Hypothesis Testing for the Risk-Sensitive Evaluation of Retrieval Systems

B. Taner Dinçer
Dept of Statistics & Computer Engineering
Mugla University
Mugla, Turkey
dtaner@mu.edu.tr

Craig Macdonald and Iadh Ounis
School of Computing Science
University of Glasgow
Glasgow, UK
{firstname.lastname}@glasgow.ac.uk

ABSTRACT

The aim of risk-sensitive evaluation is to measure when a given information retrieval (IR) system does not perform worse than a corresponding baseline system for any topic. This paper argues that risk-sensitive evaluation is akin to the underlying methodology of the Student's t test for matched pairs. Hence, we introduce a risk-reward tradeoff measure T_{Risk} that generalises the existing U_{Risk} measure (as used in the TREC 2013 Web track's risk-sensitive task) while being theoretically grounded in statistical hypothesis testing and easily interpretable. In particular, we show that T_{Risk} is a linear transformation of the t statistic, which is the test statistic used in the Student's t test. This inherent relationship between T_{Risk} and the t statistic, turns risk-sensitive evaluation from a descriptive analysis to a fully-fledged *inferential* analysis. Specifically, we demonstrate using past TREC data, that by using the inferential analysis techniques introduced in this paper, we can (1) decide whether an observed level of risk for an IR system is statistically significant, and thereby infer whether the system exhibits a *real* risk, and (2) determine the topics that individually lead to a significant level of risk. Indeed, we show that the latter permits a state-of-the-art learning to rank algorithm (LambdaMART) to focus on those topics in order to learn effective yet risk-averse ranking systems.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval; G3.3 [Probability and Statistics]: Experimental design

Keywords: Risk-Sensitive Evaluation, Student's t Test

1. INTRODUCTION

Various paradigms for the evaluation of information retrieval (IR) systems rely on many topics to produce reliable estimates of their effectiveness. For instance, in the TREC series of evaluation forums, 50 topics is generally seen as the minimum for producing a reliable test collection [2, 25]. However, in more recent times, the evaluation of systems has increasingly focused upon their robustness - ensuring that a given IR system performs well on difficult topics (as investigated by the TREC Robust track [24]), or at least as well as a baseline system (which is known as risk-sensitive evaluation [26]). Recently, the TREC 2013 Web track introduced a risk-sensitive task, which assessed how systems could perform effectively yet without exhibiting large losses compared to a pre-determined baseline system [10].

In such a risk-sensitive evaluation, the *risk* associated with an IR system is defined as the risk of performing a given particular topic less effectively than a given baseline system [8, 9, 26]. In particular, the U_{Risk} risk-sensitive evaluation measure [26] calculates the absolute difference of an effectiveness measure (e.g. NDCG) between a given retrieval system and the baseline system, in a manner that more strongly emphasises decreases with respect to the baseline (known as risk) than gains (reward). A parameter $\alpha \geq 0$ controls the risk-reward tradeoff towards losses in effectiveness compared to the baseline, where $\alpha = 0$ weights risk and rewards equally.

In this paper, we argue that in the current practice of risk-sensitive evaluation based on U_{Risk}, any amount of loss in an IR system's average effectiveness, observed on a particular set of topics, is considered enough in magnitude to infer that the system exhibits a "*real* risk". However, from a statistical viewpoint, such an inferential decision may be said to be valid only if the observed amount of loss cannot be attributed to chance fluctuation. Otherwise, it will be equally likely that the corresponding system may or may not be under a real risk, meaning that it is possible that the system can perform every topic with a score higher than that of the baseline system on another set of topics that could be drawn from the population of topics. On the other hand, it is also possible that the observed amount of loss in a particular system's average effectiveness can be attributed to a chance fluctuation, while the corresponding performance losses for some individual topics are statistically significant in magnitude. In other words, significant performance losses for a few topics may not result in a significant total loss on average, given a relatively large set of topics.

Hence, we advocate that risk-sensitive evaluation can actually provide the necessary basis for (i) testing the significance of the observed amount of loss in a given IR system's average effectiveness, called *inferential risk analysis* in this paper, and (ii) testing the significance of the corresponding losses for individual topics, called *exploratory risk analysis*.

Indeed, we show that the U_{Risk} risk-reward tradeoff measure is actually a linear transformation of the t statistic, as used in the Student's t test. Therefore, using this statistical interpretation of U_{Risk} based upon hypothesis testing, this paper proposes a new risk-reward tradeoff measure, T_{Risk}, which is a linear transformation of the existing U_{Risk} measure, yet is theoretically grounded upon the Student's t test

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609625 .

for testing the significance of the observed amount of loss in a given IR system's average effectiveness. For $\alpha = 0$, T_{Risk} is equivalent to the standard t statistic used typically in the Student's t test for testing the null hypothesis of equality in the population mean effectiveness for two IR systems. However, for $\alpha > 0$, the U_{Risk} measure emphasises performance losses compared to the baseline effectiveness. This raises challenges in the estimation of the standard error of the calculated U_{Risk} scores. For this reason, we propose the use of the Jackknife technique (or leave-one-out) [11], which is a re-sampling technique for estimating the bias and the standard error of any estimate. The Jackknife technique serves two purposes: firstly, to allow the empirical verification of the estimation of the standard error of U_{Risk} as valid; and secondly, for testing the significance of the corresponding performance losses for individual topics.

From a practical perspective, a risk-sensitive evaluation serves two objectives: firstly, as a step further than the classical evaluation of IR systems, which takes into account the stability or variance of retrieval results across queries as well as for the average retrieval effectiveness [8, 9]; and secondly, as a technique for jointly optimising the retrieval effectiveness and robustness of retrieval frameworks such as learning to rank [26]. Indeed, compared to the existing U_{Risk} measure, this paper contributes to both objectives, by exploiting the theory of statistical hypothesis testing for allowing meaningful interpretation of risk-sensitive evaluation scores, and also by allowing a learning to rank technique, namely LambdaMART, to focus on those topics that lead to a significant level of risk, in order to learn effective yet risk-averse ranking systems. The remainder of this paper is structured as follows: Section 2 provides an overview of risk-sensitive evaluation practices, including U_{Risk}; Section 3 relates the U_{Risk} measure to the t statistic, and hence proposes the new T_{Risk} risk-sensitive evaluation measure, and discusses the estimation of the standard error. Section 4 and Section 5 describe new forms of analysis, inferential and exploratory respectively, that arise from the T_{Risk} measure, and demonstrate their application upon the TREC 2012 Web track. Next, Section 6 shows how T_{Risk} can improve the robustness of the LambdaMART state-of-the-art learning to rank technique. Finally, we review some related work and provide concluding remarks in Sections 7 & 8, respectively.

2. RISK-SENSITIVE EVALUATION

Different approaches in IR such as query expansion [1, 5] and learning to rank [17] 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 [23]), or more recently, on the risk-sensitive evaluation of systems across many topics [8, 9, 26].

Originally, the aim of risk-sensitive evaluation [9] was to provide new analysis techniques for quantifying and visualising the risk-reward tradeoff of any retrieval strategy that requires a balance between risk and reward. Hence, it facilitates the quest for ranking strategies that are more *robust* in retrieval effectiveness compared to a baseline retrieval strategy – robust in the sense of the stability or variance of the retrieval results across topics, while achieving good average performance over all topics.

