Guillaume Cabanac, Gilles Hubert, Mohand Boughanem, Claude Chrisment. Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. In Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010), Padua, Italy, 20/09/10-23/09/10, Maristella Agosti, Nicola Ferro, Carol Peters, Maarten de Rijke, F. Smeaton (Eds.), Springer, LNCS 6360, p. 112-123, Sept. 2010.
http://dx.doi.org/10.1007/978-3-642-15998-5_13

Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation*

Guillaume Cabanac, Gilles Hubert,
Mohand Boughanem, and Claude Chrisment

Université de Toulouse — IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
{cabanac, hubert, boughanem, chrisment}@irit.fr

Abstract. We consider Information Retrieval evaluation, especially at Trec with the trec_eval program. It appears that systems obtain scores regarding not only the relevance of retrieved documents, but also according to document names in case of ties (i.e., when they are retrieved with the same score). We consider this tie-breaking strategy as an uncontrolled parameter influencing measure scores, and argue the case for fairer tie-breaking strategies. A study of 22 Trec editions reveals significant differences between the Conventional unfair Trec's strategy and the fairer strategies we propose. This experimental result advocates using these fairer strategies when conducting evaluations.

1 Introduction

Information Retrieval (IR) is a field with a long tradition of experimentation dating back from the 1960s [1]. The IR community notably benefited from Trec evaluation campaigns and workshops. Since 1992, these have been offering researchers the opportunity to measure system effectiveness and discuss the underlying theoretical aspects [2, 3]. At Trec, evaluation results of IR systems (IRSs), a.k.a. search engines, are computed by the trec_eval [4] program. Many subsequent IR evaluation initiatives also rely on trec_eval, such as tasks of Ntcir [5], Clef [6], and ImageClef [7].

As a general rule when conducting a scientific experiment, one should identify all the parameters at stake and control all but one to be able to test its effect on the measured artifact. Controlling parameters is a key concern since conclusions may be biased when two or more parameters vary at the same time during the experiment. Following on from studies on IR evaluation methodology such as Voorhees's [8] and Zobel's [9] we identified an uncontrolled parameter in Trec through trec_eval: evaluation results not only depend on retrieved documents, but also on how they were named in case of ties (i.e., *ex aequo* documents). This is a major issue since 'lucky' ('unlucky') IRSs can get better (worse) results than they would deserve in an unbiased evaluation.

* This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

This paper is organized as follows. In Sect. 2, we present how IRSs are commonly evaluated according to the Cranfield paradigm. Then, we detail in Sect. 3 the issue we identified in Trec methodology, that we call the 'tie-breaking bias.' In Sect. 4, we propose alternative reordering strategies for canceling out the effect of the considered uncontrolled parameter. A comparison of the current Trec strategy with our proposal is detailed in Sect. 5. Our findings and limitations of our analyses are discussed in Sect. 6. Finally, related works are reviewed in Sect. 7 before concluding the paper and giving insights into research directions.

2 Tie-breaking Prior to Evaluating IRSs Effectiveness

This section introduces the concepts considered throughout the paper. Our brief description may be complemented by [10], which details IR evaluation at Trec and its realization with the trec_eval program. At Trec, at least one *track* a year is proposed. A track is comprised of 50+ *topics*; each one is identified with a qid. Participants in a track contribute at least one *run* file. Among the fields of this file, trec_eval only considers the following: qid, the document identifier docno, and the similarity sim of docno regarding qid, as provided by the IRS. In addition, the *query relevance judgments* file (i.e., *qrels*) results from manual assessment. The trec_eval program only considers the 3 following fields from it: qid, docno, and rel, which represents the relevance of docno regarding qid. The rel $\in [1, 127]$ value is assigned to relevant documents. Other values (e.g., rel $= 0$) represent non-relevant documents. Prior to computing effectiveness measures, trec_eval pre-processes the run file. Since it ignores its rank field, documents are reordered as follows: "internally ranks are assigned by sorting by the sim field with ties broken deterministicly (using docno)" [4]. Buckley and Voorhees comment this rationale and underline the importance of tie-breaking:

> "For TREC-1 ... Each document was also assigned a rank by the system, but this rank was deliberately ignored by trec_eval. Instead, trec_eval produced its own ranking of the top two hundred documents[1] based on the RSV [sim] values to ensure consistent system-independent tie breaking among documents that a system considered equally likely to be relevant (the ordering of documents with tied RSV values was arbitrary yet consistent across runs). Breaking ties in an equitable fashion was an important feature at the time since many systems had large number of ties—Boolean and coordination-level retrieval models could produce hundreds of documents with the same RSV." [10, p. 55]

Finally, trec_eval uses qrels and the reordered run to compute several effectiveness measures. In the remainder of this paper, let us consider a system s, a topic t, and a document d of the run, and the following measures: Reciprocal Rank $RR(s, t)$ of top relevant document, Precision at d cutoff $P(s, t, d)$, Average Precision $AP(s, t)$, and Mean Average Precision $MAP(s)$. Due to space limitation, we do not elaborate on these measures and refer the reader to [11, ch. 8] for a comprehensive definition.

