Risk-Sensitive Evaluation and Learning to Rank using Multiple Baselines

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

[1] Sitki Kocman University of Mugla, Mugla, Turkey
[2] University of Glasgow, Glasgow, UK

dtaner@mu.edu.tr[1],{craig.macdonald, iadh.ounis}@glasgow.ac.uk[2]

ABSTRACT

A robust retrieval system ensures that user experience is not damaged by the presence of poorly-performing queries. Such robustness can be measured by risk-sensitive evaluation measures, which assess the extent to which a system performs worse than a given baseline system. However, using a particular, single system as the baseline suffers from the fact that retrieval performance highly varies among IR systems across topics. Thus, a single system would in general fail in providing enough information about the real baseline performance for every topic under consideration, and hence it would in general fail in measuring the real risk associated with any given system. Based upon the Chi-squared statistic, we propose a new measure Z_{Risk} that exhibits more promise since it takes into account multiple baselines when measuring risk, and a derivative measure called GeoRisk, which enhances Z_{Risk} by also taking into account the overall magnitude of effectiveness. This paper demonstrates the benefits of Z_{Risk} and GeoRisk upon TREC data, and how to exploit GeoRisk for risk-sensitive learning to rank, thereby making use of multiple baselines within the learning objective function to obtain effective yet risk-averse/robust ranking systems. Experiments using 10,000 topics from the MSLR learning to rank dataset demonstrate the efficacy of the proposed Chi-square statistic-based objective function.

1. INTRODUCTION

The classical evaluation of information retrieval (IR) systems has focused upon the arithmetic mean of their effectiveness upon a sample of queries. However, this does not address the *robustness* of the system, i.e. its effectiveness upon the worst performing queries. For example, while some retrieval techniques (e.g. query expansion [2, 8]) perform effectively for some queries, they can orthogonally cause a decrease in effectiveness for other queries. To address this, various research into robust and *risk-sensitive* measures has taken place. For instance, in the TREC Robust track, systems were measured by geometric mean average precision [23, 25] to determine the extent to which they perform well on all queries. More recently, the notion of risk-sensitivity has been introduced, in that an evaluation measure should consider per-query losses and gains compared to a particular baseline technique [11]. Within this framework, measures such as U_{Risk} [27] and T_{Risk} [13] have been proposed. Both measures can be adapted to integrate with the state-of-the-art LambdaMART learning to rank technique.

Since risk-sensitive measures compare to a specific baseline, such measures are most naturally applied in experiments using a *before-and-after* design, where different treatments are applied to a particular baseline system, e.g. query expansion. However, when simply considering a single baseline, a full knowledge of the difficulty of a particular query cannot be obtained. For instance, a single baseline system may perform lowly for a query that other systems typically perform well. For this reason, the inference of risk based upon a population of baseline systems is attractive. One can easily draw an analogy with the building of ranking methods that combine multiple weighting models, such as data fusion or learning to rank, to obtain a more effective final ranking. Moreover, the use of multiple baselines permits a deployed search engine to evaluate the risk of an alternative retrieval approach not only with respect to its own baseline, but also to other competitor systems.

In this paper, we show how a risk-sensitive evaluation based on the Chi-square test statistic permits the consideration of multiple baselines, unlike the existing measures U_{Risk} & T_{Risk} which can only consider a single baseline. In doing so, we argue that a robust system should not be less effective for a given topic than an *expectation* of performance given a population of other (baseline) systems upon that topic. In particular, this paper contributes: a new risk-sensitive evaluation measure, namely Z_{Risk}, based on Chi-square test statistic, and a derivative called GeoRisk that enhances Z_{Risk} by also taking into account the overall magnitude of effectiveness; Moreover, we demonstrate the use of Z_{Risk} and GeoRisk upon a TREC comparative evaluation of Web retrieval systems; Finally, we show how to directly and effectively integrate GeoRisk within the state-of-the-art LambdaMART learning to rank technique.

This paper is organised as follows: Section 2 provides a background on robust and risk-sensitive evaluation; Section 3 defines Z_{Risk} based upon Chi-squared statistic, as well as the GeoRisk derivative; Section 4 & Section 5 demonstrate the proposed measures upon synthetic & real TREC data, while Section 6 shows the integration of GeoRisk within the LambdaMART learning to rank technique; Related work and concluding remarks follow in Sections 7 & 8.

2. RISK-SENSITIVE EVALUATION

Risk-sensitive evaluation [11] aims at quantifying the trade-off between risk and reward for any given retrieval strat-

SIGIR '16, July 17–21, 2016, Pisa, Italy.

DOI: http://dx.doi.org/10.1145/2911451.2911511

egy. Information retrieval performance, which is usually measured by a given retrieval effectiveness measure (e.g. NDCG@20, ERR@20 [9]) over a set of topics Q, can be expressed in terms of risk and reward as a risk function. Such a risk function takes into account the downside-risk of a new system s with respect to a given baseline system b (i.e. a loss: performing a topic q worse than the baseline according to the effectiveness measure, $s_q < b_q$) and an orthogonal reward function that takes into account the upside-risk (i.e. a win: performing a topic better than the baseline, $s_q > b_q$).

A single measure, U_{Risk} [27], which allows the tradeoff between risk and reward to be adjusted, is defined as:

$$\mathrm{U}_{Risk} \quad = \quad \frac{1}{c}\left[\sum_{q\in Q_+}\delta_q + (1+\alpha)\sum_{q\in Q_-}\delta_q\right], \qquad (1)$$

where $c = |Q|$ and $\delta_q = s_q - b_q$. The left summand in the square brackets, which is the sum of the score differences δ_q for all q where $s_q > b_q$ (i.e. $q \in Q_+$), gives the total win (or upside-risk) with respect to the baseline. On the other hand, the right summand, which is the sum of the score differences δ_q for all q where $s_q < b_q$, gives the total loss (or downside-risk). The risk sensitivity parameter $\alpha \geq 0$ controls the tradeoff between reward and risk (or win and loss): $\alpha = 0$ calculates the average change in effectiveness between s and b, while for higher α, the penalty for under-performing with respect to the baseline is increased: typically $\alpha = 1, 5, 10$ [12] to penalise risky systems, where $\alpha = 1$ doubles the emphasis of down-side risk compared to $\alpha = 0$.

Recently, Dinçer et al. [13] introduced a statistically-grounded risk-reward tradeoff measure, T_{Risk}, as a generalisation of U_{Risk}, for the purposes of hypothesis testing:

$$\mathrm{T}_{Risk} = \frac{\mathrm{U}_{Risk}}{SE(\mathrm{U}_{Risk})}, \qquad (2)$$

where $SE(\mathrm{U}_{Risk})$ is the *standard error* in the risk-reward tradeoff score U_{Risk}. Here, T_{Risk} is a linear monotonic transformation of U_{Risk}. This transformation is called *studentisation* in statistics (c.f., t-scores) [16], and T_{Risk} can be used as the test statistic of the Student's t-test. Moreover, the aforementioned work shows that T_{Risk} permits a state-of-the-art learning to rank algorithm (LambdaMART) to focus on those topics that lead to a significant level of risk in order to learn effective yet risk-averse ranking systems.

On the other hand, the comparative risk-sensitive evaluation of different IR systems is challenging, as the systems may be based upon a variety of different (base) retrieval models – such as learning to rank or language models – or upon different IR platforms (Indri, Terrier etc.). It has been shown that using a particular system as the baseline in a comparative risk-sensitive evaluation of a set of diverse IR systems – as attempted by the TREC 2013 and 2014 Web track – yields biased risk-reward tradeoff measurements [14], especially when the systems under evaluation are not variations of the provided baseline system. To address this, the use of the within-topic mean system performance was proposed as an unbiased baseline (as well as the within-topic median system performance and the within-topic maximum system performance). Given a particular topic q and a set of r systems, the arithmetic mean of the r performance scores according to an evaluation measure observed on q is the unbiased baseline score:

$$\mathrm{Mean}_q = \frac{1}{r}\sum_{i=1}^{r} s_i(q), \qquad (3)$$

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

However, as shown in [14], the use of Mean_q exposes a problem about the validity of the comparative risk-sensitive evaluation of different IR systems. This issue is related to the risk-based rankings of the systems obtained using Mean_q. Indeed, such a comparison of the risk-sensitive performances of different IR systems actually implies the comparison of the retrieval effectiveness of the individual systems based on the underlying effectiveness measure, i.e. ERR-@20 [14]. That is, the ranking of the systems obtained by using the underlying effectiveness measure will be the same as the risk-based ranking of the systems obtained using the unbiased baseline Mean_q, irrespective of the value of the risk sensitivity parameter α.