The variance with respect to a given baseline system b over a given set of topics Q with c topics can then be measured as a risk function F_{Risk}, which takes into account the downside-risk of a new system r (i.e. performing a topic worse than the baseline) is defined in [26] as follows:

$$\mathrm{F}_{Risk} = \frac{1}{c}\sum_{i=1}^{c} \max\left[0, (b_i - r_i)\right], \tag{1}$$

where r_i and b_i are respectively the score of the system r and the score of the baseline system b on topic i, as measured by a retrieval effectiveness measure (e.g. NDCG@20, ERR@20 [6]). Similarly, a reward function F_{Reward}, which takes into account the upside-risk (i.e. performing a topic better than the baseline) is defined as:

$$\mathrm{F}_{Reward} = \frac{1}{c}\sum_{i=1}^{c} \max\left[0, (r_i - b_i)\right]. \tag{2}$$

Thereby, the overall *gain* in the retrieval effectiveness of r with respect to b can be expressed as:

$$\mathrm{U}_{Gain} = \mathrm{F}_{Reward} - \mathrm{F}_{Risk}. \tag{3}$$

Next, a single measure, U_{Risk} [26], which allows the risk-reward tradeoff to be adjusted, was defined:

$$\begin{aligned} \mathrm{U}_{Risk} &= \mathrm{U}_{Gain} - \alpha \cdot \mathrm{F}_{Risk} \\ &= \frac{1}{c}\left[\sum_{q\in Q_+} \delta_q + (1+\alpha)\sum_{q\in Q_-} \delta_q\right], \end{aligned} \tag{4}$$

where $\delta_q = r_q - b_q$. The left summand in the square brackets, which is the sum of the score differences δ_q for all q where $r_q > b_q$ (i.e., $q \in Q_+$), gives the total win (or upside-risk) with respect to the baseline. Orthogonally, the right summand, which is the sum of the score differences δ_q for all q where $r_q < b_q$, 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): $\alpha = 0$ results in a pure gain model, while for higher α, the penalty for under-performing with respect to the baseline is increased: typically $\alpha = 1, 5, 10$ [10].

In this paper, we extend the original aforementioned aim of risk-sensitive evaluation with the following contributions:
1. A well-established statistical hypothesis testing theory for risk-sensitive evaluations from which arises a new risk measure T_{Risk} (Section 3), to turn risk-sensitive evaluation from a descriptive analysis to a fully-fledged *inferential* analysis (Section 4).
2. A method for *exploratory* risk analysis that can identify the topics that commit real levels of risk (Section 5).
3. Adaptations of the proposed T_{Risk} measure that can enhance the robustness of the state-of-the-art LambdaMART learning to rank technique, compared to U_{Risk}, without degradations in overall effectiveness, where the learned model adaptively adjusts with respect to the risk level committed by individual topics (Section 6).

3. THE NEW $\mathrm{T}_{\mathrm{RISK}}$ MEASURE

Without loss of generality, at $\alpha = 0$, the risk-reward tradeoff measure U_{Risk} reduces to the U_{Gain} formula in Eq. (3), which can be expressed as the *average gain* over c topics:

$$\mathrm{U}_{Gain} = \frac{1}{c}\sum_{i=1}^{c} \delta_i = \frac{1}{c}\sum_{i=1}^{c} (r_i - b_i). \tag{5}$$

In the context of statistics, U_{Gain} refers to the sample mean of paired score differences, $\bar{d}$, for two IR systems (the system under evaluation r and the baseline system b):

$$\bar{d} = \bar{r} - \bar{b} = \frac{1}{c}\sum_{i=1}^{c}(r_i - b_i) = \mathrm{U}_{Gain} \quad (6)$$

and in the context of evaluating IR systems, this refers to the difference in average effectiveness between two IR systems, $\bar{r}-\bar{b}$, where $\bar{r}$ and $\bar{b}$ are respectively the average effectiveness of system r and the average effectiveness of the baseline system b over c topics.

On the other hand, the Student's t statistic for matched pairs, as is commonly applied when testing the significance of results between two systems, can be expressed as:

$$t = \frac{\bar{d}}{\mathrm{SE}(\bar{d})} \quad \left[= \frac{\bar{r}-\bar{b}}{\mathrm{SE}(\bar{d})}\right], \quad (7)$$

Within Eq. (7), the standard error of paired sample mean, $SE(\bar{d})$, can be estimated as follows:

$$\mathrm{SE}(\bar{d}) = \frac{s_d}{\sqrt{c}}, \quad (8)$$

where $s_d = \sqrt{c^{-1}\sum(\delta_i - \bar{d})^2}$ is the paired sample standard deviation. Hence, we argue that the Student's t statistic of Eq. (7) is actually a linear transformation of U_{Gain} from Eq. (3), which we call T_{Gain}:

$$\mathrm{T}_{Gain} = \frac{\mathrm{U}_{Gain}}{SE(\mathrm{U}_{Gain})} = \frac{\sqrt{c}}{s_d} \times \mathrm{U}_{Gain}. \quad (9)$$

This transformation can be referred to as *studentisation* (c.f., t-scores) [14], which in fact is a type of *standardisation* (i.e., z-scores). Standardisation is a monotonic linear transformation, which transforms any given set of data to a set with zero mean and unit variance, while preserving the original data distribution in shape.

The t-score of a raw U_{Gain} measurement, T_{Gain}, differs from the raw measurement in two important aspects. First, given a set of IR systems, a test collection, and a baseline system, the systems' ranking to be obtained on the basis of T_{Gain} will not necessarily be concordant with the systems' ranking to be obtained on the basis of U_{Gain}, since the t statistic takes into account the inherent variation in the observed paired score differences $r_i - b_i$ across the topics, i.e., $SE(\mathrm{U}_{Gain})$. Second, given a particular baseline system, the two T_{Gain} scores to be obtained on two different test collections for the same IR system are comparable with each other in magnitude, at least in theory [7], while the two U_{Gain} scores are not, as typical in the case of the two raw effectiveness scores to be yielded from a standard effectiveness measure, such as mean average precision [28].

Having shown how T_{Gain} can be defined as a linear transformation of U_{Gain}, based upon the t statistic, we now examine U_{Risk}, which allows the risk-reward tradeoff to be controlled by the α parameter. For $\alpha \geq 0$, the t statistic based on U_{Risk}, which we call T_{Risk}, can be expressed as follows:

$$\mathrm{T}_{Risk} = \frac{\mathrm{U}_{Risk}}{SE(\mathrm{U}_{Risk})}. \quad (10)$$

Although both the T_{Gain} formula in Eq. (9) and the T_{Risk} formula in Eq. (10) stem from the classical t statistic in Eq. (7), the estimation of the standard error in U_{Risk}, the estimation of $SE(\mathrm{U}_{Risk})$ within T_{Risk}, is not as straightforward as in the case of $SE(\mathrm{U}_{Gain})$, for the reason that the U_{Risk} formula reweighs the score differences δ_i in averaging, proportionally to α, for each topic i where $r_i < b_i$, as opposed to U_{Gain}. Hence, in the remainder of this section, we propose two methods to estimate $SE(\mathrm{U}_{Risk})$: A speculative parametric estimator $SE_{\bar{x}}$ that is an analogy to the paired sample standard deviation s_d (Section 3.1); and a nonparametric Jackknife Estimator SE_J, based on the leave-one-out Jackknife technique (Section 3.2). Indeed, later in Section 3.3, we use the Jackknife Estimator SE_J to show the validity of the speculative $SE_{\bar{x}}$ estimator.

On the other hand, T_{Risk} has several advantages over U_{Risk}. Firstly, it can be easily interpreted for an inferential analysis of risk. Indeed, we will later show in Section 4 that in order to test the significance of an observed risk-reward tradeoff score between a particular IR system and a provided baseline system, one can use T_{Risk} as the test statistic of the Student's t test for matched pairs.

Secondly, T_{Risk} permits the identification of topics that commit significant risk or not – we call this *exploratory* risk analysis – which we present later in Section 5.