[1] Since TREC-2, the top 1,000 documents is kept [10, p. 58].

The next section presents an issue related to the way document ties are broken at TREC. We argue that this issue makes the current tie-breaking strategy an uncontrolled parameter in IR experiments.

3 On How the Tie-breaking Bias Influences IR Evaluation

Let us consider, in Fig. 1(a), a sample of a run concerning the top 3 documents retrieved by an IRS for a given topic t (qid = 3). Suppose that 5 documents are relevant documents for t in the collection, including **WSJ5** (in bold). Since trec_eval ignores ranks it reorders the run by ascending qid, descending sim, and descending docno for tie-breaking purpose. The resulting document list is presented in Fig. 1(b) where the relevant **WSJ5** document is assigned rank #1. Notice that reciprocal rank is $RR(s,t) = 1$, precision at **WSJ5** is $P(s,t,\textbf{WSJ5}) = 1$ and $AP(s,t) = 1/5$. Now, without making any changes to the document contents, which still remain relevant for topic t, suppose that **WSJ5** had been named **AP8** instead. So, relevant document **AP8** is initially ranked #2 (i.e., the same position as **WSJ5**), as shown in Fig. 1(c). Then, due to the reordering process, LA12 remains ranked #1 by descending docno, remaining above **AP8**. Notice that reciprocal rank and average precision have been halved.

(a)

qid	docno	sim	rank
3	LA12	0.8	1
3	**WSJ5**	0.8	2
3	FT8	0.5	3

$\longrightarrow$ (b)

qid	docno	sim	$RR(s,t)$	$P(s,t,d)$	$AP(s,t)$
3	**WSJ5**	0.8	1	1	1/5
3	LA12	0.8		1/2	
3	FT8	0.5		1/3	

(c)

3	LA12	0.8	1
3	**AP8**	0.8	2
3	FT8	0.5	3

$\longrightarrow$ (d)

3	LA12	0.8	1/2	0	1/10
3	**AP8**	0.8		1/2	
3	FT8	0.5		1/3	

Fig. 1. Effect of document naming on reordered run and measure values

This minimal example illustrates the issue addressed in the paper: IRS scores depend not only on their ability to retrieve relevant documents, but also on document names in case of ties. Relying on docno field for breaking ties here implies that the *Wall Street Journal* collection (WSJ$\star$ documents) is more relevant than the *Associated Press* collection (AP$\star$ documents) for whatever the topic, which is definitely wrong. This rationale introduces an uncontrolled parameter in the evaluation regarding all rank-based measures, skewing comparisons unfairly. Let us justify our statement by considering the example of AP:

1) Tie-breaking effect on *inter-system* comparison, where $AP(s_1,t)$ of system s_1 and $AP(s_2,t)$ of system s_2 are considered for a given topic t. This comparison is unfair since AP values can be different although both the systems returned the same result $[R_{0.8}, N_{0.8}, N_{0.5}]$ where R_x is a relevant document (N_x is a nonrelevant document) retrieved with sim $= x$. This is the case when we associate the run in Fig. 1(a) with s_1, and the run in Fig. 1(c) with s_2. Indeed $AP(s_1,t) = 1/1 \cdot 1/5$ whereas $AP(s_2,t) = 1/2 \cdot 1/5 = 1/10$, thus showing a 200% difference.

2) Tie-breaking effect on *inter-topic* comparison, where we consider $AP(s, t_1)$ and $AP(s, t_2)$ of a single system for two topics t_1 and t_2. Such a comparison is made in TREC's *robust* [12] track for characterizing easy and difficult information needs. It is unfair since TREC reordering process may have benefited system s for t_1 (by re-ranking relevant tied documents upwards in the list) while having hindered it for t_2 (by re-ranking relevant tied documents downwards in the list). As a result, the IRS designers may conduct failure analysis to figure out why their system poorly performed on some topics. Poor results, however, may only come from misfortune when relevant documents are reorganized downwards in the result list only because of their names. Imagining that every relevant document comes from the AP collection, they will be penalized since they will be re-ranked at the bottom of the tied group when reordering by decreasing docno.

