Tackling Biased Baselines in the Risk-Sensitive Evaluation of Retrieval Systems

B. Taner Dinçer[1], Iadh Ounis[2], and Craig Macdonald[2]

[1] Department of Statistics & Computer Engineering, Muğla University, 48000, Muğla, Turkey
dtaner@mu.edu.tr
[2] School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK
{iadh.ounis,craig.macdonald}@glasgow.ac.uk

Abstract. The aim of optimising information retrieval (IR) systems using a risk-sensitive evaluation methodology is to minimise the *risk* of performing any particular topic less effectively than a given baseline system. Baseline systems in this context determine the reference effectiveness for topics, relative to which the effectiveness of a given IR system in minimising the risk will be measured. However, the comparative risk-sensitive evaluation of a set of diverse IR systems – as attempted by the TREC 2013 Web track – is challenging, as the different systems under evaluation may be based upon a variety of different (base) retrieval models, such as learning to rank or language models. Hence, a question arises about how to properly measure the risk exhibited by each system. In this paper, we argue that no model of information retrieval alone is representative enough in this respect to be a true reference for the models available in the current state-of-the-art, and demonstrate, using the TREC 2012 Web track data, that as the baseline system changes, the resulting risk-based ranking of the systems changes significantly. Instead of using a particular system's effectiveness as the reference effectiveness for topics, we propose several remedies including the use of mean within-topic system effectiveness as a baseline, which is shown to enable unbiased measurements of the risk-sensitive effectiveness of IR systems.

1 Introduction

Different approaches in information retrieval (IR) such as query expansion [1, 2] and learning to rank [3] behave differently across topics, often improving the effectiveness for some of the topics while degrading performance for others. This results in a high variation in effectiveness across the topics. To address such variation, there has been an increasing focus on the effective tackling of difficult topics in particular (e.g. through the TREC Robust track [4]), or more recently, on the risk-sensitive evaluation of systems across many topics [5].

In general, the evaluation of risk is performed on the variation of a particular system against a baseline. In all the previous works, the baseline is taken as a predefined configuration of the system under consideration [5–7]. The TREC Web track has recently introduced the risk-sensitive task, to achieve a comparative evaluation of the risk across many systems [8]. In the proposed TREC risk-sensitive evaluation, the baseline is not necessarily a variation of the system deployed by an individual participating group. In

this paper, we argue that this makes the unbiased evaluation of risk challenging, as we show that using a baseline that is not a variation of the same system under consideration has implications on the validity of the risk-sensitive measurements.

Indeed, this paper shows that the choice of an appropriate baseline is of paramount importance in ensuring an unbiased risk-sensitive measurement of the performance of individual systems. We show that the higher the correlation between any given system and the baseline system across queries, the higher the measured risk-sensitive scores of that system on average, leading to a *de facto* bias in the estimation of the systems' risks.

More precisely, in all the previous works, experiments are performed on the variation of the same system, taking a particular configuration as the baseline. Although using a particular system configuration as the baseline is valid in the context of an experiment concerning a single IR system, we show that if this experimental setup is applied to multiple systems, it results in biased measurements for those systems that are not a variation of this baseline. Systems with performance scores that are highly correlated in a positive direction with the baseline scores will get their risk-sensitive performances overestimated. In contrast, systems with performance scores that are highly correlated in a negative direction will have their risk-sensitive performances under-estimated.

To address this bias in risk measurements, this paper argues for and contributes a number of definitions for alternative baselines, such as the per-topic "average system performance" that gives equal weight to every system under consideration in determining the baseline performance for each topic, and the per-topic "maximum system performance", which is akin to the achievable retrieval performance on every topic in the current state-of-the-art (SOTA). Using the TREC 2012 Web track participating systems, as well as a TREC-provided baseline system, we show that for a world-wide experimental evaluation effort such as TREC, our alternative baselines lead to unbiased evaluation of the risk-sensitive performance of the participating systems. As demonstrated in the study presented in this paper, by using the per-topic maximum baseline, the risk-sensitive evaluation of IR systems can be turned into a loss-in-SOTA evaluation where the systems are compared with each other based on measuring to what degree their observed performances diverge from the performance achievable in SOTA.

The remainder of this paper is structured as follows: In Section 2, we empirically show, through a study on the TREC 2012 Web track, the inherent bias obtained using a single baseline for multi-system risk-sensitive evaluation; Section 3 proposes several alternative baselines and shows through the same methodology their unbiasedness within a risk-sensitive evaluation. We discuss the unbiasedness property and the limitations of the proposed baselines in Section 4. We provide concluding remarks in Section 5.

2 The Bias in Risk-Sensitive Evaluation

This section first discusses the evaluation of single IR systems in a risk-sensitive fashion (Section 2.1), then their comparative evaluation within a TREC setting (Section 2.2). Later, we show that the choice of baseline for a multi-system risk-sensitive evaluation can favour some systems over others (Section 2.3), which is explained through the use of a Principal Components Analysis (Section 2.4).