Finally, this exploratory risk analysis leads to new risk-sensitive measures that can be directly integrated into the LambdaMART learning to rank technique, to produce learned models that exhibit less risk than those obtained from U_{Risk} whilst not degrading effectiveness, as explained in Section 6.

3.1 Parametric Estimator of $SE(\mathrm{U}_{Risk})$

Let the random variable X_i denote the risk-reward tradeoff score between system r and baseline b for topic i:

$$X_i = \begin{cases} \delta_i & \text{if } r_i > b_i \\ (1+\alpha)\delta_i & \text{if } r_i < b_i \end{cases} \quad (11)$$

for $i = 1, 2, \ldots, c$ and a predefined value of $\alpha \geq 0$. Then, the standard error of U_{Risk}, $SE(\mathrm{U}_{Risk})$ can be approximated by the standard error of the sample mean $\bar{x}$:

$$SE_{\bar{x}} = \frac{s_x}{\sqrt{c}}, \quad (12)$$

where $s_x^2 = c^{-1}\sum(x_i - \bar{x})^2$. Here, the sample mean $\bar{x}$ corresponds to the U_{Risk} score considered as the arithmetic mean of the sample of the *observed* individual topic risk-reward tradeoff scores $x_1, x_2, \ldots, x_c$ at a predefined value of α:

$$\bar{x} = \mathrm{U}_{Risk} = \frac{1}{c}\sum_{i=1}^{c} x_i. \quad (13)$$

This parametric estimator of $SE(\mathrm{U}_{Risk})$, $SE_{\bar{x}}$, is speculative and hence its validity might be compromised to some extent. Therefore, we empirically verify the validity of $SE_{\bar{x}}$ in estimating $SE(\mathrm{U}_{Risk})$ by means of comparing it with a nonparametric re-sampling technique, called the Jackknife [21], which we present in Section 3.2. Indeed, by comparing the two estimates of $SE(\mathrm{U}_{Risk})$ (i.e., the parametric estimate $SE_{\bar{x}}$ of Eq. (12) and the nonparametric Jackknife estimate of $SE(\mathrm{U}_{Risk})$), one can decide whether an inference to be made on the basis of the T_{Risk} statistic is valid. If the two estimates agree with each other, such an inference may be said to be valid, otherwise its validity is compromised.

3.2 Jackknife Estimate of $SE(\mathrm{U}_{Risk})$

In this paper, the Jackknife technique is employed for a purpose which serves two different aims: 1) as a mechanism of the empirical verification of the validity of an inference to be made based on the T_{Risk} statistic in Eq. (10), and 2) as a mechanism for exploratory risk analysis.

Jackknife, which is also known as the Quenouille-Tukey Jackknife or leave-one-out, was first introduced by Quenouille [18] and then developed by Tukey [21]. Tukey used the Jackknife technique to determine how an estimate is affected by the subsets of observations when discordant values

(i.e., outlier data) are present. In the presence of discordant values, it is expected that the Jackknife technique could reduce the bias in the estimate. Although the original objective of Jackknife is to detect outliers, in principle it is a re-sampling technique for estimating the bias and the standard error of any estimate [11]. In Jackknife, the same test is repeated by leaving one subject out each time: this explains why this technique is also referred to as leave-one-out.

Let the random variables $X_1, X_2, \ldots, X_c$ denote a random sample of size c, such that X_i is drawn identically and independently from a distribution F for $i = 1, 2, \ldots, c$. Suppose that the goal is to estimate an unknown parameter θ of F. It can be shown that θ can be estimated by a statistic $\hat{\theta}$, which is derived from an observed sample $x_1, x_2 \ldots, x_c$ from F, with a measurable amount of sampling error [15].

An *unbiased* estimator $\hat{\theta}$ is a statistic whose expected value $E(\hat{\theta})$ is equal to the true value of the population parameter of interest θ, i.e., $E(\hat{\theta}) = \theta$. The amount of *bias* associated with an estimator is therefore given by:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta. \quad (14)$$

We denote as $X_{(i)}$ the sub-sample without the datum X_i. There are in total c sub-samples of size $c-1$ for $i = 1, 2, \ldots, c$: $X_{(i)} = X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_c$.

Next, let the estimate derived from the i^{th} sub-sample $X_{(i)}$ be denoted as $\hat{\theta}_{(i)}$, and the mean over c sub-samples be:

$$\hat{\theta}_{(.)} = \frac{1}{c}\sum_{i=1}^{c}\hat{\theta}_{(i)}. \quad (15)$$

The Jackknife estimate of *bias*, which is actually a nonparametric estimate of $E(\hat{\theta} - \theta)$, is defined as follows [21]:

$$\text{bias}_J(\hat{\theta}) = (c-1)(\hat{\theta}_{(.)} - \hat{\theta}) = \frac{(c-1)}{c}\sum_{i=1}^{c}(\hat{\theta}_{(i)} - \hat{\theta}).$$

and, in accordance, the *bias-reduced* Jackknife estimate of θ is defined as $\tilde{\theta} = \hat{\theta} - \text{bias}_J(\hat{\theta}) = c\hat{\theta} - (c-1)\hat{\theta}_{(.)}$.

Tukey [21] showed that the Jackknife technique can also be used to estimate the *variance* of $\hat{\theta}$ by introducing the so-called *pseudo-values*, $\tilde{\theta}_{(i)} = c\hat{\theta} - (c-1)\hat{\theta}_{(i)}$, such that

$$\text{var}_J(\hat{\theta}) = \frac{1}{c(c-1)}\sum_{i=1}^{c}\left[\tilde{\theta}_{(i)} - \tilde{\theta}\right]^2 = \frac{(c-1)}{c}\sum_{i=1}^{c}\left[\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right]^2.$$

This nonparametric Jackknife estimate of variance gives the empirical estimate of the standard error of $\hat{\theta}$:

$$SE(\hat{\theta}) = \sqrt{\text{var}_J(\hat{\theta})}. \quad (16)$$

For the T_{Risk} statistic in Eq. (10), the standard error of U_{Risk}, $SE(\text{U}_{Risk})$, can hence be estimated by substituting U_{Risk} into Eq. (16) as $\hat{\theta}$:

$$SE_J = \sqrt{\text{var}_J(\text{U}_{Risk})}. \quad (17)$$

3.3 Empirical Validation of $SE(\text{U}_{Risk})$

The nonparametric estimator SE_J is an alternative to the parametric estimator $SE_{\bar{x}}$ (Eq. (12)). In this section, we empirically compare these estimates of $SE(\text{U}_{Risk})$ with each other, to assess the validity of the result of a hypothesis test to be performed using T_{Risk} as the test statistic. In general, if the two estimates agree, the test result may be said to be valid, and otherwise its validity will be compromised. As a result, nonparametric methods can help to alleviate doubts about the validity of the analysis performed [14].

In the following, we compare the estimates using the submitted runs to the TREC Web track. In particular, the provided baseline run for the TREC 2013 Web track risk-sensitive task is based on the *Indri* retrieval platform. However, as the submitted runs and results for the TREC 2013 campaign were not yet publicly available 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[1] that correspond to the TREC 2013 baseline runs - in our results, we use the 2012 equivalent run to the 2013 pre-determined baseline, the so-called *indriCASP*. We report the U_{Risk} values obtained using the official TREC 2012 evaluation measure, ERR@20.