Breaking ties as currently proposed at TREC introduces an uncontrolled parameter affecting IR evaluation results. In order to avoid this tie-breaking issue, the next section introduces our proposal: alternative reordering strategies.

4 Realistic and Optimistic Tie-breaking Strategies

The current tie-breaking strategy (qid asc, sim desc, docno desc) introduces an uncontrolled parameter, as it relies on the docno field for reordering documents with the same sim value. Another strategy would be to randomize tied documents; this is not suitable as evaluations would be unfair and not reproducible (non deterministic). However, evaluations must measure how well a contribution performed, not how well chance benefited an IRS. Alternatively, relying on the initial ranks (from run) implies the same issue: IRS designers may have untied their run by assigning random ranks, as they were not able to compute a discriminative sim for those documents. As a result, random-based and initial rank-based approaches do not solve the tie-breaking issue.

In this section, we propose two tie-breaking strategies that are not subject to the bias presented in this paper. Figure 2 shows merged runs and qrels, as well as the result of current TREC Conventional strategy for reordering ties and the two strategies that we propose:

1. *Realistic reordering* stipulates that tied nonrelevant documents should come above relevant documents in the ranked list because the IRS was not able to differentiate between them. The reordering expression meeting this requirement is "qid asc, sim desc, rel asc, docno desc."

$$\text{Example: } [R_x, N_x, R_x] \xrightarrow[\text{qid asc, sim desc, rel asc, docno desc}]{\textit{Realistic} \text{ reordering}} [N_x, R_x, R_x].$$

2. *Optimistic reordering* stipulates that tied relevant documents should come above nonrelevant documents in the ranked list because the IRS may present them together, within clusters for instance. The reordering expression meeting this requirement is "qid asc, sim desc, rel desc, docno desc."

$$\text{Example: } [R_x, N_x, R_x] \xrightarrow[\text{qid asc, sim desc, rel desc, docno desc}]{\textit{Optimistic} \text{ reordering}} [R_x, R_x, N_x].$$

Regarding the selected reordering strategy (Realistic, Conventional or Optimistic) the value of a measure can differ. Notice that Optimistic reordering is fair but game-able, which is bad: a run comprised of ties only would be evaluated just like if it ranked all relevant documents at the top. As a result, we recommend the use of Realistic strategy for conducting fair evaluations. In the remainder of the paper, 'M_S' denotes measure M with reordering strategy $S \in \{R, C, O\}$. Notice the total order $M_R \leqslant M_C \leqslant M_O$ between measure values.

qid	docno	sim	rank	rel
8	**CT5**	0.9	1	1
8	AP5	0.7	2	0
8	WSJ9	0.7	3	0
8	**AP8**	0.7	4	1
8	FT12	0.6	5	0

↙ ↓ ↘

qid	docno	sim	rel
8	**CT5**	0.9	1
8	WSJ9	0.7	0
8	AP5	0.7	0
8	**AP8**	0.7	1
8	FT12	0.6	0

(b) Realistic reordering

qid	docno	sim	rel
8	**CT5**	0.9	1
8	WSJ9	0.7	0
8	**AP8**	0.7	1
8	AP5	0.7	0
8	FT12	0.6	0

(c) Conventional reordering

qid	docno	sim	rel
8	**CT5**	0.9	1
8	**AP8**	0.7	1
8	WSJ9	0.7	0
8	AP5	0.7	0
8	FT12	0.6	0

(d) Optimistic reordering

Fig. 2. Realistic, Conventional, and Optimistic reordering strategies for a run

We demonstrated in this section how an uncontrolled parameter (i.e., document naming) affects IRSs scores. In order to foster fairer evaluation, we proposed alternative Realistic and Optimistic reordering strategies. In the next section, we conduct an analysis of past Trec datasets to measure the effect of the chosen reordering strategy on evaluation results.

5 Effect of the Tie-breaking Bias on IR Evaluation

We studied the effect of the tie-breaking bias on the results of 4 Trec tracks: *ad hoc* (1993–1999), *routing* (1993–1997), *filtering* (limited to its *routing* subtask, 1998–2002, as other subtasks feature binary sim values, making them inappropriate for our study), and *web* (2000–2004). The corresponding 22 editions comprise 1,360 *runs* altogether, whose average length is 50,196 lines. This represents 3 Gb of raw data retrieved from `trec.nist.org` and analyzed as follows. In Sect. 5.1, we evaluate to what extent runs are concerned with the uncontrolled parameter issue by assessing the proportion of document ties within runs. Then, in Sect. 5.2, we report the differences between scores obtained with the proposed fair reordering strategies *vs* the Conventional strategy promoted at Trec.