Most importantly, the previously proposed risk measures are only sensitive to the mean and the variance of the observed losses and wins, i.e. U_{Risk} is sensitive to mean and T_{Risk} is sensitive to mean and variance (c.f. $SE(\mathrm{U}_{Risk})$). However in a comparative risk-sensitive evaluation, we argue that it is necessary to be sensitive to the shape of the score distributions, as well as the mean and the variance. As such, in the next section, we propose the Z_{Risk} measure, which satisfies the aforementioned variance and shape requirements of a comparative risk-sensitive IR evaluation, while the derivative GeoRisk measure enhances Z_{Risk} by naturally incorporating the overall effectiveness of the considered system.

3. MEASURES OF RISK FROM CHI-SQUARE

Each existing robust and risk-sensitive evaluation measure each encodes properties about what a good (or bad) IR system should exhibit. Firstly, the classical mean measure (e.g. MAP or mean NDCG) stipulates that a good system should perform well on a population of topics on average; The geometric mean (e.g. as proposed in [24] for Mean Average Precision as GMAP) says that a good system should avoid performing lowly on any topics, while comparing GMAP values permits identifying improvements in low performing topics, in contrast to mean, which gives equal weight to absolute changes in per-topic scores, regardless of the relative size of the change [4]. Risk-sensitive evaluation measures such as U_{Risk} and T_{Risk} use the notion of a baseline - a good system should perform well, but preferably no worse than the given baseline. Hence U_{Risk} responds to changes in the mean effectiveness of the system, but emphasises those worse than the baseline. Building upon U_{Risk}, T_{Risk} is also sensitive to the variance exhibited by a system across the population of topics. These attributes are highlighted in Table 1.

In this section, we argue for a risk measure that considers the 'shape' of a system's performance across topics. In particular, we consider that the distribution of the effectiveness scores of a set of baseline systems across the topics, mapped to the same overall mean effectiveness as the system at hand, represents an expected performance for each topic that the system should not underperform. In other words, we calculate the *expectation* of the system's performance for each topic, by considering the overall performance of the current system and the observed performances of other baseline systems. This allows to determine topics that the system should be performing better on. It follows that our proposal encap-

Measure	Baseline	Penalty of low topics	Sensitive to: Mean	Var.	Shape
Mean AP	None	None	✔	✗	✗
Geo. MAP	None	Focus on lowest topics	✔	✗	✗
U$_{Risk}$	Single	$1 + \alpha$	✔	✗	✗
T$_{Risk}$	Single	$1 + \alpha$	✔	✔	✗
Z$_{Risk}$	Multiple	$1 + \alpha$	✗	✔	✔
GeoRisk	Multiple	$1 + \alpha$	✔	✔	✔

Table 1: Comparison of existing and proposed robustness/risk-sensitive measures.

		Topics					
	Systems	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\dots$	$\mathbf{t}_c$	Total
	$\mathbf{s}_1$	x_{11}	x_{12}	x_{13}	$\dots$	x_{1c}	S_1
	$\mathbf{s}_2$	x_{21}	x_{22}	x_{23}	$\dots$	x_{2c}	S_2
$\mathbf{X} =$	$\mathbf{s}_3$	x_{31}	x_{32}	x_{33}	$\dots$	x_{3c}	S_3
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\mathbf{s}_r$	x_{r1}	x_{r2}	x_{r3}	$\dots$	x_{rc}	S_r
	Total	T_1	T_2	T_3	$\dots$	T_c	N

Table 2: Data matrix for an IR experiment.

sulates two separate measures: Z$_{Risk}$, introduced in Section 3.1, which measures the shape of the system's performance irrespective of the overall magnitude of effectiveness; and later in Section 3.2 we show how to create a risk-measure responsive to mean effectiveness called GeoRisk. The subsequent Section 4 & Section 5 demonstrate the Z$_{Risk}$ and GeoRisk measures upon artificial and TREC data.

The first measure, Z$_{Risk}$, is inspired by the Chi-square statistic used in the Chi-square test for goodness-of-fit, which is one of the well-established nonparametric hypothesis tests in categorical data analysis [1]. In statistics, goodness-of-fit tests are used to decide whether two distributions are significantly different from each other in shape/form. In relation to risk-sensitive evaluation, this means that, given a sample of topics, a risk measure based on Chi-square statistic permits quantifying the difference in the performance profiles of two IR systems across the topics. As mentioned above, none of the previously proposed risk measures are sensitive to the score distributions of IR systems on topics. However, risk-sensitive evaluation, by nature, should take into account all of shape, mean and variance, while Z$_{Risk}$ is independent of overall mean effectiveness. Hence, building upon Z$_{Risk}$, we propose the GeoRisk measure, which covers all of the aforementioned aspects including the overall mean effectiveness of the system at hand, as highlighted in Table 1.

3.1 The Chi-square Statistics & Z$_{\text{Risk}}$

Z$_{Risk}$ is best explained by deriving it directly from the Chi-square statistic used in the Chi-square test for goodness-of-fit. In particular, the Chi-square statistic is calculated over a data matrix of $r \times c$ cells, called the contingency table. The result of an IR experiment involving r systems and c topics can be represented by a $r \times c$ data matrix $\mathbf{X}$, whose rows and columns correspond respectively to the r systems and c topics, where the cells x_{ij} (for $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$) contain the observed performances of the corresponding systems for the associated topics, measured by an effectiveness measure such as ERR@20. Table 2 provides a graphical portrayal of data matrix $\mathbf{X}$.

For such a data matrix, the row and the column marginal totals are given by $S_i = \sum_{j=1}^{c} x_{ij}$ and $T_j = \sum_{i=1}^{r} x_{ij}$ respectively, and the grand total is given by $N = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}$. The average effectiveness of a system i over c topics is given by S_i/c and similarly, the within-topic mean system effectiveness is given by T_j/r.

Given a data matrix $\mathbf{X}$, the Chi-square statistic, G^2, can be expressed as

$$G^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}, \tag{4}$$

where the expected value for cell (i, j), e_{ij}, is given by

$$e_{ij} = \frac{S_i \times T_j}{N} = S_i \times \frac{T_j}{N} = S_i \times p_j. \tag{5}$$

In Equation (5), $p_j = \frac{T_j}{N}$ can be described as the density or mass of column j, for $j = 1, 2, \dots, c$. If a row total S_i is distributed on columns proportional to the column masses p_j, then the Chi-squared differences of the associated cell values from the corresponding expected values will sum up to zero, i.e. $\sum_{j=1}^{c} (x_{ij} - e_{ij})^2 = 0$. Note that $e_{ij} = x_{ij}$ when $r = 1$, where $p_j = x_{ij}/S_i$ since $N = S_i$. Intuitively, when there is only one IR system, the expected system performance for any topic j will be equal to the score observed for that system. When $r = 1$, $G^2 = 0$, meaning that the observed score distribution of the system across topics is perfectly fit to itself. Thus, G^2 values that are greater than zero indicate a discordance between two distributions, above or below expectations. This makes G^2 not directly applicable as a risk-sensitive evaluation measure, since it equally and uniformly penalises both downside (losses) *and* upside risk (wins). In contrast, risk-sensitive measures should favour wins and orthogonally penalise losses. Hence, we propose below a measure derived from G^2 that addresses this limitation.

For large samples, the Pearson's Chi-square statistic G^2 in Eq. (4) follows a Chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom and the observed cell values x_{ij} follow a *Poisson* distribution with mean e_{ij} and variance e_{ij} [1]. This means that the Chi-square statistic can also be expressed as the sum of the square of standard normal deviates [1]:

$$G^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} z_{ij}^2 \quad \text{where} \quad z_{ij} = \frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}}.$$

The square root of the components of Chi-square statistic, z_{ij}, gives the *standardised* deviation in cell (i, j) from the expected value e_{ij} (i.e. z-scores). Thus, for large samples, the distribution of z_{ij} values on the population can be approximated by the *standard normal* distribution with zero mean and unit variance.