2.1 Measuring Risk

The risk-sensitive performance of a retrieval system (a *run* in TREC terminology) is typically measured as the risk-reward tradeoff between the system itself and a baseline configuration of that system. In particular, given a topic set Q with c topics, the risk-reward tradeoff between a run r and a baseline run br is given by:

$$\mathrm{U}_{Risk}(r|br, Q) = \frac{1}{c}\left[\sum_{q\in Q_+}(r_q - br_q) - (\alpha+1)\sum_{q\in Q_-}(br_q - r_q)\right], \qquad (1)$$

where r_q and br_q are respectively the score of the run and the score of the baseline run on q measured by a retrieval effectiveness measure [7] (e.g. NDCG@20, ERR@20 [9]). The left summand in the square brackets gives a total win (or upside-risk) with respect to the baseline and the right summand the total loss (or downside-risk). The risk-reward tradeoff score of a run refers to the average difference of the total win from a weighted total loss with weight $(1+\alpha)$ over c topics. For higher α, the penalty for under-performing with respect to the baseline is increased: typically $\alpha = 1, 5, 10$ [8].

2.2 Comparative Risk-sensitive Evaluation of Systems

The TREC 2013 Web track aims to make a comparative evaluation of the risk of retrieval systems, in order to identify the systems that are able to consistently outperform a "provided baseline run"[1]. The provided baseline run for the TREC 2013 Web track risk-sensitive task is based on the *Indri* retrieval platform, which is developed under the Lemur project[2]. However, as the TREC 2013 campaign has yet to conclude at the time of writing, in the following we perform an empirical study based on runs submitted to the TREC 2012 Web track. Indeed, the 2013 track coordinators have made available a set of Indri runs on the TREC 2012 Web track topics[3] that correspond directly to the TREC 2013 baseline runs.

Table 1 lists the risk-sensitive scores calculated for the top 8 TREC 2012 adhoc runs and the corresponding performance ranks of the runs, for varying values of risk-sensitive parameter $\alpha = 1, 5, 10$. Here, the baseline is *IndriCASP*, an Indri run on the ClueWeb09 Category A document collection, with the Waterloo spam-page filter [10] applied - in effect, this equates to the TREC provided Indri baseline. The retrieval effectiveness measure used for the calculation of the U_{Risk} scores in Table 1 is the Expected Reciprocal Rank at 20 documents, $ERR@20$.

As can be seen from Table 1, increasing the risk-sensitive parameter α changes the risk-based performance ranking of the runs. For example, at $\alpha = 5$, run *uogTrA44xi* is demoted from rank 1 to rank 4, while *srchvrs12c09* is promoted from rank 2 to rank 1. According to the notion of risk-sensitive evaluation, as defined in [5–7], a system is promoted over another, if it minimises the risk over all topics better than the other system, i.e. a risk in the sense of showing a performance worse than that of the baseline system for any particular topic. Thus, the results given in Table 1 suggest, in theory, a conclusion that can be stated roughly as *srchvrs12c09* is the best run in minimising the

[1] http://research.microsoft.com/en-us/projects/trec-web-2013 [2] http://www.lemurproject.org

[3] https://github.com/trec-web/trec-web-2013

Table 1. U_{Risk} scores calculated for the top 8 TREC 2012 runs, based on ERR@20 measure, and the corresponding ranks of the runs (R), for varying values of risk-sensitive parameter α, where IndriCASP is the baseline.

			U_{Risk}					
Run	ERR@20	R	$\alpha = 1$	R	$\alpha = 5$	R	$\alpha = 10$	R
uogTrA44xi	0.313	1	0.0556	2	-0.1959	4	-0.5104	4
srchvrs12c09	0.305	2	0.0679	1	-0.1015	1	-0.3133	1
DFalah121A	0.292	3	0.0467	3	-0.1558	2	-0.4089	2
QUTparaBline	0.290	4	0.0385	4	-0.1893	3	-0.4740	3
utw2012fc1	0.219	5	-0.0558	6	-0.3782	6	-0.7813	6
ICTNET12ADR2	0.215	6	-0.0495	5	-0.3286	5	-0.6774	5
indriCASP	0.195	*	0	*	0	*	0	*
irra12c	0.172	7	-0.1182	7	-0.5014	7	-0.9805	7
qutwb	0.166	8	-0.1342	8	-0.5560	8	-1.0832	8

risk amongst the top 8 TREC 2012 adhoc runs, followed by *DFalah121A*, which is better than *QUTparaBline*, which in turn is better than *uogTrA44xi*, and so on. In addition, a risk-sensitive evaluation is meant to provide information on the robustness/stability of IR systems in terms of retrieval effectiveness across topics, such that *srchvrs12c09* is more robust than *DFalah121A*, which is in turn more robust than *QUTparaBline*, and so on.