5.1 Proportion of Document Ties as Observed in 22 Trec Editions

In the remainder of the paper, we call a *result-list* the sample of a run concerning a specific topic qid submitted in a given *year*, and denote it $\texttt{runid}^{\,year}_{\mathsf{qid}}$. Since differences in scores arise when a result-list contains tied documents, this section assesses how often such a phenomenon happened in the considered TREC dataset. Table 1 shows statistics related to each track: the considered editions (Year) and number of submitted runs (detailed for each year, and overall). Two other indicators are provided, regarding ★the percentage of ties in result-lists, and ☆the average number of tied documents when grouped by equal similarity (sim). Statistics related to minimum (Min), average (Avg), maximum (Max) and standard deviation (SD) are also reported. For instance, the result-list featured in Fig. 2 contains ★$3/5 = 60\%$ of tied documents, and ☆presents an average of $(1{+}3{+}1)/3 = 1.7$ tied documents per sim.

Table 1. Proportion of document ties as observed in the runs of 4 TREC tracks

Track	Year	# of runs	★Tied docs in a result-list (%)				☆Avg # of tied docs per sim			
			Min	Avg	Max	SD	Min	Avg	Max	SD
ad hoc	1993	36	0.0	30.3	100.0	36.0	2.2	4.4	28.0	4.2
	1994	40	0.0	28.4	100.0	35.9	1.9	9.5	37.3	11.2
	1995	39	0.0	29.2	99.9	32.8	1.0	2.8	26.2	4.2
	1996	82	0.0	24.1	100.0	32.3	2.0	4.1	35.1	4.7
	1997	79	0.0	24.7	100.0	34.7	1.8	4.5	25.8	5.1
	1998	103	0.0	19.0	100.0	27.4	1.0	2.5	33.8	4.4
	1999	129	0.0	15.6	100.0	24.6	1.5	3.7	22.9	4.4
	Avg over 508 *runs* →		0.0	24.5	100.0	32.0	1.6	4.5	29.9	5.5
filtering	1998	47	0.0	26.8	100.0	40.8	41.0	42.0	51.8	2.2
	1999	55	0.0	7.5	100.0	23.8	2.1	2.1	2.7	0.1
	2000	53	0.0	21.1	100.0	38.1	15.3	22.3	37.1	10.0
	2001	18	0.0	25.6	100.0	30.3	19.8	33.3	69.6	17.0
	2002	17	0.0	34.6	100.0	37.2	2.5	23.3	97.9	33.2
	Avg over 190 *runs* →		0.0	23.1	100.0	34.0	16.1	24.6	51.8	12.5
routing	1993	32	0.0	32.9	100.0	39.9	1.1	4.1	38.2	6.0
	1994	34	0.0	31.0	100.0	37.6	2.3	5.5	30.9	5.9
	1995	27	0.0	24.9	99.2	27.4	1.0	1.5	14.7	1.4
	1996	26	0.0	21.3	100.0	24.5	1.4	7.2	40.0	10.7
	1997	34	0.0	27.4	100.0	33.7	6.7	13.0	54.3	10.9
	Avg over 153 *runs* →		0.0	27.5	99.8	32.6	2.5	6.3	35.6	7.0
web	2000	104	0.0	29.3	100.0	34.3	2.9	9.3	79.6	16.6
	2001	96	0.0	32.0	100.0	31.9	25.8	27.8	63.8	5.7
	2002	71	0.0	25.8	100.0	33.5	1.0	3.6	44.7	6.3
	2003	164	0.0	18.8	100.0	27.8	1.4	2.3	12.0	1.8
	2004	74	0.0	24.9	100.0	34.4	1.5	4.3	39.6	6.2
	Avg over 509 *runs* →		0.0	26.2	100.0	32.4	6.5	9.5	47.9	7.3
Total avg over 1,360 *runs* →			0.0	25.2	100.0	32.7	6.2	10.6	40.3	7.8

Overall, IRSs participating in early TREC *ad hoc* editions contributed more result-lists with tied documents than later on. This is in line with Buckley and Voorhees's observation [10, p. 55] quoted in Sect. 2.