Table 1 reports the parametric estimates ($SE_{\bar{x}}$) and the nonparametric Jackknife estimates (SE_J) of the standard errors associated with the average risk-reward tradeoff scores (U_{Risk}), calculated for each of the TREC 2012 Web track top 8 ad-hoc runs over $c = 50$ topics, with respect to the *indriCASP* baseline, applying several risk-sensitivity parameter values of $\alpha = 0, 1, 5, 10$. From the results, it can be observed that the two estimates, $SE_{\bar{x}}$ and SE_J agree with each other for each of the 8 runs. In fact, over all of the 48 runs submitted to the TREC 2012 Web track, we observe a Root Mean Square Error (RMSE) of 0.000 between $SE_{\bar{x}}$ and SE_J. Thus, we conclude that it is highly likely that it would be valid to conduct an inferential risk analysis upon those TREC 2012 runs based on the new risk-reward tradeoff measure T_{Risk} (Eq. (10)), regardless of how $SE(\text{U}_{Risk})$ is estimated. An example of inferential risk analysis based on T_{Risk} follows in the next section.

4. INFERENTIAL RISK ANALYSIS

The goal of the classical evaluation of IR systems is to decide whether one IR system is better in retrieval effectiveness than another on the population of topics. This goal can be formulated into a (two-sided) null hypothesis, as given by:

$$H_0 : \mu_r = \mu_b \quad \text{or} \quad H_0 : \mu_r - \mu_b = 0, \quad (18)$$

against the alternative hypothesis $H_1 : \mu_r \neq \mu_b$, where μ_r and μ_b represent respectively the population mean performance of the system r and the population mean performance of the baseline system b. The test statistic for this null hypothesis is the t statistic (Eq. (7)), since the larger values of t are evidence against the null hypothesis $H_0 : \mu_r - \mu_b = 0$. Below, we describe the hypothesis testing of H_0 in abstract terms, before explaining how it can be applied to T_{Risk} (Section 4.1) and illustrating its application upon the TREC 2012 Web track runs (Section 4.2).

In order to decide how much difference between the two sample means $\bar{r}$ and $\bar{b}$ is assumed to be large enough to reject the null hypothesis, we should first determine how much difference can be attributed to a *chance fluctuation*. It can be shown that, under the null hypothesis H_0, the sampling (or null) distribution of the test statistic t can be approximated by a Student's t distribution with $df = c - 1$ degrees of freedom for any population distribution with finite mean μ and variance $\sigma^2 > 0$, because of the central limit theorem [12]. Thus, at a predefined significance level of γ (typically $\gamma = 0.05$ for 95% confidence), two standard deviations ($\pm t_{(\gamma/2,df)} \times \text{SE}(\bar{d})$) determine the maximum difference that can be attributed to chance fluctuation, where in between the critical values $\pm t_{(\gamma/2,df)}$ the area under the Student's t

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

Table 1: Calculated risk-reward tradeoff scores, U_{Risk} for the TREC 2012 Web track top 8 ad-hoc runs at the risk-sensitivity parameter values of $\alpha = 0, 1, 5, 10$, along with the parametric estimates $SE_{\bar{x}}$ and the nonparametric Jackknife estimates SE_J of the associated standard errors $SE(\mathrm{U}_{Risk})$. *indriCASP* is the baseline.

		$\alpha = 0$			$\alpha = 1$			$\alpha = 5$			$\alpha = 10$		
	ERR@20	U_{Risk}	$SE_{\bar{x}}$	SE_J	U_{Risk}	$SE_{\bar{x}}$	SE_J	U_{Risk}	$SE_{\bar{x}}$	SE_J	U_{Risk}	$SE_{\bar{x}}$	SE_J
uogTrA44xi	0.3132	0.1185	0.0528	0.0528	0.0556	0.0739	0.0739	-0.1959	0.1755	0.1755	-0.5104	0.3091	0.3091
srchvrs12c09	0.3049	0.1102	0.0479	0.0479	0.0679	0.0644	0.0644	-0.1015	0.1489	0.1489	-0.3133	0.2619	0.2619
DFalah121A	0.2920	0.0974	0.0425	0.0425	0.0467	0.0632	0.0632	-0.1558	0.1588	0.1588	-0.4089	0.2827	0.2827
QUTparaBline	0.2901	0.0954	0.0448	0.0448	0.0385	0.0672	0.0672	-0.1893	0.1703	0.1703	-0.4740	0.3033	0.3033
utw2012fc1	0.2195	0.0248	0.0449	0.0449	-0.0558	0.0705	0.0705	-0.3782	0.1849	0.1849	-0.7813	0.3314	0.3314
ICTNET12ADR2	0.2149	0.0203	0.0416	0.0416	-0.0495	0.0637	0.0637	-0.3286	0.1648	0.1648	-0.6774	0.2950	0.2950
indriCASP	0.1947	*	*	*	*	*	*	*	*	*	*	*	*
irra12c	0.1723	-0.0223	0.0410	0.0410	-0.1182	0.0693	0.0693	-0.5014	0.1904	0.1904	-0.9805	0.3437	0.3437
qutwb	0.1659	-0.0287	0.0462	0.0462	-0.1342	0.0791	0.0791	-0.5560	0.2194	0.2194	-1.0832	0.3969	0.3969

distribution sums up to $(1-\gamma)$. If an observed t-score is greater than $t_{(\gamma/2,df)}$, or less than $-t_{(\gamma/2,df)}$, one can reject H_0 with $100\%(1-\gamma)$ confidence, denoted as the p-value.

4.1 Inference Based on T_{Gain} and T_{Risk}

The above protocol of hypothesis testing is referred to as the Student's t test for matched pairs, or paired t test for short, in statistics. Hence, in the context of risk-sensitive evaluation, the T_{Gain} formula in Eq. (9) stands for the test statistic t. In fact, at $\alpha = 0$, testing the significance of an observed risk-reward tradeoff score between r and b (i.e. an observed U_{Gain} score) is akin to testing the significance of the observed difference between $\bar{r}$ and $\bar{b}$.

To test the significance of an observed U_{Gain} score, one can therefore compare the corresponding T_{Gain} score with the two-sided critical $\pm t_{(\gamma/2,df)}$ values at a desired level of significance γ. If $-t_{(\gamma/2,df)} \leq \mathrm{T}_{Gain} \leq t_{(\gamma/2,c-1)}$, the observed U_{Gain} score can be attributed to chance fluctuation, meaning that the observed gain in the performance of the system r with respect to the baseline system b is not statistically significant. In such a case, it is equally likely that the observed U_{Gain} score may or may not occur on another topic sample drawn from the population. Otherwise, if $\mathrm{T}_{Gain} \leq -t_{(\gamma/2,c-1)}$ or $\mathrm{T}_{Gain} \geq t_{(\gamma/2,c-1)}$, one can however be sure that a U_{Gain} score at least as extreme as the observed score would occur on $100(1-\gamma)\%$ of the topic samples that could be drawn from the population.

Both T_{Gain} and T_{Risk} stem from the t statistic. Indeed, for $\alpha = 0$, $\mathrm{T}_{Gain} = \mathrm{T}_{Risk}$, while for $\alpha > 0$, $SE(\mathrm{U}_{Risk})$ was shown to be valid in Section 3.3. Hence, we argue that an equivalent inferential analysis can be conducted upon the T_{Risk} scores that have been calculated based on U_{Risk}. In the following, we provide an illustration of such inferential analysis upon runs submitted to the TREC 2012 Web track, but the same inferential analysis methodology could be applied for any risk-sensitive evaluation scenario.

4.2 Inferential Analysis of Web Track Runs

Given a particular IR system, a baseline system, and a set of c topics, one can use the paired t test for testing the significance of the calculated average tradeoff score between risk and reward over the c topics, U_{Risk}, by comparing the corresponding t-score, T_{Risk}, with the critical values $\pm t_{(\gamma/2,df)}$ at a desired level of significance γ. To illustrate such an analysis, Table 2 reports the U_{Risk} risk-reward tradeoff scores based on ERR@20, and the corresponding T_{Risk} scores for the 8 highest performing TREC 2012 ad-hoc runs, given the baseline run *indriCASP* (we omit other submitted runs for brevity, however the following analysis would be equally applicable to them). As the TREC 2012 Web track has 50 topics, for a significance level of $\gamma = 0.05$, the critical values for T_{Risk} are $\pm t_{(0.025,49)} = \pm 2$.