In measuring the risk-sensitive performance of IR systems, we will show in the next section that the baseline system is a major factor in depicting the final risk-sensitive performance of a given system. Indeed, one could argue that, except for spam-page filtering, IndriCASP is a plain, out-of-the-box run of the Indri system that employs no advanced retrieval technology other than its core term weighting and ranking methods. Moreover, it has a moderate ERR@20 score, 0.195, compared to the top TREC 2012 runs. Thus, one can claim that it is a *fair* baseline, fair in the sense of being a true reference IR system relative to which the risk-sensitive performance of other IR systems can objectively be measured and compared with each other. There is no doubt that, to be a fair baseline, an IR system must satisfy certain conditions including but not limited to the ones considered thus far. However, the question that arises is whether this choice of a particular baseline system is unbiased with respect to the other systems under evaluation. We study this issue in details in Section 2.3.

2.3 Bias in the TREC baseline

We argue that a baseline system for risk-sensitive evaluation must not only be fair, but must also be *unbiased*, in the sense of favouring no system over another in a systematic fashion. Risk-sensitive evaluation is originally proposed for those IR experiments that involve only the variants of a single retrieval strategy, where a supplementary method, such as query expansion, is applied to the system of base retrieval methods particular to that strategy, e.g., term weighting model and ranking function. This kind of experiment follows an experiment design called the *before and after* design in statistics, where measurements are made on the response variable before and after the treatment's exposure in order to decide whether the treatment has an effect on the response of the subject of interest. In relation a to risk-sensitive evaluation, treatments correspond to the supplementary methods to be applied to the system of base retrieval methods. For this kind of risk-sensitive evaluation, the baseline system is, by default, composed of the base

Table 2. U_{Risk} scores of top 8 TREC 2012 runs at $\alpha = 10$, when each run in turn is chosen as the baseline.

	U_{Risk}							
Run	$\alpha = 10$	R	$\alpha = 10$	R	$\alpha = 10$	R	$\alpha = 10$	R
uogTrA44xi	0	*	-0.928	1	-0.882	2	-1.072	3
srchvrs12c09	-1.027	1	0	*	-0.944	3	-0.830	1
DFalah121A	-1.135	2	-1.098	3	0	*	-0.835	2
QUTparaBline	-1.349	3	-1.008	2	-0.858	1	0	*
utw2012fc1	-1.443	5	-1.412	4	-1.291	4	-1.387	4
ICTNET12ADR2	-1.434	4	-1.577	5	-1.542	5	-1.513	5
irra12c	-1.758	6	-1.941	7	-1.766	7	-1.706	6
qutwb	-1.814	7	-1.826	6	-1.715	6	-1.788	7

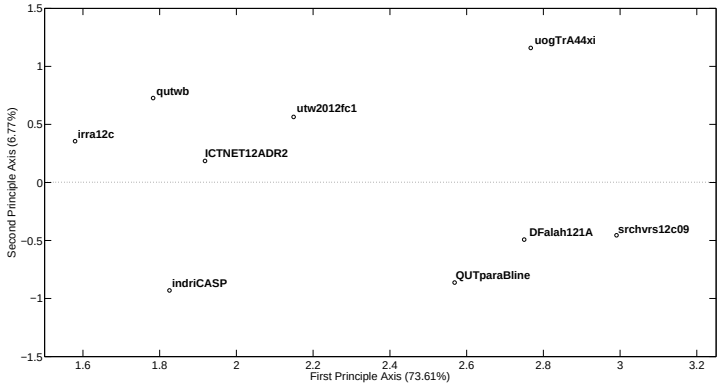
	U_{Risk}							
Run	$\alpha = 10$	R	$\alpha = 10$	R	$\alpha = 10$	R	$\alpha = 10$	R
uogTrA44xi	-0.319	1	-0.255	1	-0.068	**1**	-0.047	1
srchvrs12c09	-0.387	2	-0.497	2	-0.351	**4**	-0.159	2
DFalah121A	-0.421	3	-0.617	4	-0.329	**3**	-0.202	3
QUTparaBline	-0.540	4	-0.611	3	-0.293	**2**	-0.298	5
utw2012fc1	0	*	-0.843	5	-0.432	5	-0.224	4
ICTNET12ADR2	-0.897	6	0	*	-0.581	6	-0.613	7
irra12c	-0.997	7	-1.092	6	0	*	-0.610	6
qutwb	-0.866	5	-1.201	7	-0.687	7	0	*

methods determining the retrieval strategy under consideration, so that, after the exposure to the treatment, any measured gain or loss in the effectiveness of the system under evaluation can be attributed only to the supplementary method applied.