Averaging over each track, 25.2% of a result-list is comprised of tied documents. This proportion is highly variable, as highlighted by an average 32.7% standard deviation (strikingly similar for each track). Moreover, each year featured result-lists with no ties at all (i.e., Min★ = 0.0). It also happened that some result-lists consisted of tied documents only (1,338 result-lists over the

4 tracks). The latter case may be illustrated at TREC *ad hoc* by `ibmge2` $^{1996}_{291}$ as an example of non-discrimination: all retrieved documents share the same $\mathsf{sim} = -126.000000$ score. Those result-lists are most likely to obtain completely different results according to the applied tie-breaking strategy.

Regarding a run, when we consider the retrieved documents grouped by sim, we notice a great variability. Some result-lists have no ties ($\text{Min}^{\star} = 1.0$, which corresponds to $\text{Min}^{\bigstar} = 0.0$) while others have on average up to 97.9 documents with the same sim value. The average group size of 10.6 documents implies that a document ranked at position $r+10$ with Realistic strategy can be re-ranked rth with another strategy if lucky enough. Generalizing this observation, the larger the tied document group is, the larger the unfair position gain or loss will be.

This section showed that result-lists are likely to contain several tied documents. Thus, the uncontrolled parameter that we identified may affect IR evaluation. This hypothesis is tested in the next section.

5.2 Result Differences Regarding the Adopted Reordering Strategy

We emphasize comparisons between M_R and M_C (i.e., Realistic *vs* Conventional) to show how luck increased the M_R score deserved in a fair and unbiased setting. Notice however that larger differences will be observed when comparing M_R and M_O as the tested measures are totally ordered: $M_R \leqslant M_C \leqslant M_O$, cf. Sect. 4. For each measure, we first present systems that most benefited from the uncontrolled parameter by showing the top 3 differences between unfair M_C and fair M_R. Then, we generalize these findings by reporting statistical significance of $M_C - M_R$ for each track as a whole (i.e., considering every contributed runs). Significance p-values result from Student's paired (difference is observed between paired M_C and M_R values) one-tailed (because $M_C \geqslant M_R$) t-test. Sanderson and Zobel [13] showed that it is more reliable than other tests, such as Wilcoxon's signed rank test. The difference between tested samples is statistically significant when $p < \alpha$, with $\alpha = 0.05$. The smaller p-value, the more significant the difference is [14]. Finally, correlation between samples is reported according to Pearson's r product-moment correlation coefficient for interval scales, and Kendall's τ rank correlation coefficient for ordinal scales.

Effect on Reciprocal Rank. The effect of the chosen reordering strategy on the rank of the first relevant document is shown in Tab. 2. We report reciprocal ranks RR_x truncated to four digits but computations were done using exact values. Rank positions $1/RR_x$ are also presented because they seem helpful for the reader. Table 2 is ordered by descending $\delta_{RC} = 1/RR_R - 1/RR_C$ to focus on most 'lucky' systems. Statistical tests reported in Tab. 5 show a significant difference between RR_C and RR_R. With a Conventional strategy, the first relevant document is significantly ranked higher in the result-list than with a Realistic Strategy although the IRS remains the same. Despite this difference, a strong correlation ($\geqslant 99\%$) exists between the measure values resulting from both strategies except for the *filtering* track, as characterized by a weaker correlation (89%). Overall, RR_C and RR_R values are correlated, showing a slight but significant difference.

Table 2. Top 3 differences between Conventional $1/RR_C$ and Realistic $1/RR_R$ ranks

Track	Result-list	RR_R	RR_C	RR_O	$1/RR_R$	$1/RR_C$	$1/RR_O$	δ_{RC}
ad hoc	padre2 $^{1994}_{195}$	0.0011	0.0667	0.0769	946	15	13	931
	anu5aut1 $^{1996}_{297}$	0.0010	0.0149	1.0000	992	67	1	925
	anu5aut2 $^{1996}_{297}$	0.0010	0.0149	1.0000	992	67	1	925
filtering	antrpohsu00 $^{2000}_{32}$	0.0000	0.5000	1.0000	988	2	1	986
	antrpnohsu00 $^{2000}_{62}$	0.0000	0.0909	1.0000	988	11	1	977
	antrpohsu00 $^{2000}_{62}$	0.0000	0.0909	1.0000	988	11	1	977
routing	cir6rou1 $^{1997}_{118}$	0.0010	0.1429	1.0000	970	7	1	963
	cir6rou1 $^{1997}_{161}$	0.0010	0.0250	0.1429	998	40	7	958
	virtue3 $^{1997}_{228}$	0.0011	0.2000	0.5000	949	5	2	944
web	irtLnut $^{2001}_{516}$	0.0010	1.0000	1.0000	993	1	1	992
	ictweb10nfl $^{2001}_{525}$	0.0010	0.1667	1.0000	992	6	1	986
	ictweb10nf $^{2001}_{525}$	0.0010	0.1667	1.0000	992	6	1	986