In Table 2, the U_{Risk} scores to which a two-sided paired t test gives significance are those that have a corresponding T_{Risk} score less than -2 or greater than $+2$. For example, at $\alpha = 0$, the calculated U_{Risk} scores of the top 4 runs are significant with a p-value less than 0.05. This means that, under the null hypothesis $H_0 : \mu_r = \mu_b$, given another sample of 50 topics from the population, the probability of observing a risk-reward tradeoff score, between any one of these 4 runs and the baseline run *indriCASP*, that is as extreme or more extreme than the one that was observed is less than 0.05, i.e. the associated p-values. Since $\mathrm{T}_{Risk} > 0$, for those runs, the declared significance counts in favour of "reward" against "risk". Thus, one can conclude, with 95% confidence, that the expected per topic effectiveness of each of the top 4 runs is, on average, higher than the expected per topic effectiveness of the baseline run *indriCASP* on the population of topics. In other words, given a topic from the population, it is highly likely that any one of the top 4 runs will not perform worse for that topic than *indriCASP*. This suggests, as a result, that those top runs do not exhibit a real risk that is generalisable to the population of topics.

On the other hand, a run with $\mathrm{T}_{Risk} < -2$ at $\alpha = 0$ will be under a real risk, though among the shown top 8 TREC 2012 runs there is no such run. For those runs with $-2 \leq \mathrm{T}_{Risk} < +2$, such as *utw2012fc1* and *qutwb*, the risk analysis performed here is inconclusive, since the associated U_{Risk} scores can be attributed to chance fluctuation, i.e. it is equally likely that they may or may not be under a real risk.

Next, we observe from Table 2 that as α increases, the observed tradeoffs between risk and reward for each run changes in favour of risk compared to reward, hence the runs exhibiting significant U_{Risk} scores change. For example, each of the runs with significant U_{Risk} scores at $\alpha = 0$ (i.e., the top 4 runs) have a U_{Risk} score that can be attributed to a chance fluctuation at $\alpha = 10$, while, in contrast, those runs whose U_{Risk} scores can be attributed to chance fluctuation at $\alpha = 0$ (i.e., the last 4 runs) have a significant U_{Risk} score at $\alpha = 10$.

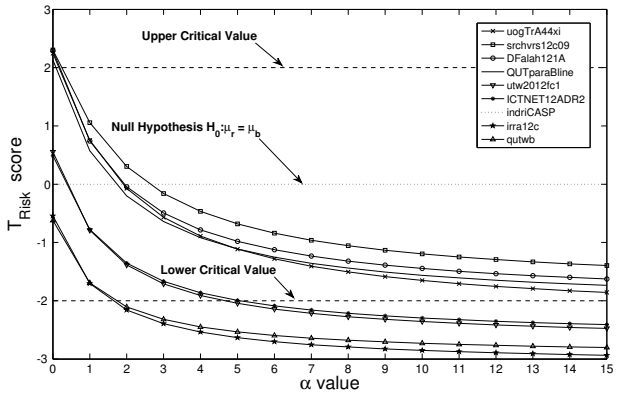
Figure 1 shows the change in the T_{Risk} scores of the TREC 2012 top 8 ad-hoc runs for several risk-sensitivity α parameter values from 0 to 15. From the figure, we observe that for $\alpha > 5$ the T_{Risk} scores for all runs are negative in sign, and for the last 4 runs the calculated U_{Risk} scores can be considered statistically significant (i.e., $\mathrm{T}_{Risk} > -2.0$ for $\alpha > 5$). It is also observed that, even for $\alpha = 15$, the calculated U_{Risk} scores of the top 4 TREC runs can still be attributed to chance fluctuation.

As a result, the inferential analysis performed so far suggests that, in general, none of the 8 top TREC 2012 ad-hoc

Table 2: U_{Risk} and T_{Risk} scores risk-reward tradeoff scores for the top 8 TREC 2012 ad-hoc runs at $\alpha = 0, 1, 5, 10$, where the baseline is *indriCASP*. The underlined U_{Risk} scores are those for which a two-tailed paired t test gives significance with $p < 0.05$ - i.e. exhibit a T_{Risk} score greater than $+2$ or less than -2.

	$\alpha = 0$			$\alpha = 1$			$\alpha = 5$			$\alpha = 10$		
	U_{Risk}	T_{Risk}	p-value	U_{Risk}	T_{Risk}	p-value	U_{Risk}	T_{Risk}	p-value	U_{Risk}	T_{Risk}	p-value
uogTrA44xi	0.1185	2.2440	0.029	0.0556	0.7528	0.455	-0.1959	-1.1163	0.270	-0.5104	-1.6512	0.105
srchvrs12c09	0.1102	2.3034	0.026	0.0679	1.0541	0.297	-0.1015	-0.6817	0.499	-0.3133	-1.1961	0.237
DFalah121A	0.0974	2.2899	0.026	0.0467	0.7401	0.463	-0.1558	-0.9808	0.332	-0.4089	-1.4466	0.154
QUTparaBline	0.0954	2.1305	0.038	0.0385	0.5723	0.570	-0.1893	-1.1116	0.272	-0.4740	-1.5626	0.125
utw2012fc1	0.0248	0.5526	0.583	-0.0558	-0.7914	0.432	-0.3782	-2.0457	0.046	-0.7813	-2.3574	0.022
ICTNET12ADR2	0.0203	0.4869	0.629	-0.0495	-0.7775	0.441	-0.3286	-1.9942	0.052	-0.6774	-2.2960	0.026
irra12c	-0.0223	-0.5446	0.588	-0.1182	-1.7038	0.095	-0.5014	-2.6335	0.011	-0.9805	-2.8525	0.006
qutwb	-0.0287	-0.6226	0.536	-0.1342	-1.6956	0.096	-0.5560	-2.5342	0.015	-1.0832	-2.7295	0.009

Figure 1: The change in standardised T_{Risk} scores for the top TREC 2012 ad-hoc runs for $0 \leq \alpha \leq 15$.



runs are under a real risk of performing any given topic from the population worse than the baseline run *indriCASP*, on average. In particular, there can be no significant reduction in risk that could be attained for the top 4 systems, given a baseline system with the average retrieval effectiveness of *indriCASP*. On the other hand, a significant reduction in risk could be attained, on average, for the last 4 systems, particularly for $\alpha > 5$.

Lastly, in Table 2, it is notable that the high U_{Risk} scores do not necessarily imply high T_{Risk} scores, because of the fact that each system would in general have a different inherent variation in $r_i - b_i$ across topics (i.e. $SE(\mathrm{U}_{Risk})$) from that of the other systems. For example, consider the runs *uogTrA44xi* and *srchvrs12c09*. At $\alpha = 0$, *uogTrA44xi* has a U_{Risk} score (0.1185) higher than the U_{Risk} score (0.1102) of *srchvrs12c09*, while *srchvrs12c09* has a higher T_{Risk} score than *uogTrA44xi*, i.e. 2.3034 vs. 2.2440. This shows that a ranking of retrieval systems obtained based on T_{Risk} will not necessarily be concordant with the ranking of systems obtained based on U_{Risk}.

5. EXPLORATORY RISK ANALYSIS

In the previous section, the risk analysis that we performed could hide significant performance losses on individual topics. Nevertheless, one can perform an *exploratory* risk analysis to determine those individual topics on which the observed risk-reward tradeoff score between a given IR system and the baseline system (i.e., x_i) is statistically significant. In the following, we provide a definition for exploratory risk analysis (Section 5.1), which we later illustrate upon the TREC 2012 Web track runs (Section 5.2).