On the other hand, as opposed to the classical before and after experimental design setup, for a TREC-like large experiment involving multiple retrieval strategies (most of which could not be necessarily considered as variants of each others) there would arguably be no single reference system that is composed of the base retrieval methods in common to every retrieval strategy under evaluation. Table 2 shows the calculated U_{Risk} scores and the corresponding systems' rankings for each of the 8 possible selections of baseline system (i.e. each run is in turn chosen as the baseline). From the table, the observed systems' rankings markedly and significantly vary as the baseline changes ($p = 0.0003$ according to the nonparametric Friedman's test).

The variance in the effectiveness of IR systems across topics explains why different baseline systems yield different risk-based system rankings. IR systems with different retrieval strategies would in general show different performance profiles over a given set of topics. Two IR systems with similar/parallel performance profiles may be considered to be variants of each other, in analogy to before and after design experiments. However, having similar performance profiles over topics basically implies a positive correlation between the observed performance scores across the topics, while having discordant performance profiles implies a negative correlation. Using the classical risk-sensitive evaluation setup, given a set of IR systems, any particular baseline system promotes in a systematic fashion those IR systems whose performance profiles are similar to its performance profile, while demoting the systems with discordant performance profiles. In the next section, we use Principal Components Analysis (PCA) [11] to demonstrate the correlation of the effectiveness of the TREC submitted runs across topics, thereby illustrating how particular choices of baselines can be correlated with particular runs.

Fig. 1. PCA plot of the top 8 TREC 2012 Web track adhoc runs, based on the first and second principal axes.



2.4 Principal Components Analysis of Per-topic Effectiveness

PCA – a dimension reduction technique – provides an intuitive way for visually exploring the performance correlations among IR systems across topics [12]. Since higher dimensional spaces are difficult to inspect, it becomes necessary to reduce them into lower dimensions. For a test collection with many topics, groups of topics would in general be performed similarly by IR systems. In PCA, each principal component is a linear combination of the original topics. The first principal component is a single axis in space. When you project the measured system performance scores for each topic on that axis, the resulting values form, in one sense, a meta topic. The variance of scores in this meta topic is the maximum among all possible choices of the first axis. The second principal component is another axis in space, perpendicular to the first. The projection of the measured system performance scores on this axis generates another meta topic. The score variance associated with this meta topic is the maximum among all possible choices of this second axis. All the principal components are orthogonal to each other, and hence, as opposed to the original topics, there is no redundant information on the performance relationships among IR systems across topics. Hence, the first two principal components together can be used to define two orthogonal dimensions, and therefore to visualise the performance relationships among IR systems across topics by means of a scatter plot that accounts for the major part of the total performance variations observed on the original topics. The PCA plot of the top 8 TREC 2012 Web track adhoc runs using the first and the second principal axes is given in Figure 1. For this PCA plot, the first principal component accounts for 73.61% of the total sum of squared deviations observed on 50 topics, and the second principal component accounts for 6.77% of the same total i.e., together accounting for 80% of the total variation in performance among the top 8 TREC 2012 Web track adhoc runs across the topics.

The first principal component is positively related to almost all topics, and hence it acts as an index variable, which mostly agrees with the ranking of systems based on the $\mathrm{ERR@20}$ measure in Table 1 (i.e., 73.61%). The discrepancy between the systems' ranking based on the first principal axis and the systems' ranking based on the $\mathrm{ERR@20}$ measure is in essence due to the differences among the performance profiles of runs, and can be explained by the successive principal components. On this account,

the second principal component, which contrasts the observed average scores of runs, explains 6.77% of the total profile deviation. Note that, if all the runs under consideration had the same performance profile across the topics, the vector of scores for each topic will be a linear combination of the vectors of scores for the other topics, resulting in a perfect correlation among the topics and among the runs, and thus the total sum of squared deviations observed on 50 topics can be completely explained by the first principal component.

The interpretation of a PCA plot is simple. Runs that are clustered around the same location in the plot are runs that have both comparable average ERR@20 scores and similar performance profiles over all (50) topics, such as *srchvrs12c09*, *DFalah121A*, and *QUTparaBline*. For the PCA plot in Figure 1, since the first principal component is positively related to all topics, a low component score for any run implies a low average score over all topics, and a high component score implies a high average score. On the other hand, the second principal component contrasts the observed scores of runs on two subsets of topics. For any two runs that are contrasted by the second principal component, say *srchvrs12c09* and *uogTrA44xi*, a high component score in a positive direction implies a high average score for one of them (i.e., *uogTrA44xi*) and simultaneously a low average score for the other (i.e., *srchvrs12c09*) on the topic subset to which the second principal component is positively related. Conversely, on the topic subset to which the second principal component is negatively related, a high component score in a negative direction implies a high average score for the latter run (*srchvrs12c09*) and a low average score for the former (i.e., *uogTrA44xi*).