It follows that a risk-reward tradeoff measure can be expressed in terms of the standard normal deviates from the expected effectiveness, as given by:

$$\text{Z}_{Risk} = \left[\sum_{q \in Q_+} z_{iq} + (1 + \alpha) \sum_{q \in Q_-} z_{iq} \right], \tag{6}$$

for any system i, $i = 1, 2, \dots, r$. Q_+ (Q_-) is the set of queries where $z_{iq} > 0$ ($z_{iq} < 0$, respectively), determined by whether system i outperforms its expectation on topic j (c.f. $x_{ij} - e_{ij}$).

Z$_{Risk}$ takes the classical form of a risk-sensitive evaluation measure, in that upside risk is rewarded and the effectiveness penalty of downside risk is amplified by α - i.e. the higher the Z$_{Risk}$, the more safe and less risky a system is. In addition, Z$_{Risk}$ calculates the risk of a system in relation to the shape of effectiveness across topics exhibited by multiple baselines. In this way, Z$_{Risk}$ brings a new dimension to the measurement of robustness, originally defined by

Voorhees [24] as *"the ability of the system to return reasonable results for every topic"*, in that for Z_{Risk}, robustness is measured compared to a per-topic expectation calculated from a population of baseline systems.

3.2 GeoRisk

As noted before, a limitation of Z_{Risk} is that it measures robustness irrespective of the mean effectiveness of IR systems. Indeed, one may consider that the baseline for any given system i is composed of the expected per-topic scores of the system, e_{ij}, such that the sum of expected per-topic scores is equal to the sum of the observed per-topic scores of the system, i.e. $\sum_j e_{ij} = S_i$. This means that Z_{Risk} measures robustness using individual baselines for every system, each of which is derived on the basis of the observed total effectiveness of the system (i.e. S_i) and the observed topic masses (i.e. T_j). This makes the robustness/risk measurements of Z_{Risk} independent of the observed mean effectiveness of the systems, i.e. $\sum_j x_{ij} = \sum_j e_{ij}$ for $i = 1, 2, \ldots, r$.

On the other hand, for the purposes of the comparative risk-sensitive evaluation of different IR systems, we combine the risk measure with the effectiveness measure in use, Z_{Risk} and ERR@20 for example, into a final measure. A natural method for such a combination is the *geometric mean*, which is expressed as the n^{th} root of the product of n numbers. The geometric mean is a type of average, like arithmetic mean, that represents the central tendency in a given set of numbers. In contrast to the arithmetic mean, the geometric mean normalises the ranges of the variables, so that each datum has an equal impact on the resulting geometric mean. Hence, the geometric mean of the ERR@20 scores and the Z_{Risk} scores represents, evenly, both the effectiveness and the robustness of system s_i under evaluation:

$$\mathrm{GeoRisk}\,(s_i) = \sqrt{S_i/c \times \Phi(Z_{Risk}/c)}, \tag{7}$$

where $0 \leq \Phi() \leq 1$ is the cumulative distribution function of the standard normal distribution. In this way, we use $\Phi()$ to normalise Z_{Risk} into [0,1], because $-\infty \leq Z_{Risk}/c \leq \infty$.

4. DEMONSTRATION

To illustrate Z_{Risk} and GeoRisk introduced in Section 3, Table 3 presents an example data matrix X composed of 8 systems and 5 topics. The effectiveness scores of the example systems are artificially determined so that the resulting performance profiles of the systems across the topics serve as a basis to exemplify some potential differences in performance profiles of IR systems in relation to their mean effectiveness. Figure 1 shows the performance profiles of the 8 systems, which can be characterised as follows:

- Systems $\mathbf{s}_1$ and $\mathbf{s}_2$ have the same mean effectiveness over the 5 topics (i.e. 0.3000) but the scores of $\mathbf{s}_1$ are monotonically increasing in magnitude across the topics, whereas, the scores of $\mathbf{s}_2$ are monotonically decreasing. That is, $\mathbf{s}_1$ and $\mathbf{s}_2$ have *contrasting* performance profiles across the topics, with respect to the same mean effectiveness score of 0.3000.
- Systems $\mathbf{s}_3$ and $\mathbf{s}_4$ have constant scores across the topics that are equal to their respective mean effectiveness scores. In other words, these systems have *constant* performance profiles, while system $\mathbf{s}_3$ has the same mean effectiveness as both $\mathbf{s}_1$ and $\mathbf{s}_2$.
- Systems $\mathbf{s}_5$ and $\mathbf{s}_6$ again have the same mean effectiveness as systems $\mathbf{s}_1$ and $\mathbf{s}_2$, but have alternating scores across the topics, such that one has a higher score in magnitude than the other for one topic and vice versa for the next topic. We describe such systems as having *alternating* performance profiles across the topics.
- Systems $\mathbf{s}_7$ and $\mathbf{s}_8$ have different mean effectiveness scores from each other and also from that of the other systems. Their performance profiles are visually parallel to each other, and *concordant* with the profile of the mean topic scores, i.e. the row "Mean" of Table 3.

	$\mathbf{t}_1$	$\mathbf{t}_2$	$\mathbf{t}_3$	$\mathbf{t}_4$	$\mathbf{t}_5$	S_i	Mean
$\mathbf{s}_1$	0.0500	0.1500	0.3000	0.4500	0.5500	1.5000	0.3000
$\mathbf{s}_2$	0.4000	0.3500	0.3000	0.2500	0.2000	1.5000	0.3000
$\mathbf{s}_3$	0.3000	0.3000	0.3000	0.3000	0.3000	1.5000	0.3000
$\mathbf{s}_4$	0.2500	0.2500	0.2500	0.2500	0.2500	1.2500	0.2500
$\mathbf{s}_5$	0.4000	0.1500	0.4000	0.1500	0.4000	1.5000	0.3000
$\mathbf{s}_6$	0.2000	0.4500	0.2000	0.4500	0.2000	1.5000	0.3000
$\mathbf{s}_7$	0.2542	0.2629	0.2802	0.2975	0.3061	1.4009	0.2802
$\mathbf{s}_8$	0.2918	0.2994	0.3147	0.3301	0.3378	1.5738	0.3148
T_j	2.1460	2.2123	2.3449	2.4776	2.5440	11.7248	
Mean	0.2683	0.2765	0.2931	0.3097	0.3180		

Table 3: Example data matrix X.

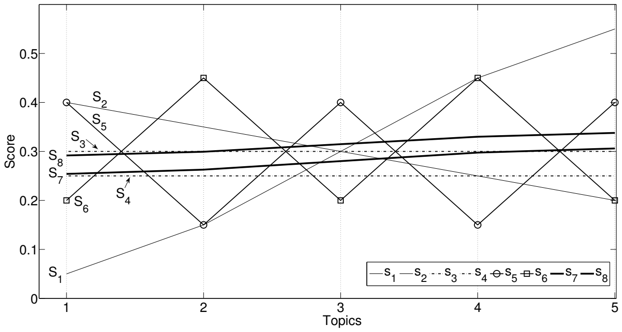

Figure 1: Example systems' performance profiles.

4.1 Single Baseline

Measuring the level of risk associated with a given IR system s with respect to a particular single baseline system b means that, in total, there are two systems under consideration, i.e. $r = 2$. For such a risk-sensitive evaluation, the Chi-square statistic G^2 is given by

$$G^2 = \sum_{j=1}^{c} \left[\frac{(x_{sj} - e_{sj})^2}{e_{sj}} + \frac{(x_{bj} - e_{bj})^2}{e_{bj}} \right],$$

and, under the null hypothesis that the observed score distributions of both systems follow a common distribution with mean μ and variance σ^2, it can be expressed as

$$\sum_{j=1}^{c} \frac{(x_{sj} - x_{bj})^2}{x_{sj} + x_{bj}}, \tag{8}$$

where x_{sj} is the observed score of the system s for topic j, and x_{bj} is the observed score of the baseline system b. Note that, when there are only two systems, $T_j = x_{sj} + x_{bj}$, and hence $x_{bj} = T_j - x_{sj}$ and $x_{sj} = T_j - x_{bj}$. Here,

$$\begin{aligned} e_{sj} &= \frac{S_s \times T_j}{N} = \frac{S_s}{N} \times (x_{sj} + x_{bj}) \\ e_{bj} &= \frac{S_b \times T_j}{N} = \frac{N - S_s}{N} \times (x_{sj} + x_{bj}) \end{aligned}$$

where $N = S_s + S_b$.