5.1 Definition

The T_{Risk} measure permits the topic-by-topic analysis of risk-reward tradeoff measurements, which we refer to as exploratory risk analysis. Such an analysis is implicitly suggested by the t statistic itself. The t statistic in Eq. (7) can be rewritten as follows:

$$t = \frac{\bar{d}}{\mathrm{SE}(\bar{d})} = \frac{\frac{1}{c}\sum_{i=1}^{c}(r_i - b_i)}{s_d/\sqrt{c}} = \frac{\sqrt{c}}{c}\sum_{i=1}^{c}\frac{r_i - b_i}{s_d}. \quad (19)$$

In here, each component of the sum $t_i = \frac{r_i - b_i}{s_d}$: gives the standardised score of the observed difference in effectiveness between the system r and the baseline system b on topic i, for $i = 1, 2, \ldots, c$.

In analogy, the T_{Risk} measure, which stems from the t statistic, can be rewritten as:

$$\mathrm{T}_{Risk} = \frac{\mathrm{U}_{Risk}}{SE_{\bar{x}}} = \frac{\frac{1}{c}\sum_{i=1}^{c}x_i}{s_x/\sqrt{c}} = \frac{\sqrt{c}}{c}\sum_{i=1}^{c}\frac{x_i}{s_x}, \quad (20)$$

where each component of the sum, in this case, gives the standardised score of the individual topic risk-reward tradeoff measurements $x_1, x_2, \ldots, x_c$:

$$\mathrm{T}_{R_i} = \frac{x_i}{s_x}. \quad (21)$$

In a similar manner that we compare the calculated T_{Risk} score of a given IR system with the two-sided critical values $\pm t_{(\gamma/2,df)}$ to decide whether the system exhibits a significant level of risk on average (Section 4), to decide whether an observed loss (or gain) on a particular topic i is significant, we can compare the component T_{R_i} score with the same critical values $\pm t_{(\gamma/2,df)}$, at a desired significance level of γ. If $-t_{(\gamma/2,df)} \leq \mathrm{T}_{R_i} \leq t_{(\gamma/2,df)}$, the observed loss (or gain) can be attributed to chance fluctuation, and otherwise it can be considered statistically significant.

Indeed, this is one of the typical methods of outlier detection in statistics [14]. Recall that the original objective of Jackknife is to detect outliers [21]. The T_{Risk} measure can also be expressed in terms of the Jackknife estimate of bias, following Wu [29]:

$$\mathrm{T}_{Risk} = \frac{\mathrm{U}_{Risk}}{SE_J} = \frac{1}{c}\sum_{i=1}^{c}\frac{\sqrt{(c-1)}\ (\hat{\theta}_{(i)} - \hat{\theta})}{SE_J}. \quad (22)$$

Here, each component of the sum:

$$\mathrm{T}_{J_i} = \frac{\sqrt{(c-1)}\ (\hat{\theta}_{(i)} - \hat{\theta})}{SE_J} = \frac{\sqrt{(c-1)}\ (\bar{x}_{(i)} - \bar{x})}{\sqrt{\mathrm{var}_J(\bar{x})}}, \quad (23)$$

gives the standardised Jackknife estimate of bias in U_{Risk} due to leaving the topic risk-reward score x_i out of the sample $x_1, x_2, \ldots, x_c$, where $\bar{x} = \mathrm{U}_{Risk}$ and $\bar{x}_{(i)}$ is the U_{Risk} score to be obtained when the i^{th} topic is leaved out of the topic set in use, for $i = 1, 2, \ldots, c$.

In general, both the T_{R_i} statistic in Eq. (21) and the T_{J_i} statistic in Eq. (23) can be used for the purpose of exploratory risk analysis. However, there is a certain difference

between them in theory. Using T_{R_i}, we can decide whether an observed performance loss on topic i is significant, by comparing the topic risk-reward score x_i with the maximum score that can be attributed to chance fluctuation, but as if the single datum x_i is the whole sample. In contrast, using T_{J_i}, we can make the same decision by comparing the observed difference between two U_{Risk} scores, $\bar{x}_{(i)} - \bar{x}$, with the maximum difference that can be attributed to chance fluctuation. Since we showed in Section 3.3 that the two estimates of the standard error for each TREC run are in perfect agreement (i.e. $SE_{\bar{x}} \approx SE_J$), we argue that this theoretical difference has no practical consequences. Hence, in the following, we provide an illustration of exploratory risk analysis on the TREC 2012 Web track runs, based on T_{J_i} alone. However, initial experiments showed no differences between T_{R_i} and T_{J_i}.

5.2 Exploratory Analysis of Web Track Runs

Figure 2 shows the standardised Jackknife estimate of bias in the U_{Risk} scores calculated for two TREC runs, namely *uogTrA44xi* and *qutwb* at $\alpha = 0, 5, 10, 15$ for the 50 TREC 2012 Web track topics, where *indriCASP* is the baseline. This standardised Jackknife estimate of bias, T_{J_i} is estimated by leaving one TREC 2012 Web track topic out of the set of topics $\{151, 152, \ldots, 200\}$ in turn. In the figure, the topics that result in a significant performance loss (gain) for the corresponding systems with respect to *indriCASP*, at the significance level of $\gamma = 0.05$, are those which have a T_{J_i} score less than -2 (greater than 2, respectively). Horizontal lines at -2 and $+2$ are shown to aid clarity.

From Figure 2, at $\alpha = 0$ it can be observed that *uogTrA44xi* has more significant wins in number than *qutwb*, and less significant losses. This shows in detail why the declared significance for *uogTrA44xi* in Section 4 counts in favour of reward against risk, while the observed tradeoff between risk and reward can be attributed to chance fluctuation for *qutwb*, with respect to the baseline *indriCASP*.

In general, both of the runs *uogTrA44xi* and *qutwb* exhibit considerable performance losses with respect to *indriCASP* on the same topics, including 166, 172, 174, 175, and 191, out of which 2 are significant for *uogTrA44xi* (i.e., 166 and 175) and 4 are significant for *qutwb* (i.e., 166, 172, 175, and 191), at $\alpha = 0$. In particular, consider the topic 166, on which the magnitude of the T_{J_i} score is nearly the same for both runs. It is notable here that, as α increases, the significance of that topic relatively doubles for *uogTrA44xi*, while for *qutwb* it nearly remains the same. The situation is also similar for topic 175, though the T_{J_i} score of *uogTrA44xi* at $\alpha = 0$ is small in magnitude compared to that of *qutwb*.

This is one of the important differences between T_{Risk} and U_{Risk} in assessing the risk associated with IR systems. Given a particular topic i, the same amount of performance loss with respect to a provided baseline effectiveness can lead to different T_{J_i} (and $\mathrm{T}_{R_i} = x_i/SE_{\bar{x}}$) scores for different IR systems, depending on the variation in the observed risk-reward tradeoff across the topics (i.e., different $SE_{\bar{x}}$ for different systems), while leading to the same topic risk-reward score, x_i, for $i = 1, 2, \ldots, c$. As α increases, the topic risk-reward score x_i increases proportionally for both of the runs *uogTrA44xi* and *qutwb*. However, the tradeoff counts, on average, significantly in favour of reward against risk for *uogTrA44xi*, whereas, it counts neither in favour of reward nor against risk for *qutwb*, as shown in Section 4. Thus, the same margin of increase in topic risk-reward tradeoff score x_i in favour of risk should lead to a relatively higher level of risk for *uogTrA44xi* than that for *qutwb*, in a way that T_{J_i} did.