Effect on Average Precision. The three most affected systems regarding AP are shown in Tab. 3, for each track and reordering strategy. Gain between paired strategies is also presented. We focus on gain_{CR}, between AP_C and AP_R, which represents the unfair gain obtained by IRSs which benefited from the uncontrolled parameter influencing TREC Conventional reordering. For instance, gain_{CR} reaches 406% for cir6rou1 $^{1997}_{194}$, which deserves $AP_R = 0.0262$ with a fair strategy. It obtained, however, $AP_C = 0.1325$ with the Conventional strategy.

Statistical tests reported in Tab. 5 show a significant difference between AP_C and AP_R for whatever the track. Nevertheless, this difference is small in percentage, which is in line with the observed strong correlation.

Table 3. Top 3 gains between AP_C and AP_R for each 4 tracks

Track	Result-list	AP_R	AP_C	AP_O	gain_{OR} (%)	gain_{CR} (%)	gain_{CO} (%)
ad hoc	ibmgd2 $^{1996}_{291}$	0.0000	0.0001	0.0074	49,867	318	11,858
	issah1 $^{1995}_{246}$	0.0001	0.0003	0.0018	2,718	311	585
	harris1 $^{1997}_{327}$	0.0139	0.0556	0.0556	300	300	0
filtering	IAHKaf12 $^{1998}_{13}$	0.0005	0.0116	0.0116	2,200	2,200	0
	IAHKaf32 $^{1998}_{13}$	0.0005	0.0116	0.0116	2,200	2,200	0
	IAHKaf12 $^{1998}_{39}$	0.0029	0.0625	0.2500	8,400	2,025	300
routing	cir6rou1 $^{1997}_{161}$	0.0000	0.0008	0.0060	11,995	1,435	688
	cir6rou1 $^{1997}_{194}$	0.0262	0.1325	0.2626	902	406	98
	erliR1 $^{1996}_{77}$	0.0311	0.1358	0.5714	1,736	336	321
web	ICTWebTD12A $^{2003}_{15}$	0.0064	0.2541	0.2544	3,861	3,856	0
	irtLnut $^{2001}_{516}$	0.0012	0.0355	0.2667	22,070	2,853	651
	iswt $^{2000}_{490}$	0.0000	0.0004	0.0007	4,248	2,173	91

Table 4. Top 3 gains between MAP_C and MAP_R for each 4 tracks

Track	Result-list	MAP_R	MAP_C	MAP_O	gain$_{OR}$ (%)	gain$_{CR}$ (%)	gain$_{CO}$ (%)
ad hoc	padre1 [1994]	0.1060	0.1448	0.2967	180	37	105
	UB99SW [1999]	0.0454	0.0550	0.0650	43	21	18
	harris1 [1997]	0.0680	0.0821	0.0895	32	21	9
filtering	IAHKaf12 [1998]	0.0045	0.0396	0.0558	1,140	779	41
	IAHKaf32 [1998]	0.0045	0.0396	0.0558	1,140	779	41
	TNOAF103 [1998]	0.0144	0.0371	0.0899	524	158	142
routing	cir6rou1 [1997]	0.0545	0.0792	0.2306	323	45	191
	erliR1 [1996]	0.1060	0.1412	0.2507	137	33	78
	topic1 [1994]	0.2062	0.2243	0.2543	23	9	13
web	ictweb10nf [2001]	0.0210	0.0464	0.4726	2,150	121	919
	ictweb10nfl [2001]	0.0210	0.0463	0.4660	2,119	120	907
	irtLnut [2001]	0.0102	0.0221	0.2343	2,202	117	960

Effect on Mean Average Precision. The three most affected systems regarding *MAP* are shown in Tab. 4, for each track and reordering strategy. Values for gain$_{CR}$ are smaller than for *AP* because several *AP* values are considered for computing their average (i.e., *MAP*). Since some result-lists are not skewed by the uncontrolled parameter, the influence of skewed *AP* values on *MAP* is limited, as counterbalanced by these non-skewed *AP*. Despite this smoothing effect due to using the arithmetic mean, we observed unjustified gains yet. For instance, `padre1` [1994] earned an extra 37% *MAP* by only benefiting from the uncontrolled parameter. Thus, without any tangible contribution it was granted $MAP_C = 0.1448$, although it only deserves $MAP_R = 0.1060$ in a unbiased setting. Provided that it had been even luckier, it could have unduly obtained up to $MAP_O = 0.2967$.