In fact, given two IR systems, the level of risk associated with any one of the two systems can be measured by taking the other system as the baseline, as implied by Eq. (8). Most importantly, Eq. (8) suggests, in this respect, that, if the

$\alpha = 0$	t_1	t_2	t_3	t_4	t_5	Z_{Risk}
s_2 vs. s_1	0.3689	0.2000	0.0000	-0.1690	-0.2858	0.1141
s_1 vs. s_2	-0.3689	-0.2000	0.0000	0.1690	0.2858	-0.1141
s_3 vs. s_1	0.2988	0.1581	0.0000	-0.1225	-0.1917	0.1427
s_1 vs. s_3	-0.2988	-0.1581	0.0000	0.1225	0.1917	-0.1427
s_4 vs. s_1	0.3077	0.1599	0.0000	-0.1209	-0.1884	0.1583
s_1 vs. s_4	-0.2809	-0.1460	0.0000	0.1103	0.1720	-0.1445
s_5 vs. s_1	0.3689	0.0000	0.0845	-0.2739	-0.1088	0.0708
s_1 vs. s_5	-0.3689	0.0000	-0.0845	0.2739	0.1088	-0.0708
s_6 vs. s_1	0.2121	0.2739	-0.1000	0.0000	-0.2858	0.1002
s_1 vs. s_6	-0.2121	-0.2739	0.1000	0.0000	0.2858	-0.1002
s_7 vs. s_1	0.2799	0.1422	0.0000	-0.1057	-0.1669	0.1496
s_1 vs. s_7	-0.2705	-0.1374	0.0000	0.1021	0.1613	-0.1446
s_8 vs. s_1	0.2792	0.1445	-0.0001	-0.1097	-0.1732	0.1408
s_1 vs. s_8	-0.2860	-0.1480	0.0001	0.1123	0.1774	-0.1442

Table 4: Single baseline example.

systems show an equal mean performance over a given set of c topics (i.e. $S_s = S_b$), the measured level of risk will be the same for both systems. In risk-sensitive evaluations, a baseline system defines what is a *robust* system, so that risk can be quantified as the degree of divergence from that baseline. However, given a set of IR systems, taking every system as a baseline, actually contributes information for the qualification of a robust (i.e. not 'risky' or safe) system on the population of topics. In this regard, multiple baselines can provide more information about the real level of risk associated with any IR system.

Let system s_1 in Table 3 be the baseline system. The level of risk associated with system s_2, which has the same mean performance with s_1, is $Z_{Risk} = 0.1141$, while the level of risk associated with s_1 for baseline s_2 is the same in magnitude but different in sign, i.e. -0.1141. The sign of the Z_{Risk} scores indicates the direction of the observed level of risk-reward tradeoff, where minus indicates down-side risk and plus indicates up-side risk. Table 4 shows the calculated values of Z_{Risk} at $\alpha = 0$ for each system $i = 2, 3, \ldots, 8$. As can be seen, for those systems whose mean performances are equal to the mean performance of s_1, only the sign of the calculated Z_{Risk} values changes when the baseline is swapped.

Based on the calculated Z_{Risk} values when the baseline is s_1, system s_4 is the least 'risky' system among the 8 example runs with the highest Z_{Risk} value of 0.1583 (i.e. the s_4 vs. s_1 row of the table). However, as can be seen in Figure 1, s_3, s_7, or s_8 are relatively less 'risky' than s_4. That is, those three systems have performance profiles that are concordant/parallel with that of s_1 and also they have relatively higher mean effectiveness scores than s_4: thus, s_4 could not be considered less "risky" than s_3, s_7, or s_8. The reason behind this counter-intuitive result is two-fold. Firstly, baseline system s_1 has performance scores that are monotonically increasing in magnitude across the topics. Thus, as a baseline, it suggests that the expected system performance on the population of topics that is represented by the sample topic t_1 would be low, and for the population of topics represented by t_2 it would be relatively higher than that of t_1, and so on. However, as seen from Figure 1, considering the observed scores of the other systems, it would appear that the expected per-topic system performances are in general different from those that the system s_1 suggests, i.e. the 'mean' row of Table 3. Secondly, the risk that is measured by Z_{Risk} is related to the distribution of the total system performance S_i on topics with respect to the expected per-topic system performances, and is not dependent on the magnitude of the mean performance of the systems across topics.

These two issues explain the above counter-intuitive result that s_4 is declared as the least 'risky' system. Indeed, the former issue can be resolved by employing multiple baselines as shown in the following Section 4.2, and the latter issue of independence from the magnitude of mean effectiveness can be resolved as shown in Section 4.3, where the risk measure Z_{Risk} and the measure of effectiveness are combined into a single measure of effectiveness, GeoRisk.

		$\alpha = 0$		$\alpha = 1$		$\alpha = 5$		$\alpha = 10$	
	Mean	Z_{Risk}	Geo	Z_{Risk}	Geo	Z_{Risk}	Geo	Z_{Risk}	Geo
s_1	0.300	-0.049	0.386	-0.727	0.364	-3.442	0.271	-6.835	0.160
s_2	0.300	0.026	0.388	-0.312	0.378	-1.668	0.333	-3.362	0.274
s_3	0.300	0.006	0.387	-0.069	0.385	-0.368	0.376	-0.742	0.364
s_4	0.250	0.005	0.354	-0.063	0.352	-0.336	0.344	-0.677	0.334
s_5	0.300	0.006	0.387	-0.541	0.370	-2.727	0.296	-5.460	0.203
s_6	0.300	0.005	0.387	-0.539	0.370	-2.718	0.297	-5.442	0.204
s_7	0.280	-0.001	0.374	-0.008	0.374	-0.036	0.373	-0.072	0.372
s_8	0.315	0.001	0.397	-0.010	0.396	-0.052	0.395	-0.106	0.393

Table 5: Z_{Risk} and GeoRisk for the example systems.

4.2 Multiple Baselines

Chi-square statistic allows the use of all systems in data matrix $\mathbf{X}$ as multiple baselines for risk-reward tradeoff measurements using Z_{Risk}. Recall that the expected value for cell (i, j), e_{ij}, is given by

$$e_{ij} = S_i \times \frac{T_j}{N} = S_i \times p_j.$$

For the case of a single baseline system b, given a particular system s, to calculate the mass or density p_j of topic j, the within topic total performance score T_j is taken as $x_{sj} + x_{bj}$, i.e. $p_j = (x_{sj} + x_{bj})/N$. Similarly, given a set of baselines, the topic masses can be calculated as

$$p_j = \frac{1}{N} \sum_{i=1}^{r} x_{ij},$$

for each topic $j = 1, 2, \ldots, c$. Intuitively this means that, given a set of r systems, the level of risk associated with every system is measured by taking the remaining $(r - 1)$ systems as the baseline. Compared to the case of taking a particular system as the baseline, as the number of baseline systems increases, the accuracy of the estimates of expected system performance for each topic increases, and hence the accuracy of the estimates of real risk increases.

Table 5 shows the calculated Z_{Risk} values for each of the 8 example runs at $\alpha = 0, 1, 5, 10$. We observe from the table that, as the risk sensitivity parameter α increases, example systems s_7 and s_8 exhibit the lowest levels of risk relative to the other systems, (i.e. $\alpha = 1, 5, 10$), while s_1 exhibits the highest level of risk ($Z_{Risk} = -6.835$ at $\alpha = 10$). As can be seen, using multiple baselines resolves the effect of the lack of information about the expected per-topic system performance in assessing the risk levels of systems, i.e. s_4 vs. s_7 and s_4 vs s_8. In the following section, we show how to combine Z_{Risk} with mean system effectiveness in order to solve the last issue about Z_{Risk}, i.e. the counter-intuitive case of s_4 vs. s_3, where the measured level of risk for s_3 is higher than that of s_4 (e.g. the Z_{Risk} score of s_3 is -0.368 and it is -0.336 for s_4 at $\alpha = 5$), while s_3 has higher effectiveness score than s_4 (i.e. 0.300 vs. 0.250) and it has also a performance profile concordant with that of s_4.

