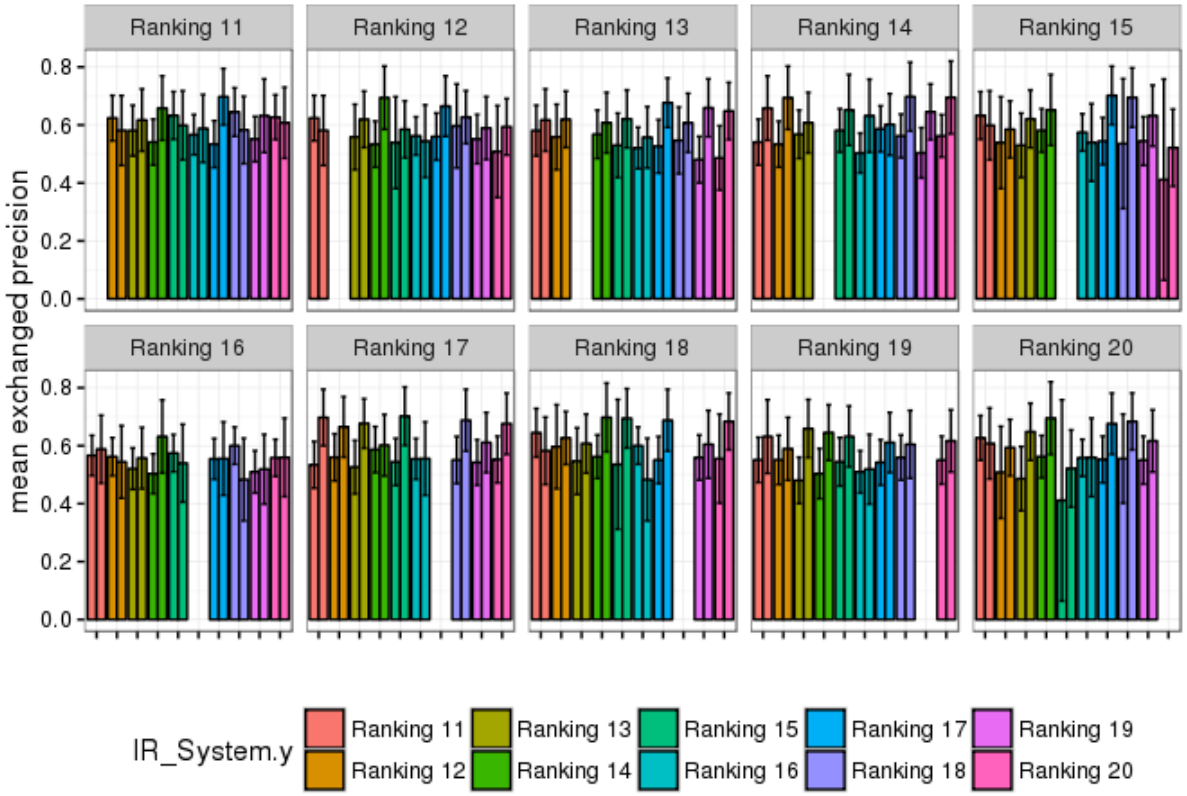


Figure 16: Mean exchanged precision.

4. Conclusion

The mean performances of the tested rankings vary only slightly. None of the test rankings achieved a statistically significant improvement over the two baseline rankings. The variability of the performance over tasks is large. This is a common observation in retrieval evaluation studies. Further system-task interaction is also large. Hence, a system is more effective than other systems on some tasks, but is less effective on other tasks. Several tasks achieved similar performance for pairs of rankings. However, for most pairs of systems the corresponding result lists are at least to some extent different.

The relevance judgments contain a several amount of uncertainty. We have only shown that the judgments between the subject indexers and the undergraduate students differ. However, also assessors from the same user group substantially disagree in their judgments. (Actually, this also applies even to topical relevance. Hence, even topical relevance is subjective.) Both effects introduce an uncertainty into performance measurements. Hence, this uncertainty places an upper bound on the achievable performance. The agreement rate varies over tasks. Hence, the upper bound also varies over tasks.

The most common problem found by the failure analysis from the RIA workshop was that systems did miss an aspect of the information need. This was seen as a failure of the systems. However, at least in some cases the missing aspect was not present in the actual query.