Assessing the level of risk that a topic commits for a given IR system relative to the level of risk associated with the system on average is a property unique to the measures T_{J_i} and T_{R_i}. Besides the use of these measures for exploratory risk analysis, this property also enables adaptive risk-sensitive optimisation within a learning to rank technique, as we explain in the next section.

6. ADAPTIVE RISK OPTIMISATION

In this section, we describe how to exploit the new risk-reward tradeoff measure T_{Risk} (Eq. (10)) in learning robust ranking models that maximises average retrieval effectiveness while minimising risk-reward ratio, in the context of the state-of-the-art LambdaMART learning to rank technique [30]. As discussed below, Wang et al. [26] proposed to integrate U_{Risk} (Eq.(4)) within LambdaMART to achieve *risk sensitive optimisation*, by using α to penalise risk during the learning process. However, U_{Risk} considers topics equally regardless of the level of risk they commit. In contrast, we propose to *adaptively* change the level of risk-sensitivity, so that the total risk-sensitivity is distributed across the topics proportionally to the level of risk each topic commits. In the following: Section 6.1 provides an overview of the LambdaMART objective function, while Section 6.3 describes the integration of U_{Risk} within LambdaMART; Section 6.3 explains our proposed adaptive risk-sensitive optimisation approaches, with the experimental setup & results following in Sections 6.4 & 6.5, respectively.

6.1 LambdaMART

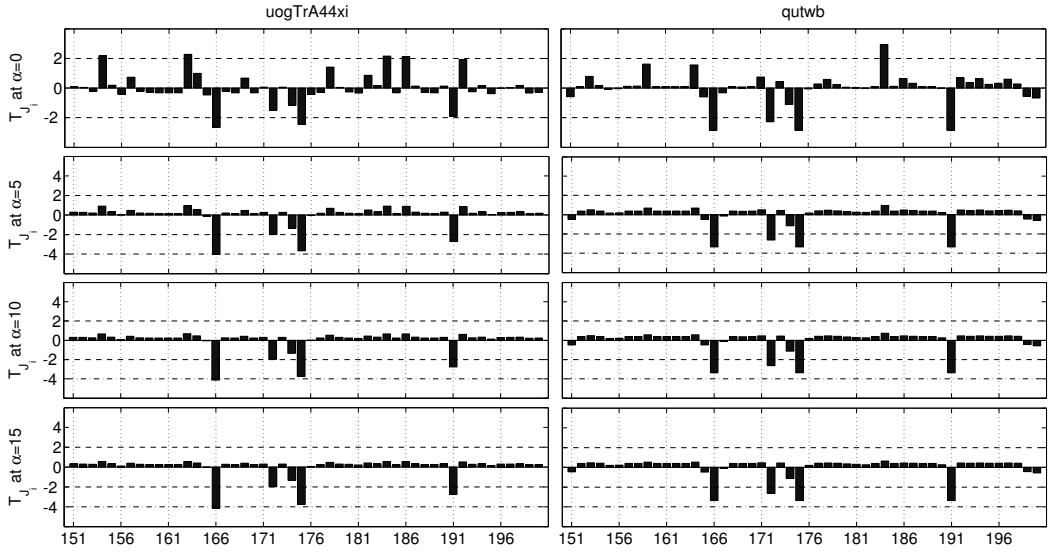
LambdaMART [30] is a state-of-the-art learning to rank technique, which won the 2011 Yahoo! learning to rank challenge. It can be described as a tree-based technique, in that its resulting learned model takes the form of an ensemble of regression trees, which is used to predict the score of each document given the document's feature values. During learning, LambdaMART creates a sequence of gradient boosted regression trees that improve an effectiveness metric. In general, for our purposes[2], it is sufficient to state that LambdaMART's objective function is based upon the product of two components: (i) the derivative of a cross-entropy that originates from the RankNet learning to rank technique [3] calculated between the scores of two documents a and b, and (ii) the absolute change ΔM in an evaluation measure M due to the swapping of documents a and b [4]. Therefore the final gradient λ_a^{new} of a document a within the objective function is obtained over all pairs of documents that a participates in for query q:

$$\lambda_a^{new} = \sum_{b \neq a} \lambda_{ab} \cdot |\Delta M_{ab}|$$

where λ_{ab} is RankNet's cross-entropy derivative, and ΔM_{ab} is the change in an evaluation measure M by swapping documents a and b. Various IR evaluation measures are suitable for use as M, including NDCG and MAP, as they have been shown to satisfy a *consistency property* [4]: for a pair of documents a and b where a is ranked higher than b, if the relevance label of a is higher than b, then a "degrading" swap of a and b must result in a decrease in M (i.e. $\Delta M \leq 0$), and orthogonally $\Delta M \geq 0$ for "improving" swaps.

[2]Further details on LambdaMART can be found in [4, 26].

Figure 2: Bar graph showing the standardised Jackknife estimate of bias in the U_{Risk}, T_{J_i}, for *uogTrA44xi* and *qutwb* at $\alpha = 0, 5, 10, 15$, where *indriCASP* is the baseline.



6.2 Risk-Sensitive Optimisation

Wang et al. [26] demonstrated that a more robust learned model could be obtained from LambdaMART if the ΔM is replaced by the difference in U_{Risk} for a given swap of two documents, denoted ΔT. In doing so, their implementation weights the value of ΔM by $\alpha + 1$ only for the topics with down-side risk, while for the topics with up-side risk it leaves ΔM as is, $\Delta T = \Delta M$. ΔT was shown to exhibit the consistency property iff the underlying evaluation measure ΔM is consistent (e.g. as obtained from NDCG).

6.3 Adaptive Risk-Sensitive Optimisation

Compared to U_{Risk}, T_{Risk} is grounded in the theory of hypothesis testing and produces values that are easily interpretable – as shown in Section 4. However, as a linear transformation of U_{Risk}, the direct application of T_{Risk} as ΔT within LambdaMART to attain risk-sensitive optimisation cannot offer marked improvements on the resulting learned models. On the other hand, the exploratory risk analysis of Section 5 offers a promising direction, as it permits the learning to rank process to *adaptively* focus on topics depending upon the level of risk that they commit. In this section, we propose two new models of adaptive risk-sensitive optimisation that exploit the standardised topic risk-reward tradeoff scores (T_{R_i}, Eq. (21)), but which differ on which individual topics they operate on. In particular, the first model, Semi-Adaptive Risk-sensitive Optimisation (SARO), focuses only on the topics with down-side risk and augments only the corresponding ΔM values. In contrast, the Fully Adaptive Risk-sensitive Optimisation (FARO) model operates on all topics and augments every ΔM value. Hence, compared to U_{Risk} as used in [26], FARO and SARO both alter the importance of riskier topics within the learning process.

In U_{Risk}, ΔM is multiplied by $\alpha + 1$ if the topic commits a downside risk[3]. This amounts to a static level of sensitivity for each topic, irrespective of the level of risk that the topic commits. In contrast, based on the standardised topic risk-reward tradeoff scores, T_{R_i} (Eq. (21)), we propose to adaptively adjust α so that the total level of sensitivity can be distributed across the topics proportional to the levels of risk that they commit. In order to achieve this, for each topic we must estimate the probability of observing a risk-reward score greater than the actual observed T_{R_i} score. Technically speaking, we need to estimate the cumulative probability $Pr\left(Z \geq \mathrm{T}_{R_i}\right)$, where T_{R_i} is the observed risk-reward tradeoff score and Z is the corresponding standard normal variable of T_{R_i} for all topics $i = 1, 2, .., c$. For large sample sizes (generally agreed to be ≥ 30), the distribution of the t statistic in Eq. (7) can be approximated by the standard normal *probability* distribution function, with zero mean and unit variance [15]. Thus, the probability $Pr\left(Z \geq \mathrm{T}_{R_i}\right)$, which is the probability of a topic risk-reward score greater than T_{R_i}, can be estimated by the standard normal *cumulative* distribution function $\Phi(\cdot)$, as follows:

$$Pr\left(Z \geq \mathrm{T}_{R_i}\right) \approx 1 - \Phi\left(\mathrm{T}_{R_i}\right), \tag{24}$$

for $i = 1, 2, \ldots, c$. $\Phi(Z)$ is a monotonically increasing function of the standard normal random variable Z, where $0 \leq \Phi(Z = z) \leq 1$ for $-\infty \leq z \leq \infty$, and at $Z = 0$, $\Phi(Z) = 0.5$.