4.3 Effectiveness vs. Risk

Table 5 shows the calculated GeoRisk values for each of the 8 example runs. As can be seen, for the case of s_4 vs. s_3, the issue of the independence of Z_{Risk} measurements from the magnitude of the mean effectiveness of IR systems is solved. The example system s_3 is now measured as less

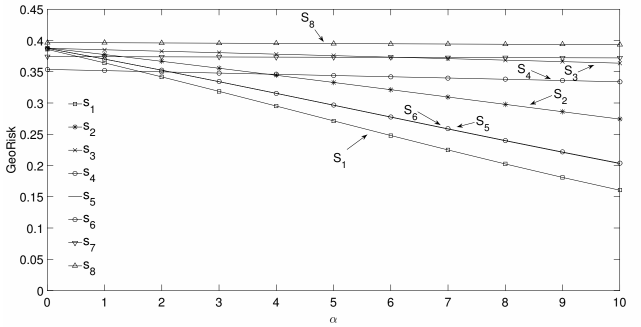

Figure 2: Example systems' GeoRisk as $0 \leq \alpha \leq 10$.

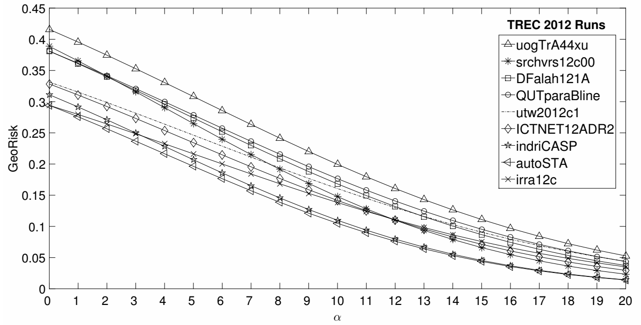

Figure 3: GeoRisk plot for 8 TREC 2012 runs.

'risky' than s_4, as suggested by the magnitude of the observed mean effectiveness scores.

Figure 2 shows the plot of GeoRisk scores for each example system for $\alpha = 0, 1, 2, \ldots, 10$, where the systems with lines sloping downward along the increasing values of α (i.e. x-axis) are those that exhibit a risk of abject failure, (i.e. s_1, s_2, s_5, and s_6) while, in contrast, the robust systems such as s_3, s_4, s_7 and s_8 have nearly a straight, horizontal lines.

In summary, the GeoRisk measure takes into account both the mean effectiveness of IR systems and the difference in the shapes of their performance profiles. As a result, GeoRisk is sensitive to mean (i.e. the component S_i/c), variance and the shape of the observed effectiveness scores across topics (i.e. the Z_{Risk} component).

5. ANALYSIS OF TREC DATA

In this section, we demonstrate the use of the risk-reward tradeoff measure derived from the Chi-square statistic, Z_{Risk}, and the aggregate measure GeoRisk, on real systems submitted to the TREC 2012 Web track [10][1], in comparison with the existing measures U_{Risk} and T_{Risk}. In the subsequent year of the Web track [12], a standard baseline called *indriCASP* and based on the *Indri* retrieval platform was provided. Similar to [13, 14], we use the same indriCASP system as the nominal single baseline on the 2012 Web track topics.

In particular, out of the 48 runs submitted to TREC 2012, we select the top runs of the highest 8 performing groups, based on the mean ERR@20 score, While we omit other submitted runs for brevity, the following analysis would be equally applicable to them. For each run, we report the risk-reward tradeoff scores obtained using the official TREC 2012 evaluation measure, ERR@20.

Table 6 lists the U_{Risk}, T_{Risk}, T^*_{Risk}, Z_{Risk} and GeoRisk risk-reward tradeoff scores for the 8 runs. For the measures U_{Risk} and T_{Risk}, the baseline run is *indriCASP*; for the measures Z_{Risk} and GeoRisk, we use as multiple baselines all 48+1 TREC 2012 runs including *indriCASP*; T^*_{Risk} denotes T_{Risk} calculated using the per-topic mean effectiveness of the 49 runs as the baseline, i.e. Mean_q in Eq. (3). Note that Dinçer et al. [13] showed that T_{Risk} is inferential, i.e. the T_{Risk} scores correspond to scores of the Student's t statistic. For this reason, for the U_{Risk} scores in TREC 2012 in Table 6 (where $c = 50$), $\mathrm{T}_{Risk} > \pm 2$ indicates that the observed U_{Risk} score exhibits a significant level of risk.

Table 6 shows in general that the notion of risk quantified by the Chi-square statistic-based risk measure Z_{Risk} differs from that of the U_{Risk}, T_{Risk} and T^*_{Risk} measures, as illustrated by the contrasting systems' rankings (the column R next to each measure) for Z_{Risk}. In particular, at $\alpha = 0$, U_{Risk} and T_{Risk} agree with the effectiveness measure ERR@20 on the rankings of the 8 TREC 2012 runs. However, at $\alpha = 5$, although U_{Risk} and T_{Risk} still agree with each other, they both diverge from the agreement with ERR@20. On the other hand, the risk measure Z_{Risk} agrees neither with ERR@20 nor with the risk measures U_{Risk} and T_{Risk}. Note that, except for the determination of baselines, the three risk measures U_{Risk}, T_{Risk}, and Z_{Risk} rely on the same notion of risk and reward, i.e. down-side risk and up-side risk. Thus, comparing Z_{Risk} with U_{Risk} and T_{Risk}, it follows that multiple baselines (i.e. 49 TREC 2012 runs) provide information that is different from the information provided by the single baseline system *indriCASP*.

According to Z_{Risk}, the most robust run is "uogTrA44xu" with a Z_{Risk} value of 0.962 at $\alpha = 0$, and the next is "irra12c" with $\mathrm{Z}_{Risk} = 0.265$, and so on, given the expected per-topic performance scores representing the baselines for each system. Based on the definition of Z_{Risk}, it is expected that "uogTrA44xu" would perform any given topic with an ERR@20 score that is better than or equal to the expected score for that topic on a population of systems with mean ERR@20 scores equal to 0.3406. Conversely, the least robust or most 'risky' run is "srchvrs12c00" with a $\mathrm{Z}_{Risk} = -0.912$.

Recall that Z_{Risk} is independent of the observed mean effectiveness scores of the systems, which is, by definition, inappropriate for the purpose of a comparative IR evaluation. Thus, as an aggregate measure, GeoRisk, the geometric mean of Z_{Risk} and ERR@20, can be used to tackle this challenge. As can be seen in Table 6, GeoRisk agrees with ERR@20 at $\alpha = 0$ on the rankings of the 8 TREC 2012 runs. Here, GeoRisk gives equal weights to ERR@20 and Z_{Risk}, and similarly, at $\alpha = 0$, Z_{Risk} gives equal weights to down-side risk and up-side risk. Thus, the observed agreement between GeoRisk and ERR@20 implies that the measured Z_{Risk} scores for each of the 8 TREC 2012 runs at $\alpha = 0$ are negligible compared to the observed differences in effectiveness between the runs. In other words, every TREC 2012 run exhibits risk, to a certain extent, but none of the measured risk levels are high enough to compensate for the observed difference in mean effectiveness between two systems, so that a swap between risk and reward for a given topic is likely to occur for two systems on the population of topics. Note that the agreements of T_{Risk}, as well as U_{Risk}, at $\alpha = 0$, with ERR@20 also give support in favour of the same conclusion, i.e. the practical insignificance of the measured levels of risk at $\alpha = 0$.

On the other hand, as α increases (i.e. as the emphasis of down-side risk increases in Z_{Risk} measurements), GeoRisk

[1] Although our analysis is equally applicable to the TREC 2013 Web track, due to the lack of space, we report results from TREC 2012, which are also directly comparable to that of previous works [13, 14].