5. References

Alonso, O. (2013). Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, *16*(2), 101–120. doi:10.1007/s10791-012-9204-1

Alonso, O., & Lease, M. (2011). Crowdsourcing for information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11* (p. 1299). New York, New York, USA: ACM Press. doi:10.1145/2009916.2010170

Bailey, P., Craswell, N., Soboroff, I., Vries, A. P. De, & Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. *In SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 667–674. doi:10.1145/1390334.1390447

Behnert, C., & Lewandowski, D. (2015). Ranking Search Results in Library Information Systems — Considering Ranking Approaches Adapted From Web Search Engines. *The Journal of Academic Librarianship*, *41*(6), 725–735. doi:10.1016/j.acalib.2015.07.010

Busa-Fekete, R., Szarvas, G., Élteto, T., & Kégl, B. (2012). An apple-to-apple comparison of Learning-to-rank algorithms in terms of Normalized Discounted Cumulative Gain. *20th European Conference on Artificial Intelligence (ECAI 2012) : Preference Learning: Problems and Applications in AI Workshop*. Ios Press.

Carterette, B., Bennett, P. N., & Dumais, S. T. (2008). Here or There. *Advances in Information Retrieval*, *4956*. doi:10.1007/978-3-540-78646-7

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, *17*(1), 134–160. doi:10.1137/S0895480102412856

Golbus, P. B., & Aslam, J. A. (2013). A Mutual Information-based Framework for the Analysis of Information Retrieval Systems. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 683–692). doi:10.1145/2484028.2484073

Kazai, G., Craswell, N., Yilmaz, E., & Tahaghoghi, S. M. . (2012). An analysis of systematic judging errors in information retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12* (p. 105). New York, New York, USA: ACM Press. doi:10.1145/2396761.2396779

Maddalena, E., Mizzaro, S., Scholer, F., & Turpin, A. (2015). Judging relevance using magnitude estimation. In *Advances in Information Retrieval* (pp. 215–220). Springer.

Ravana, S. D., & Moffat, A. (2009). Score Aggregation Techniques in Retrieval Experimentation. In *Conferences in Research and Practice in Information Technology Series* (Vol. 92, pp. 59–67).

Robertson, S. (2006). On GMAP: and other transformations. In *Proceedings of the 15th ACM international conference* (pp. 78–83). doi:10.1145/1183614.1183630

Sakai, T. (2014). Designing test collections that provide tight confidence intervals. *Forum on Information Technology 2014 RD*.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. doi:10.1002/asi.20681

Schaer, P. (2013). *Der Nutzen informetrischer Analysen und nicht-textueller Dokumentattribute für das Information Retrieval in digitalen Bibliotheken*.

Scholer, F., Turpin, A., & Sanderson, M. (2011). Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1063–1072. doi:10.1145/2009916.2010057

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), 1–38. doi:10.1145/1852102.1852106

6. Appendix

Rankings

	Run #1										Run #2										Run #3									
Ranking factor \ Ranking	**R1**	**R2 EB**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12 EB**	**R13**	**R14**	**R15**	**R16**	**R17**	**R18**	**R19**	**R20**	**R21**	**R22 EB**	**R23 LtR**	**R24**	**R25**	**R26**	**R27**	**R28**	**R29**	**R30**
Overall relative weight of query independent features over query dependent features	0	1									0	1	1,000,000								0	1	8	1,000,000						
TEXT STATISTICS	**X**	(x)									**X**	(x)									**X**	(x)	(x)							
POPULARITY				**X**										**X**										**X**						
Usage			14	25				50	50				14	25				50	50					25				17	14	30
Number of views																							85							
Number of accesses/clicks																							64							
Number of clicks on further content																							103							
Number of captures																							627							
Number of loans																							6							
Purchasing behavior			14	25									14	25										25						
Number of editions																														
Number of owning libraries																							0.4							
Publisher authority			14	25									14	25										25						
Number of items of a publisher																							17							
Peer reviewed status																							30							
References			14	25					50	50			14	25					50	50				25				17	43	30
Number of citations																							14							
Citation impact of journal																							7							
Author metrics																							15							
FRESHNESS					X										X										X					
Publication date		1	14		100			50		50		1	14		100			50		50		1	103		100			50	29	10

LOCALITY & AVAILABILITY						**X**										**X**										**X**				
Availability		1	14			100					1	14				100					1					100		14	14	30
CONTENT PROPERTIES							**X**										**X**										**X**			
Additional information			14				100					14					100										100			
Added value																						26								
Metadata quality		0.0001									0.0001										0.0001	163								
File format																														
Document type (thesis)		1									1										1									

Table 13: Overview of ranking factors integrated in the test rankings R. The numbers describe the weighting of the particular factor; EB marks the baseline Ranking, LtR a learning-to-rank-option.