In summary, comparing the *uogTrA44xi* and *IndriCASP* runs, the PCA plot in Figure 1 reveals that *uogTrA44xi* performs most of the 50 topics better than *IndriCASP* (c.f. the position of the runs along the first principal axis) but on a particular subset of topics, to which the second principal axis is negatively related, *IndriCASP* has relatively high scores while *uogTrA44xi* has relatively low scores, compared to the within-topic average system scores. This is actually the case for all the runs having positive component scores on the second principal axis, such as *utw2012fc1* and *ICTNET12ADR2*. In contrast, the opposite case is true for the runs with negative component scores. For instance, for the runs *srchvrs12c09* and *DFalah121A*, which are comparable to *uogTrA44xi* in terms of average ERR@20 score, the PCA plot reveals that they show their low and high scores on every topic in synchronisation with *IndriCASP*. Due to the nature of a risk-sensitive evaluation, the use of *IndriCASP* as a baseline will favour, in a systematic fashion, those runs that have negative component scores on the second principal axis over the runs with positive component scores. This is the basic reason why one of the two comparable runs in terms of average ERR@20 score, *srchvrs12c09*, is promoted to rank 1 while *uogTrA44xi* is demoted to rank 4 in the risk-based systems' ranking obtained using *IndriCASP* as the baseline. Simultaneously, this is also why *uogTrA44xi* is prompted to rank 1 and *srchvrs12c09* is demoted to rank 4 of the ranking obtained when *irra12c* is used as the baseline (both the *uogTrA44xi* and *irra12c* runs can be considered as variants of each other with respect to their base retrieval strategies, *Divergence From Randomness* [13] and *Divergence From Independence* [14], as well as both being based upon the Terrier retrieval platform [15][4]). Hence, within a *comparative* risk-sensitive evaluation of different IR systems, we conclude that the choice of *IndriCASP* as the baseline run benefits systems similar to that run, and hinders the risk-sensitive performance of

[4] http://terrier.org

other runs. In the next section, we propose several alternatives to define the baseline performance for comparative risk-sensitive evaluations, that are both fair and unbiased.

3 Unbiased Comparative Risk-Sensitive Evaluation

As shown above, no IR system's retrieval strategy alone is representative enough to be used as a baseline for the risk-sensitive evaluation of different retrieval strategies. In this respect, every retrieval strategy biases towards its variants. To fairly measure and compare the risk-sensitive performance of multiple retrieval strategies that are stemmed from different base models of information retrieval (e.g., vector space vs. probabilistic) in an unbiased manner, there needs to be a baseline that is generalisable enough to be applied to each retrieval strategy under evaluation. In classical statistics, a remedy for such issues is the use of an estimate of the parameter of interest. Here, the parameter of interest is the performance of an unbiased baseline system for any given topic. Given a particular topic q and a set of r runs, the arithmetic mean of the r performance scores according to an evaluation measure observed on q is one of the possible estimates of the unbiased baseline score UBS_q:

$$\mathrm{UBS}_q = \frac{1}{r} \sum_{i=1}^{r} s_i(q), \tag{2}$$

where $s_i(q)$ is the performance score of run i on topic q for $i = 1, 2, \ldots, r$ for a given evaluation measure (e.g. $\mathrm{ERR@20}$). Since the arithmetic mean gives equal weight to every retrieval strategy in determining the UBS_q, a baseline system that is determined by the UBS_q scores, MEAN for short, will be unbiased with respect to the retrieval strategies yielding the r run scores.

We note that the notion of risk defined here by MEAN is different from the original notion of risk, in that the U_{Risk} measure now measures the risk-sensitive performance of an IR system in terms of the risk of a given topic being less effective than the mean system effectiveness *expected* on that topic, instead of a single system's effectiveness.

Next, the *median* of the within-topic scores, MEDIAN, can also be used as the unbiased baseline score for each topic. In addition, it is also possible to take the *maximum* topic scores as the baseline, MAX. Since such a baseline will represent the achievable system performance for each topic in the current state-of-the-art, it turns the risk-sensitive evaluation of IR systems into a loss-in-SOTA evaluation where the systems are compared with each others based on measuring to what degree their observed performances diverge from the performance achievable in SOTA. Table 3 shows the risk-based ranking of the top 8 TREC 2012 Web track adhoc runs, using MEAN, MEDIAN and MAX as the baseline, respectively.

In general, given a set of IR systems with different retrieval strategies for a comparative risk-sensitive evaluation, an unbiased baseline system can be thought of as the system that is jointly determined by the given systems, such that being selected as the baseline is equally likely for each system. This definition of unbiasedness applies to the three baselines that we have proposed. To demonstrate the unbiasedness of our proposed baselines, we again turn to PCA.

Figure 2 provides the PCA plot showing the MEAN, MEDIAN and MAX baselines as three virtual runs. In the PCA plot of Figure 2, a run with scores equal to the mean

Table 3. U_{Risk} scores of the top 8 TREC 2012 runs, using MEAN, MEDIAN and MAX as the baseline, and the corresponding ranks of the runs (R), for $\alpha = 5, 10$.