		$\alpha = 0$							$\alpha = 5$										$\alpha = 20$	
	ERR	U_{Risk}	T_{Risk}	T^*_{Risk}	Z_{Risk}	R	Geo	R	U_{Risk}	R	T_{Risk}	R	T^*_{Risk}	R	Z_{Risk}	R	Geo	R	Geo	R
uogTrA44xu	0.3406	0.146	$\underline{2.822}$	4.833	0.962	1	0.4158	1	-0.130	2	-0.798	2	2.767	1	-29.255	1	0.308	1	0.053	1
srchvrs12c00	0.3067	0.112	$\underline{2.332}$	2.985	-0.912	9	0.3887	2	-0.100	1	-0.673	1	-0.678	2	-37.069	7	0.265	4	0.024	7
DFalah121A	0.2920	0.097	$\underline{2.290}$	2.895	-0.328	7	0.3811	3	-0.156	3	-0.981	3	-0.861	3	-32.635	5	0.274	3	0.037	4
QUTparaBline	0.2901	0.095	$\underline{2.130}$	2.870	0.004	4	0.3809	4	-0.189	4	-1.112	4	-1.006	4	-30.966	4	0.279	2	0.044	3
utw2012c1	0.2203	0.026	0.561	0.917	-0.172	6	0.3314	5	-0.388	6	$\underline{-2.046}$	6	-2.987	5	-29.807	2	0.246	5	0.044	2
ICT. . . DR2	0.2149	0.020	0.487	0.569	0.233	3	0.3284	6	-0.329	5	-1.994	5	-3.238	6	-32.887	6	0.234	6	0.030	6
indriCASP	0.1947	*	*	-0.120	-0.339	8	0.3111	7	*	*	*		*	*	-38.619	9	0.207	8	0.014	8
autoSTA	0.1735	-0.021	-0.498	-0.968	-0.143	5	0.2942	8	-0.509	8	$\underline{-2.518}$	7	-4.195	7	-38.215	8	0.196	9	0.014	9
irra12c	0.1723	-0.022	-0.545	-1.214	0.265	2	0.2942	9	-0.501	7	$\underline{-2.634}$	8	-4.510	8	-30.410	3	0.216	7	0.035	5

Table 6: U_{Risk}, T_{Risk}, T^*_{Risk} and Z_{Risk} risk-reward tradeoff scores for the top 8 TREC 2012 runs, along with GeoRisk **at $\alpha = 0, 1, 5, 10, 20$. For U_{Risk} and T_{Risk} the baseline is *indriCASP*, and for T^*_{Risk} it is Mean_q in Eq. (3) over all $48 + 1$ TREC 2012 runs including *indriCASP*, and for Z_{Risk} and** GeoRisk**, the baselines are estimated for the 8 runs over the same set of 49 runs and 50 Web track topics. The underlined T_{Risk} scores correspond to those U_{Risk} scores for which a two-tailed paired t-test gives significance with p-value < 0.05 - i.e. $\mathrm{T}_{Risk} > \pm 2$.**

diverges from ERR@20 and ranks the 8 TREC 2012 runs in a way that is different from all of the risk measures under consideration including Z_{Risk} (for example, $\alpha = 5$), and at a high value of α, it converges to an agreement with Z_{Risk}, e.g. the systems' rankings in Table 6 at $\alpha = 20$ for GeoRisk and at $\alpha = 5$ for Z_{Risk}. The tendency of GeoRisk towards Z_{Risk} as α increases is expected from the definition of GeoRisk.

Figure 3 plots the GeoRisk scores calculated for each run at $\alpha = 0, 1, \ldots, 20$. As can be seen, each of the 8 TREC 2012 runs has a decreasing GeoRisk score in magnitude, as the risk sensitivity parameter α increases. This means in general that - to a varying extent - every run under evaluation is subject to the risk of committing an abject failure, as the importance of getting a reasonable result for every topic increases. In particular, the runs "uogTrA44xu" and "ICTNET12ADR2" keep their relative positions in the observed rankings across all α values, while the ranks of the other runs considerably change. For example, the rank of "srchvrs12c00" changes from 2 at $\alpha = 0$ to 7 at $\alpha = 5$. At $\alpha = 0$, the run with rank 7 is *indriCASP*. However, the calculated risk for "srchvrs12c00" at any level of α cannot be considered as empirical evidence to favour *indriCASP* over "srchvrs12c00" for any given topic from the population, since the mean effectiveness of "srchvrs12c00" is significantly higher than the mean effectiveness of *indriCASP* ($p < 0.0239$, paired t-test).

Note that, for two IR systems whose mean effectiveness scores are significantly different from each other, a measured risk level could have no particular meaning from a user perspective. This is because the system with higher mean effectiveness would be the one that can fulfil the users' information needs better than the other on average, no matter what level of risk is associated with it. The system with significantly low mean effectiveness would, on average, fail to return a "reasonable" result for any given topic, compared to the other system's effectiveness. For a declared significance with a p-value of 0.05, a swap in scores between the two systems for a topic (i.e. a transition from risk to reward or vice versa between the systems) is likely to occur 5% of the time on average [26].

Nevertheless, the same case is not true for runs "DFalah121A" and "QUTparaBline", whose observed mean effectiveness scores are not significantly different from the mean effectiveness of "srchvrs12c00". The paired t-test, which is performed at a significance level of 0.05, fails to give significance to the observed difference in mean effectiveness between "DFalah121A" and "srchvrs12c00" with a p-value of 0.7592, and similarly for "QUTparaBline" with a p-value of 0.7003. This means that, for a given topic, a transition from risk to reward, or vice versa, between the runs "DFalah121A" and "srchvrs12c00", or between runs "QUTparaBline" and "srchvrs12c00", is highly likely to occur on the population of topics. Thus, both systems can be considered less "risky" or more robust than "srchvrs12c00".

In summary, this analysis of the TREC 2012 Web track runs demonstrates the suitability of GeoRisk for balancing risk-sensitivity and overall effectiveness, and the importance of using multiple baselines within the appropriate statistical framework represented by Z_{Risk}. The analysis performed for the 8 TREC runs shows overall that, in a comparative IR evaluation effort, relying only on the observed mean effectiveness of the systems may be misleading, even for a best performer system like "srchvrs12c00", where the risk associated with such a system is high enough that it can cause an over-estimation of the real performance of the system. However, we showed that GeoRisk provides a solution for identifying systems exhibiting such risks.

6. RISK-SENSITIVE OPTIMISATION

In this section, we show how GeoRisk can be integrated within the state-of-the-art LambdaMART learning to rank technique [7, 28]. Indeed, Wang et al. [27] showed how their U_{Risk} measure could be integrated within LambdaMART. Similarly, Dinçer et al. [13] proposed variants of T_{Risk} that resulted in learned models that exhibited less risk.

The method of integration of risk-sensitive measures into LambdaMART requires adaptation of its objective function. In short, LambdaMART's objective function is a product of (i) the derivative of a cross-entropy that was originally defined in the RankNet learning to rank technique [6], based on the scores of two documents a and b, and (ii) the absolute change ΔM_{ab} in an evaluation measure M due to the swapping of documents a and b [28]. Various IR measures (e.g. NDCG) can be used as M, as long as the measure is *consistent*: for each pair of documents a and b with differing relevance labels, making an *"improving swap"* (moving the higher labelled document above the lesser) must result in $\Delta M_{ab} \geq 0$, and orthogonally for *"degrading swaps"*.

In adapting LambdaMART to be more robust within a risk-sensitive setting, the ΔM is replaced by a variant $\Delta M'$ that considers the change in risk observed by swapping documents a and b, according to the underlying evaluation measure M, e.g. NDCG. In the following, we summarise existing instantiations of $\Delta M'$ arising from U_{Risk} and T_{Risk} (called U-CRO, T-SARO and T-FARO), followed by our proposed instantiation of GeoRisk within LambdaMART.

U-CRO: Constant Risk Optimisation based upon the U_{Risk} measure [27] (U-CRO) maintains a constant risk-level, regardless of the topic. In particular, let M_m define the effectiveness of the learned model m according to measure M, and let M_b define the effectiveness of the baseline b. Correspondingly, $M_m(j)$ ($M_b(j)$) is the effectiveness of the learned model (baseline) for query j, then:

$$\Delta M' = \begin{cases} \Delta M & \text{if } M_m(j) + \Delta M \geq M_b(j); \\ \Delta M \cdot (1+\alpha) & \text{otherwise.} \end{cases} \quad (9)$$

T-SARO & T-FARO: Adaptive Risk-sensitive Optimisation makes use of the fact that T_{Risk} can identify queries that exhibit real (significant) levels of risk [13] compared to the baseline b. Dinçer et al. [13] proposed two *Adaptive* Risk-sensitive Optimisation adaptations of LambdaMART, namely T-SARO and T-FARO, which use this observation to focus on improving those risky queries. Indeed, in U-CRO, ΔM is multiplied by $(1+\alpha)$, for a static $\alpha \geq 0$. In T-SARO and T-FARO, α is replaced by a variable α', which varies according to the probability of observing a query with a risk-reward score greater than that observed. By modelling this probability using the standard normal cumulative distribution function $Pr\left(Z \geq \mathrm{T}_{R_j}\right) \approx 1 - \Phi\left(\mathrm{T}_{R_j}\right)$, T-SARO replaces the original α in Eq. (9) with α' as:

$$\alpha' = \left[1 - \Phi\left(\mathrm{T}_{R_j}\right)\right] \cdot \alpha, \quad (10)$$

where $\mathrm{T}_{R_j} = \delta_j / SE(\mathrm{U}_{Risk})$ determines the level of risk exhibited by topic j. T-SARO and T-FARO contrast on the topics for which α' is adjusted – while T-SARO only adjusts topics with downside risk as per Eq. (9), T-FARO adjusts all topics:

$$\Delta M' = \Delta M \cdot (1 + \alpha')$$

The experiments in [13] found that T-FARO exhibited higher effectiveness than T-SARO, thus we compare GeoRisk to only U-CRO and T-FARO in our experiments.