Although correlated (Tab. 5), MAP_C and MAP_R values are significantly different. Regarding ranks, however, Kendall's τ shows that IRS ranks computed from *MAP* do not differ significantly for whatever the track or the reordering strategy. This is due to the fact that difference in *MAP* is not large enough to change IRS ranks. Moreover, we studied the effect of the tie-breaking strategy (MAP_R *vs* MAP_C) on the statistical significance of differences between paired systems, for each edition. There are up 9% wrong conclusions: *t*-test would have concluded to significant differences ($p < 0.05$) with Conventional strategy, but to the contrary with Realistic strategy, and vice versa. As another observation, we found that the rank of 23% of the systems is different when computed on MAP_R or MAP_C. When removing the 25% worst systems, there is still 17% of affected systems. This contradicts the assumption that most ties would have been provided by bad systems. Moreover, we noticed that, for instance, *ad hoc* `uwmt6a0` [1997] was ranked 1st, although containing 57% of ties.

We showed in this section that RR_R, AP_R and MAP_R are statistically different from Conventional counterparts, meaning that there is a noticeable difference between the proposed Realistic fair reordering strategy and TREC's strategy. We discuss the implications of these findings in the next section.

Table 5. Correlation and significance of $M_C - M_R$ ($p < 0.001$ are marked with '*')

Track	RR_C vs RR_R		AP_C vs AP_R		MAP_C vs MAP_R	
	δ_{RC} (%)	corr. r	δ_{RC} (%)	corr. r	δ_{RC} (%)	corr. r
ad hoc	0.60*	0.99	0.37*	1.00	0.37*	1.00
filtering	9.39*	0.89	3.14*	0.99	3.12*	0.99
routing	1.14*	0.99	0.57*	1.00	0.58*	1.00
web	0.55*	1.00	0.40*	1.00	0.45*	1.00

6 Discussion: Tie-breaking and 'Stuffing' Phenomenon

In Sect. 5.2 we showed that IRS scores are influenced by luck. This is an issue when evaluating several IRSs. Comparing them according to evaluation measures may be unfair, as some may just have been luckier than others. In order to foster fairer evaluations, it may be worth supplying trec_eval with an additional parameter allowing reordering strategy selection: Realistic, Conventional and Optimistic. In the end, other evaluation initiatives based on trec_eval (e.g., NTCIR and CLEF) would also benefit from this contribution.

In addition to the tie-breaking bias, we identified a 'stuffing' phenomenon practiced by several IRSs. At TREC, a result-list is at most comprised of 1,000 documents. We noticed that 10.5% of the studied IRSs retrieve less than 1,000 documents for a topic and 'stuff' their result-lists with documents associated with sim = 0. This is conceptually intriguing: why would a system return an irrelevant document? One rational answer may be: among these stuffed documents some may be relevant and thus contribute to the score, even slightly. And yet, with TREC's current reordering strategy, relevant sim = 0 documents may be top ranked in the 'stuffed' part of the result-list. As a result, they unduly contribute more than if they had been ranked further down the list by the Realistic strategy that we propose. Consequently, it seems mandatory to discourage this 'stuffing trick' aiming to artificially increase measure values. This represents another case for the Realistic reordering strategy that we propose.

7 Related Works

The issue of evaluating runs comprising ties with the common, tie-oblivious, measures (e.g., precision, recall, *F1*, *AP*, *RR*, *NDCG*) was reported in [15, 16]. A way to address this issue is the design of tie-aware measures. Raghavan et al. [15] proposed Precall as an extension of precision at varying levels of recall, taking into account groups of tied documents. McSherry and Najork [16] extended the six aforementioned popular measures by averaging over all permutations of tied documents in the result-list. Both of these approaches allow the deterministic comparison of IRS results.

As an alternative solution, we did not design new measures, but tackled the tie-breaking problem by means of reordering strategies applied to the runs instead. The current reordering strategy, that we called Conventional, has been implemented in TREC since its inception. Besides being deterministic, the Realistic and Optimistic strategies that we propose allow the measurement of how much improvement (loss) in effectiveness can be reached when correctly (wrongly) ordering tied documents. A difference between these two bounds can be interpreted as a lack of the system in handling ties properly.