		U_{Risk}											
		MEAN				MEDIAN				MAX (SOTA)			
Run	ERR@20	R	$\alpha = 5$	R	$\alpha = 10$	R	$\alpha = 5$	R	$\alpha = 10$	R	$\alpha = 5$	R	$\alpha = 10$	R
uogTrA44xi	0.313	1	-0.1061	1	-0.2845	1	0.0059	1	-0.0969	1	-1.3901	1	-2.5486	1
srchvrs12c09	0.305	2	-0.1668	2	-0.3848	2	-0.0067	2	-0.1140	2	-1.4398	2	-2.6396	2
DFalah121A	0.292	3	-0.1760	3	-0.4161	3	-0.0469	3	-0.1815	3	-1.5168	3	-2.7809	3
QUTparaBline	0.290	4	-0.1976	4	-0.4444	4	-0.0559	4	-0.1976	4	-1.5286	4	-2.8024	4
utw2012fc1	0.219	5	-0.3574	5	-0.6935	5	-0.1235	5	-0.2621	5	-1.9523	5	-3.5792	5
ICTNET12ADR2	0.215	6	-0.4204	6	-0.8149	6	-0.2757	6	-0.5620	6	-1.9796	6	-3.6293	6
indriCASP	0.195	7	-0.5343	7	-1.0224	7	-0.3500	7	-0.6616	7	-2.1011	7	-3.8521	7
irra12c	0.172	8	-0.5555	8	-1.0425	8	-0.3921	9	-0.7522	8	-2.2351	8	-4.0977	8
qutwb	0.166	9	-0.5908	9	-1.1067	9	-0.3825	8	-0.7553	9	-2.2736	9	-4.1682	9

scores for each topic will be shown at the origin with respect to all contrasting principal axes. This is the reason why the MEAN baseline is close to the origin with respect to the second principal axis in Figure 2[5]. Here, based on the given definition of unbiasedness, one may argue that the closer a baseline is to the origin, the higher its degree of unbiasedness w.r.t. the systems under evaluation. However, this is not the case. As can be seen from Figure 2, the MEDIAN and MAX baselines diverge from the origin, although they are unbiased. As statistical estimates, both MEDIAN and MAX have the same property as MEAN in giving equal weight to every system in determining the baseline scores for each topic, and hence, as a baseline, they have the property of unbiasedness. However, for the MEDIAN baseline, the PCA plot shows in essence that 1) the median scores for most of the topics are slightly below the corresponding mean scores (i.e., the positions of MEAN and MEDIAN with respect to the first axis), and that 2) the median scores tend to be higher in magnitude than the corresponding mean scores for those topics to which the second principal axis is negatively related. Similarly, the MAX baseline exhibits the same issue. Overall, the PCA plot of Figure 2 shows that our three proposed baselines are unbiased, but each serves a different purpose in minimising the risk attached to the IR systems under evaluation.

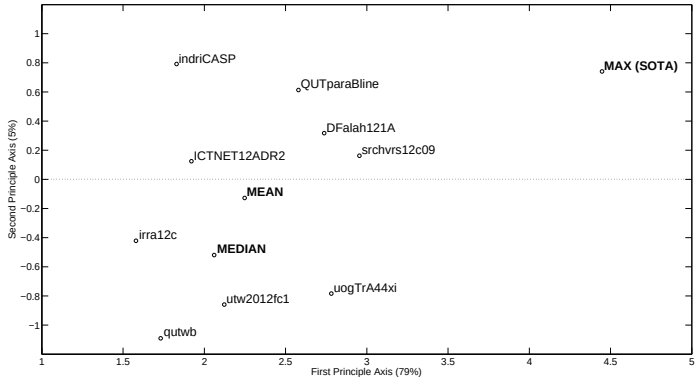
Our introduced definition of unbiasedness actually exposes a problem about the validity of the *comparative* risk-sensitive evaluation of IR systems, which we will discuss in detail in the next section. This issue is related to the rankings of risk-sensitive systems obtained using the unbiased baselines. Indeed, we will show that such a comparison of the risk-sensitive performances of different IR systems in an objective manner actually implies the comparison of the retrieval effectiveness of the individual systems based on the underlying measure, i.e. ERR@20.

4 Discussion

As can be seen in Table 3, the risk-based ranking of systems remains concordant with the original systems' ranking and also with each others for all values of the risk parameter α, as opposed to the risk-based systems' rankings given in Table 1 where *indriCASP*

[5] The MEAN baseline is not shown at the origin because the mean scores, median scores, and maximum scores for the corresponding baseline runs are calculated using the scores of 9 actual runs and then PCA is applied to a (run$\times$topic) matrix of 12 runs (9 actual + 3 virtual) and 50 topics. Removing the median and max scores from the matrix would fix this issue.