GeoRisk: Our adaptation of ΔM for the newly proposed GeoRisk is more straightforward than the T_{Risk} Adaptive Risk-sensitive Optimisations, in that we use GeoRisk directly as the measure within LambdaMART.

$$\begin{aligned} \Delta M' &= GeoRisk(M_m + \Delta M) - GeoRisk(M) \\ &= \sqrt{(S_i + \Delta M)/c \times \Phi(Z_{Risk}/c)} - GeoRisk(M) \end{aligned}$$

Indeed, GeoRisk is suitable for LambdaMART as it exhibits the consistency property: an improving 'swap' will increase both the left factor (Si/c) and the right factor Z_{Risk} and therefore the value of GeoRisk for $M_m + \Delta M$ will likewise increase. Moreover, as ΔM is calculated repeatedly during the learning process, the speed of the implementation is critical for efficient learning. In this respect, it is important to note that retaining the values of the separate z_{ij} summands for each query in the Z_{Risk} calculation (see Equation (6)) allows the new Z_{Risk} value to be calculated by only recomputing the z_{ij} for the query affected, then recalculating GeoRisk.

Recall from Section 3 that z_{ij} encapsulates differential weighting of downside and upside risks, but with respect to the expected performance on the topic. Hence, by using GeoRisk, the objective function of LambdaMART will favour learned models that make improving swaps of documents on topics where the learned model performs below expectation as calculated on the set of baselines.

Naturally, the instantiation of GeoRisk within a learning to rank setting depends on the set of baselines $\mathbf{X}$, to allow the estimation of the topic expectations e_{ij} (see Eq. (5)). The choice of baselines to provide for learning can impact upon which topics the learner aims to improve. Previous works on risk-sensitive learning [13, 27] have used the performance of a single BM25 retrieval feature as the baseline. Indeed, single weighting models such as BM25 are typically used to identify the initial set of documents, which are then re-ranked by the learned model [19, 20], and hence it is a baseline that the learner should outperform. However, it does not represent the typical performance of a learned approach upon the queries, as it cannot encapsulate the effectiveness of refined ranking models using many other features.

Hence, instead of using GeoRisk to learn a model more effective than a set of BM25-like baselines, we argue for the use of state-of-the-art baselines, which portray representative estimations of query difficulty to the learner. Such baselines are more akin to the systems submitted to TREC (which themselves have been trained on previous datasets), rather than a single weighting model feature. In a deployed search engine, such state-of-the-art-baselines could represent the effectiveness of the currently deployed search engine, or other deployed search engines for the same query. In an experimental setting, such as in this paper, we use held-out data to pre-learn several learned models before conducting the main comparative experiments. Finally, for comparison purposes, we also deploy T*-FARO in our experiments, where the *mean* performance of the state-of-the-art baselines for each topic is used as the *single* learning baseline.

6.1 Experimental Setup

Our experiments use the open source Jforests learning to rank tool [15][2], which implements U-CRO, and T-FARO, as well as plain LambdaMART. Our implementations of T*-FARO & GeoRisk are also built upon Jforests[3]. As baselines, in addition to LambdaMART, we also deploy a plain gradient boosted regression trees learner (also implemented by Jforests), and two linear learned models, Automatic Feature Selection (AFS) [21] & AdaRank [19, Ch. 4].

We conduct experiments using the MSLR-Web10k dataset[4]. This learning to rank dataset contains 136 query-dependent and query-independent feature values for documents retrieved for 10,000 queries, along with corresponding relevance labels. As highlighted above, our baselines require separate training. For this reason, we hold out 2000 queries for initial training (two thirds) and validation (one third). The remaining 8000 queries are then split into 5 folds, each with a balance of 60% queries for training, 20% for validation, and 20% for testing. Hence, our reported results are not comparable to previous works using all 10000 queries of the MSLR dataset, but instead performances for LambdaMART, U-CRO & T-FARO are presented on the 8000 queries. The underlying evaluation measure M in each learner is NDCG.

For GeoRisk & T*-FARO, the multiple baselines are evaluated for each query in the main 5 folds, which represent 'unseen' queries for those systems. For U-CRO, T-SARO and T-FARO, the baseline is depicted by the performance of the `BM25.wholedocument` feature.

Finally, we note that LambdaMART has several hyper-parameters, namely the minimum number of documents in each leaf m, the number of leaves l, the number of trees in the ensemble nt and the learning rate r. Our experiments use a uniform setting for all parameters across all folds, namely $m = 1500$, $l = 50$, $nt = 500$, $l = 0.1$, which are similar to those reported in [27] for the same dataset.

6.2 Results

In Table 7, we report the NDCG@10 effectiveness and robustness for LambdaMART, U-CRO, T-FARO, T*-FARO

[2] https://github.com/yasserg/jforests

[3] We have contributed GeoRisk as open source to Jforests.

[4] http://research.microsoft.com/en-us/projects/mslr/

and GeoRisk for $\alpha = \{1, 5\}^5$. The table is split into two halves: comparison to the effectiveness of the single BM25 baseline, and comparison viz. the 4 baseline learned models. For each of the baselines, we report the reward to risk ratio (denoted "Reward/Risk"), which measures the gain over the effectiveness of the baseline. Similarly, the win to loss ratio (denoted "W/L") measures the number of queries that the risk-sensitive optimisation contributed to reward against risk. Finally, the number of queries that each model wins or looses relative to the baseline are also shown for each α value, along with the number of queries that experience a relative loss greater than 20% NDCG@10. For clarity, the header of Table 7 denote arrows to show the favourable direction of each measure, e.g. $\uparrow$ means that higher is better.

On analysis of the top half of Table 7, we observe that GeoRisk generates the highest mean NDCG effectiveness, marginally improving over LambdaMART. This is also observable for the Reward measure, in comparison to the BM25 baseline. However, for the risk measures, T-FARO and U-CRO obtain the best values. For α, $\alpha = 1$ is deemed the appropriate setting across all risk-sensitive learners, which has the effect of doubling the penalty of a query underperforming the corresponding baseline for that learner.

In the bottom half of Table 7, we examine the performance profiles of the different learners compared to the effectiveness of the 4 learned model baselines - the measures reported are the macro-average, i.e. the mean when each measure is calculated with respect to each learned baseline in turn (rather than compared to the micro-averaged effectiveness of the 4 learned baselines). In this half of the table, we note that, for $\alpha = 1$, GeoRisk demonstrates the highest Reward and number of Wins and lowest Losses (and the resulting best Reward/Risk & Win/Loss ratios (the latter is a 2.7% improvement over LambdaMART). These improvements in the risk profile of the systems are achieved while still attaining the highest mean NDCG effectiveness among the systems. All differences are statistically significant ($n = 8000$ queries).

Finally, we note that the effectiveness and risk profiles attained by T*-FARO are markedly different, with T*-FARO attaining the lowest Reward/Risk & Win/Loss ratios. This verifies that the use of expected topic performance by GeoRisk rather than a mean per-topic performance (as used by T*-FARO) results in a learned model more attuned to the normal performances of state-of-the-art baseline systems.

Overall, this empirical evidence confirms our claim that the new risk-sensitive objective function GeoRisk for the LambdaMART learning to rank technique allows effective yet robust learned models to be obtained using multiple baselines. Moreover, we would highlight the more impressive risk-profiles attained in the bottom half of Table 7, which demonstrate that given a set of state-of-the-art baselines, using GeoRisk can generate an effective model that is as effective as LambdaMART with better risk profiles, and allows learning-to-rank to benefit from natural incremental improvements as practiced in real deployment settings.