8 Conclusion and Future Work

This paper considered IR evaluation using the trec_eval program, which is used in major evaluation campaigns (e.g., TREC, NTCIR, CLEF) for computing IRS scores (i.e., measure values such as MAP). We underlined that scores depend on two parameters: i) the relevance of retrieved documents, and ii) document names when documents are tied (i.e., retrieved with a same sim value). We argue that the latter represents an uncontrolled parameter influencing computed scores. Indeed, luck may benefit a system when relevant documents are re-ranked higher than non relevant ones, only because of their names.

Counteracting this unfair tie-breaking strategy, we proposed two alternative strategies, namely Realistic and Optimistic reordering. A thorough study of 22 editions of TREC *ad hoc*, *routing*, *filtering*, and *web* tracks showed a statistically significant difference between the Realistic strategy that we propose *vs* TREC's current Conventional strategy for RR, AP, and MAP. However, measure values are not skewed enough to significantly change IRS ranks computed over MAP. This means that the ranking of systems is not affected. We suggest the integration of the two proposed strategies into trec_eval, allowing the experimenter to choose the proper behavior, enabling and fostering fairer evaluations. In addition, this would enable the identification of claimed 'improvements' that only result from chance.

Future work concern three main aspects. First, we plan to test whether the tie-breaking bias affected CLEF and NTCIR, just as it does for TREC. The DIRECT [17] service will be of great help in this respect. Second, as biased evaluation results may have skewed 'learning to rank' approaches [18], it would be worth checking them against fairer evaluation conducted with the proposed Realistic strategy. Third, the study of the 'stuffing' phenomenon discussed in Sect. 6 will quantify the proportion of scores obtained by exploiting side effects related to good knowledge of the evaluation protocol—instead of by improving IRS effectiveness.

Acknowledgments

We thank the reviewers for their thorough evaluations and insightful remarks.

References

1. Robertson, S.: On the history of evaluation in IR. J. Inf. Sci. **34**(4) (2008) 439–456

2. Harman, D.K., ed.: TREC-1: Proceedings of the First Text REtrieval Conference, Gaithersburg, MD, USA, NIST (February 1993)
3. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, MA, USA (2005)
4. NIST: README file for trec_eval 8.1. `http://trec.nist.gov/trec_eval`
5. Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., Hidaka, S.: Overview of IR Tasks at the First NTCIR Workshop. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, NACSIS (1999) 11–44
6. Peters, C., Braschler, M.: European Research Letter – Cross-Language System Evaluation: the CLEF Campaigns. J. Am. Soc. Inf. Sci. Technol. **52**(12) (2001) 1067–1072
7. Clough, P., Sanderson, M.: The CLEF 2003 Cross Language Image Retrieval Track. In: CLEF'03: Proceedings of the 4th Workshop of the Cross-Language Evaluation Forum. Volume 3237 of LNCS., Springer (2004) 581–593
8. Voorhees, E.M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In: SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference, New York, NY, USA, ACM (1998) 315–323
9. Zobel, J.: How Reliable are the Results of large-scale Information Retrieval Experiments? In: SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference, New York, NY, USA, ACM (1998) 307–314
10. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. [3] chapter 3 53–75
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (July 2008)
12. Voorhees, E.M.: Overview of the TREC 2004 Robust Track. In Voorhees, E.M., Buckland, L.P., eds.: TREC'04: Proceedings of the 13th Text REtrieval Conference, Gaithersburg, MD, USA, NIST (2004)
13. Sanderson, M., Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In: SIGIR'05: Proceedings of the 28th annual international ACM SIGIR conference, New York, NY, USA, ACM (2005) 162–169
14. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference, New York, NY, USA, ACM Press (1993) 329–338
15. Raghavan, V., Bollmann, P., Jung, G.S.: A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst. **7**(3) (1989) 205–229
16. McSherry, F., Najork, M.: Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In: ECIR'08: Proceedings of the 30th European Conference on IR Research. Volume 4956 of LNCS., Springer (2008) 414–421
17. Di Nunzio, G.M., Ferro, N.: DIRECT: A System for Evaluating Information Access Components of Digital Libraries. In: ECDL'05: Proceedings of the 9th European Conference on Digital Libraries. Volume 3652 of LNCS., Springer (2005) 483–484
18. Joachims, T., Li, H., Liu, T.Y., Zhai, C.: Learning to Rank for Information Retrieval (LR4IR 2007). SIGIR Forum **41**(2) (2007) 58–62