Fig. 2. PCA plot showing the performance relationships among the top TREC 2012 Web track runs, *indriCASP*, and the mean and median topic scores across 50 Web track topics.



was used as the baseline. Indeed, this simply suggests that the use of *indriCASP* as a baseline is not unbiased. Note that, for each value of α, a risk-sensitive evaluation is applied to the same set of systems, where in fact the (true) risk associated with each system that the U_{Risk} measure is intended to measure remains constant for all α values. Under this particular condition, unless the baseline favours some systems over the others in a systematic fashion, it is expected that the calculated U_{Risk} scores for each system will increase in magnitude evenly, proportional to the α values, because of the magnification of the downside-risk, while the rankings of the top 8 TREC 2012 runs remain unchanged relative to each other. This is actually the case in Table 3 but not in Table 1.

Figure 3 shows (a) the observed per-topic ERR@20 scores of *uogTrA44xi* and *srchvrs12c09*, (b) the U_{Risk} scores for *uogTrA44xi* when *srchvrs12c09* is the baseline, and (c) the U_{Risk} scores for *srchvrs12c09* when *uogTrA44xi* is the baseline. *srchvrs-12c09* has 11 topic scores greater than 0.8, while *uogTrA44xi* has 8 topic scores greater than 0.8 (a). However, the per-topic loss of *srchvrs12c09* with respect to *uogTrA44xi* (a) is greater in number than the per-topic loss of *uogTrA44xi* with respect to *srchvrs12c09* (c). Thus, it can be argued that the risk of showing a performance worse than that of a fair baseline is higher for *srchvrs12c09* than for *uogTrA44xi*.

For *uogTrA44xi*, *srchvrs12c09* and *DFalah121A*, Figure 4 shows the per-topic U_{Risk} scores when MEAN is used as the baseline. From the figure, it can be observed that the loss distribution on the topics for *uogTrA44xi* is steeper than that of *srchvrs12c09*, which in turn is steeper than *DFalah121A*, suggesting that the risk attached to *uogTrA44xi* is less than that of *srchvrs12c09*, which in turn is less than that of *DFalah121A*. As a result, we argue that the comparison of the risk-sensitive performances of different IR systems in an objective manner actually implies comparing the retrieval effectiveness of individual systems based on the underlying effectiveness measure, i.e. ERR@20.

5 Conclusions

Following the proper practice of risk-sensitive evaluation in before and after design experiments concerning single IR systems, it would appear, for the experimental evaluation of multiple IR systems, that every IR system employing a particular retrieval

Fig. 3. Comparison of *uogTrA44xi* and *srchvrs12c09*: a) based on per topic ERR@20 scores, b) per topic U_{Risk} scores of *srchvrs12c09* for $\alpha = 5$ when *uogTrA44xi* is the baseline, and c) per topic U_{Risk} scores of *uogTrA44xi* for $\alpha = 5$ when *srchvrs12c09* is the baseline.

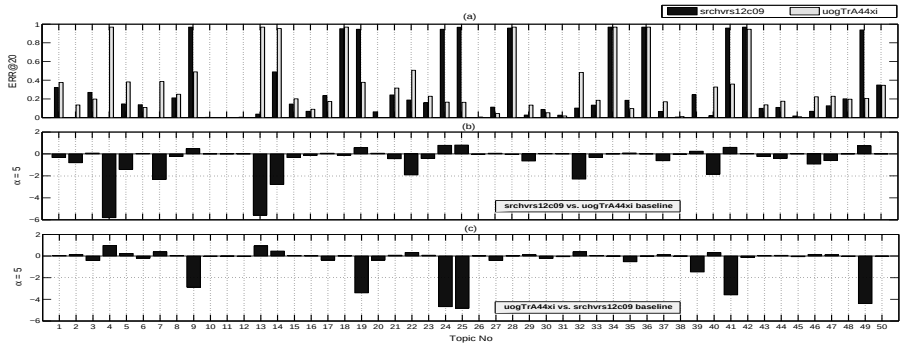
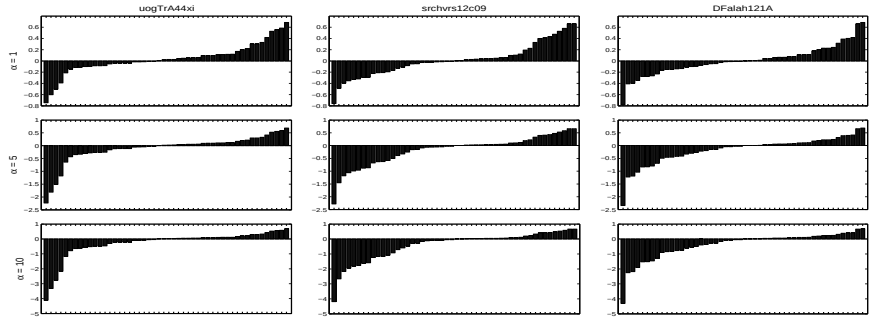