7. RELATED WORK

One aspect of this paper is the assessment of the robustness of IR systems, initiated first by the TREC Robust track [24] based on the geometric mean average precision measure [23, 25], and developed further by the introduction of new measures of risk/robustness, such as U_{Risk} [27] and T_{Risk} [13]. In this respect, while Dinçer et al. [13] investigated the use of the Student's t-test for risk-sensitive evaluation, this is the first work to investigate the use of Pearson's Chi-square statistic for risk-sensitive evaluation, thereby facilitating the use of multiple baselines. Instead, previous usages of the Chi-square statistic has encompassed index term weighing [18] and document classification [22].

In IR experimentation, Armstrong et al. [3] noted that many papers appeared to show improvement upon older, weaker baselines. More recent work by Kharazmi et al. [17] showed that testing upon state-of-the-art baselines is necessary to demonstrate an advance. Similarly, this paper advocates the use of multiple state-of-the-art baselines, both in experimental and learning settings.

We also note several attempts to develop robust learning to rank techniques: of note, the AdaRank technique [19, Ch. 4] focuses on improving hard queries using boosting. Since then, risk-sensitive optimisation techniques such as U-CRO [27], T-SARO & T-FARO [13] have aimed to adapt the LambdaMART technique by identifying risky topics with respect to a single baseline. Our work goes further by making use of multiple state-of-the-art baseline systems when calculating risk-estimation in the learning to rank objective function. Finally, Bennett et al.[5] take a different route, by developing personalised risk-averse ranking strategies upon the LambdaMART technique. As they build upon U-CRO, it may be possible to combine both approaches in the future.

8. CONCLUSIONS

This is the first paper that thoroughly investigated the use of multiple baselines in risk-sensitive evaluation. It argued for a new definition of risk-sensitivity related to the expected performance upon a given topic, calculated from a population of existing baseline systems. In particular, the paper introduced two new risk-sensitive evaluation measures, Z_{Risk} and GeoRisk that are based upon the Chi-square statistic. Moreover, while Z_{Risk} estimates risk independent of the overall mean retrieval effectiveness, GeoRisk enhances Z_{Risk} by additionally accounting for overall effectiveness.

Our new measures were demonstrated on the results of a comparative system evaluation from the TREC Web track. Finally, the paper showed how the proposed GeoRisk measure can be directly integrated within the objective function of the state-of-the-art LambdaMART learning to rank technique. Experiments upon 8000 queries from a learning to rank dataset showed that the resulting learned models were as effective as LambdaMART, but also more risk-averse when compared to four learned baselines.

9. ACKNOWLEDGMENTS

This work is partially supported by TUBITAK, The Scientific and Technological Research Council of Turkey (Project No: 114E558). All opinions are that of the authors.

10. REFERENCES

[1] A. Agresti. *Categorical Data Analysis*. Wiley, 2002.

[2] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Proceedings of ECIR*, 2004.

[3] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of ACM CIKM*, 2009.

[4] S. Beitzel, E. Jensen, and O. Frieder. GMAP. In L. Liu and M. Özsu, eds., *Encyclopedia of Database Systems*, pp 1256–1256, 2009.

[5] α=0 is equivalent to the normal LambdaMART algorithm.

	α	NDCG↑	Reward↑	Risk↓	Reward/Risk↑	Wins↑	Losses↓	W/L↑	L > 20%↓
Compared to BM25 Baseline									
L'MART	0	0.4578	0.195	0.036	5.401	5892	1715	3.44	982
U-CRO	1	0.4558	0.192	0.035	5.472	5913	1684	3.51	965
U-CRO	5	0.4461	0.180	**0.032**	5.552	5913	**1666**	**3.55**	**880**
T-FARO	1	0.4576	0.194	0.035	**5.568**	**5928**	1671	**3.55**	963
T-FARO	5	0.4569	0.194	0.036	5.435	5889	1719	3.43	995
T*-FARO	1	0.4557	0.194	0.037	5.25	5857	1753	3.34	1016
T*-FARO	5	0.4562	0.194	0.037	5.26	5861	1737	3.37	999
GeoRisk	1	**0.4581**	**0.196**	0.037	5.327	5868	1732	3.39	992
GeoRisk	5	0.4557	0.195	0.038	5.179	5876	1728	3.40	1037
Compared to 4 Learned Models Baselines									
L'MART	0	0.4578	**0.061**	0.053	1.138	3978.3	3669.0	1.12	1722.5
U-CRO	1	0.4558	0.059	0.053	1.103	3954.5	3690.0	1.11	1718.8
U-CRO	5	0.4461	0.051	0.056	0.941	3748.8	3887.3	1.01	1825.8
T-FARO	1	0.4576	0.059	**0.052**	1.137	3965.5	3674.3	1.12	**1677.0**
T-FARO	5	0.4569	0.059	**0.052**	1.121	3967.3	3676.5	1.12	1692.0
T*-FARO	1	0.4557	0.042	0.051	0.820	3562.0	4079.5	0.87	1821.5
T*-FARO	5	0.4562	0.041	0.053	0.823	3539.0	4099.5	0.86	1783.5
GeoRisk	1	**0.4581**	**0.061**	0.054	**1.142**	**4017.0**	**3628.8**	**1.15**	1709.8
GeoRisk	5	0.4557	0.060	0.055	1.095	3974.0	3674.0	1.12	1756.5

Table 7: Learning to rank results, with risk results calculated w.r.t. BM25 & the 4 learned models. All differences are statistically significant over the $n = 8000$ queries.

[5] P. N. Bennett, M. Shokouhi, and R. Caruana. Implicit preference labels for learning highly selective personalized rankers. In *Proceedings of ACM ICTIR*, 2015.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of ICML*, 2005.

[7] C. J. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

[8] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer. Automatic query refinement using lexical affinities with maximal information gain. In *Proceedings of ACM SIGIR*, 2002.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of ACM CIKM*, 2009.

[10] C. L. A. Clarke, N. Craswell, and E. Voorhees. Overview of the TREC 2012 Web track. In *Proceedings of TREC*, 2012.

[11] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of ACM CIKM*, 2009.

[12] K. Collins-Thompson, P. Bennett, F. Diaz, C. Clarke, and E. M. Voorhees. Overview of the TREC 2013 Web track. In *Proceedings of TREC*, 2013.

[13] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proceedings of ACM SIGIR*, 2014.

[14] B. T. Dinçer, I. Ounis, and C. Macdonald. Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In *Proceedings of ECIR*, 2014.

[15] Y. Ganjisaffar, R. Caruana, and C. Lopes. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceedings of ACM SIGIR*, 2011.

[16] D. Hoaglin, F. Mosteller, and J. Tukey, eds. *Understanding robust & exploratory data analysis.* Wiley, 1983.

[17] S. Kharazmi, F. Scholer, D. Vallet and M. Sanderson. Examining Additivity and Weak Baselines. *TOIS*, to appear, 2016.

[18] I. Kocabaş, B. T. Dinçer, and B. Karaoğlan. A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. *Information Retrieval*, 17(2):153–176, 2014.

[19] T.-Y. Liu. Learning to rank for information retrieval. *Foundation and Trends in Information Retrieval*, 3(3):225–331, 2009.

[20] C. Macdonald, R. L. Santos, and I. Ounis. The whens and hows of learning to rank for web search. *Information Retrieval.*, 16(5):584–628, 2013.

[21] D. A. Metzler. Automatic feature selection in the markov random field model for information retrieval. In *Proceedings of ACM CIKM*, 2007.

[22] M. Oakes, R. Gaaizauskas, H. Fowkes, A. Jonsson, V. Wan, and M. Beaulieu. A method based on the chi-square test for document classification. In *Proceedings of ACM SIGIR*, 2001.

[23] S. Robertson. On GMAP - and other transformations. In *Proceedings of ACM CIKM*, 2006.

[24] E. M. Voorhees. Overview of the TREC 2003 Robust retrieval track. In *Proceedings of TREC*, 2003.

[25] E. M. Voorhees. The TREC Robust retrieval track. *SIGIR Forum*, 39(1):11–20, June 2005.

[26] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of ACM SIGIR*, 2002.

[27] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proceedings of ACM SIGIR*, 2012.

[28] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao. Ranking, boosting, and model adaptation. Technical Report MSR-TR-2008-109, Microsoft, 2008.