Fig. 4. Distribution of the U_{Risk} scores on 50 Web track topics for *uogTrA44xi*, *srchvrs12c09*, and *DFalah121A* runs, measured using topic mean scores as the baseline for $\alpha = 1, 5, 10$. U_{Risk} scores are shown in ascending order for each run of the ease in visual comparison.



strategy requires a baseline system that is composed of the base retrieval methods particular to that strategy, for the purpose of the proper measurement of its risk-sensitive performance. However, since this experiment design asks for a different baseline for every IR system having a particular retrieval strategy that could not be considered the variant of the retrieval strategies represented by the baseline systems at hand, it would cause existing test collections not to be reusable for new IR systems. On the other hand, selecting a particular IR system as the baseline – as attempted by the TREC 2013 Web track – is also not a viable remedy for the issue of test collection reusability. Moreover, we show in this study that selecting a particular IR system as the baseline will result in biased performance measurements for all systems that are not a variation of the provided baseline. Finally, we also demonstrate that a comparative risk-sensitive evaluation of multiple IR systems using unbiased baselines actually implies the typical adhoc type evaluation of the systems based on a retrieval effectiveness measure like ERR@20.

Nevertheless, the benefit of the proposed baselines is that individual IR systems can be optimised for risk minimisation, using MEAN or MEDIAN as the baseline, with

respect to the retrieval effectiveness expected for every topic on average by current SOTA IR technology, and, using MAX as the baseline, with respect to the retrieval effectiveness achievable in SOTA.

In summary, we question whether it is possible to conduct a comparative evaluation campaign for risk-sensitive approaches within a TREC-like setting, as it is impossible to derive a common, unbiased baseline that can measure risk-sensitivity separately from average effectiveness. Alternatively, a comparative evaluation for risk-sensitivity could be formulated with each participating risk-sensitive run being measured with respect to the effectiveness of its own declared baseline. However, such an operationalisation would make it difficult to combine measures of baseline effectiveness and risk-sensitivity into a theoretically defined final measure that is comparable across baselines/participating groups. A similar problem has been faced by some previous tasks, e.g., the TREC 2004 Terabyte track efficiency task used tradeoff graphs for comparing the efficiency and effectiveness of participating runs on different axes [16]. We leave the theoretical combination of baseline effectiveness and risk-sensitivity to future work.

References

1. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: Proc. ECIR. pp. 127–137 (2004)
2. Carmel, D., Farchi, E., Petruschka, Y., Soffer, A.: Automatic query refinement using lexical affinities with maximal information gain. In: Proc. SIGIR. pp. 283–290. (2002)
3. Macdonald, C., Santos, R., Ounis, I.: The whens and hows of learning to rank for web search. Information Retrieval 16(5), 584–628 (2013)
4. Voorhees, E.M.: Overview of the TREC 2003 robust retrieval track. In: Proc.TREC(2003).
5. Collins-Thompson, K.: Accounting for stability of retrieval algorithms using risk-reward curves. In: Proceedings of SIGIR Workshop on the Future of Evaluation in Information Retrieval. (2009)
6. Collins-Thompson, K.: Reducing the risk of query expansion via robust constrained optimization. In: Proc. CIKM. pp. 837–846. (2009)
7. Wang, L., Bennett, P.N., Collins-Thompson, K.: Robust ranking models via risk-sensitive optimization. In: Proc. SIGIR. pp 761–770 (2012)
8. Collins-Thompson, K., Bennett, P.N., Diaz, F., Clarke, C., Voorhees, E.: TREC 2013 Web Track Guidelines, http://research.microsoft.com/en-us/projects/trec-web-2013/
9. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proc. CIKM. pp. 621–630. (2009)
10. Cormack, G., Smucker, M., Clarke, C.: Efficient and effective spam filtering and re-ranking for large web datasets. Information Retrieval 14(5), 441–465 (2011)
11. Jackson, J.E.: A users guide to principal components. John Wiley & Sons (1990)
12. Dinçer, B.T.: Statistical principal components analysis for retrieval experiments. Journal of the American Society for Information Science and Technology 58(4), 560–574 (2007)
13. Amati, G., van Rijsbergen, C.: Probabilistic models of information retrieval based on measuring the divergence from randomness. Transactions on Information Systems 20(4), 357–389 (2002)
14. Dinçer, B.T.: IRRA at TREC 2012: Index term weighting based on divergence from independence model. In: Proc. TREC. (2012)
15. Macdonald, C., McCreadie, R., Santos, R., Ounis, I.: From puppy to maturity: experiences in developing Terrier. In: Proc. OSIR at SIGIR (2012)
16. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2004 Terabyte track. In: Proc. TREC (2004)