Hence, we can replace the original α in ΔT as α' as follows:

$$\alpha' = [1 - \Phi(\mathrm{T}_{R_i})] \cdot \alpha. \tag{25}$$

where $0 \leq \alpha' \leq \alpha$. As the level of risk T_{R_i} committed by topic i increases, α' also increases. By substituting α' into ΔT (as defined by Wang et al. [26]), this augments the ΔM values for every topic with a weight proportional to the level of risk that each topic commits.

The application of α' differs between the SARO and FARO models. In particular, SARO only addresses the down-side risk, as in the case of U_{Risk}. Indeed, under the null hypothesis $H_0 : \mu_r = \mu_b$, the higher the level of down-side risk (i.e. the larger the size of the difference $r_i - b_i < 0$), the higher the probability of observing a topic risk-reward tradeoff score greater than the observed score ($Pr\left(Z \geq \mathrm{T}_{R_i}\right)$). Hence, SARO varies α' from 0 to α, according to the down-side risk of each topic.

On the other hand, FARO operates on all topics. Indeed, for the topics with up-side risk, FARO gives lower weights

[3]This follows directly from the definition of Eq. (4), however the consistency proof in Section 4.3.2 of [26] defines ΔT for different scenarios.

to the topics that more strongly outperform the baseline system (i.e. as the difference $r_i - b_i > 0$ increases). At the extreme, if topic i exhibits maximal improvements over the baseline (i.e. $r_i - b_i = 1$), then $\Phi(\mathrm{T}_{R_i}) = 1$, and hence topic i has minimal emphasis on the learner. In other words, the learner focuses on improving the riskier topics. FARO operates on all topics, by redefining ΔT as follows:

$$\Delta T' = (1 + \alpha') \times \Delta M, \tag{26}$$

Moreover, for $\alpha = 0$, $\alpha' = 0$, hence $\Delta T' = \Delta M$, i.e. the gain-only LambdaMART, as for U_{Risk}.

Finally, we informally comment on the consistency of SARO and FARO: For both models, we calculate $SE(U_{Risk})$ after the first iteration of boosting within LambdaMART, and not for each considered swap – we found this to be sufficient to obtain accurate estimates of $SE(U_{Risk})$; Next, the consistency of SARO follows from U_{Risk}, as our replacement of α with α', as $0 \leq \alpha' \leq \alpha$. For FARO, $\Delta T'$ only changes sign with ΔM, again as $0 \leq \alpha' \leq \alpha$. Hence, as long as ΔM is consistent, both SARO and FARO are also consistent.

6.4 Experimental Setup

We implement the U_{Risk}, SARO and FARO models within the Jforests implementation [13] of LambdaMART[4]. Experiments are conducted using the large MSLR-Web10k learning to rank dataset[5], as used by Wang et al. [26]. This dataset encompasses 9,685 queries with labelled documents obtained from a commercial web search engine. For each ranked document for each query, a range of 136 typical query-independent, query-dependent and query features are provided.

We use identical hyper-parameters for LambdaMART to those described by Wang et al. [26], namely: the minimum number of documents in each leaf $m = 500, 1000$, the number of leaves $l = 50$, the number of trees in the ensemble $nt = 800$ and the learning rate $r = 0.075$. The best m value is chosen for each of the five folds using the validation topic set, based on the NDCG@10 performance of the original LambdaMART algorithm, and used for all experiments for that fold thereafter. For the calculation of risk measures, like [26], we use the ranking obtained from the `BM25.whole.document` feature as the baseline system. The NDCG@10 performance of this baseline is 0.309.

The performances obtained for LambdaMART upon the MSLR-Web10k in terms of NDCG@1 and NDCG@10 are similar in magnitude to those reported by Wang et al. [26], however we note some differences in the risk profile. Such differences are expected given the different implementations: Wang et al. [26] used a private implementation of LambdaMART, while we use and adapt an open source machine learning toolkit for U_{Risk}, SARO and FARO. Nevertheless, the reported results allow valid conclusions to be drawn, including identical conclusions to [26] on the impact of using U_{Risk} within LambdaMART.

6.5 Results for SARO and FARO

Table 3 reports the effectiveness and robustness results for FARO and SARO along with U_{Risk}, for $\alpha = 1, 5, 10, 20$[6]. In the table, the gain over the baseline effectiveness is expressed as the risk (Eq. (1)) to reward (Eq. (2)) ratio (i.e., the "Risk/Reward" rows). Similarly, the number of topics that the risk-sensitive optimisation contributed to reward against risk is expressed as the loss to win ratio (i.e., the "Loss/Win" rows). Raw numbers of losses and wins associated with each α value for each model are also shown. Finally the "Loss > 20%" rows show, for each model, the number of topics on which the relative loss in performance over the BM25 baseline was higher than 20%[7].

Table 3: Results for SARO, FARO and U_{Risk}.

	$\alpha = 0$	$\alpha = 1$	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$
NDCG@1 (U_{Risk})	0.472	0.468	0.458	0.442	0.423
NDCG@1 (SARO)	-	**0.470**	0.463	0.455	0.439
NDCG@1 (FARO)	-	0.468	**0.467**	**0.470**	**0.469**
NDCG@10 (U_{Risk})	0.480	0.478	0.470	0.458	0.448
NDCG@10 (SARO)	-	**0.479**	0.474	0.468	0.458
NDCG@10 (FARO)	-	**0.479**	**0.477**	**0.479**	**0.478**
Risk/Reward (U_{Risk})	0.172	0.168	**0.164**	0.176	0.185
Risk/Reward (SARO)	-	**0.167**	**0.164**	**0.169**	0.177
Risk/Reward (FARO)	-	0.170	0.171	0.171	**0.172**
Loss/Win (U_{Risk})	0.281	0.278	0.267	0.272	0.275
Loss/Win (SARO)	-	**0.267**	**0.266**	0.272	**0.270**
Loss/Win (FARO)	-	0.272	0.274	**0.271**	0.277
Loss (U_{Risk})	2080	2059	**1992**	**2019**	2040
Loss (SARO)	-	**1996**	**1992**	2024	**2010**
Loss (FARO)	-	2025	2040	2040	2060
Win (U_{Risk})	7400	7417	7468	7427	7406
Win (SARO)	-	**7470**	**7476**	7437	**7441**
Win (FARO)	-	7451	7452	**7469**	7429
Loss > 20% (U_{Risk})	1180	1130	**1036**	**1036**	1042
Loss > 20% (SARO)	-	**1124**	1124	1046	**1032**
Loss > 20% (FARO)	-	1152	1145	1155	1172

As expected, since the semi-adaptive risk-sensitive optimisation (SARO) and the risk-sensitive optimisation based on U_{Risk} focus on only those topics with down-side risk, there is a steady decrease in average retrieval effectiveness (i.e., NDCG@1 and NDCG@10), as the risk-sensitivity parameter value of α increases. Nevertheless, SARO results in a decrease in average retrieval effectiveness that is less than U_{Risk}, for all α values. In contrast, the fully adaptive risk-sensitive optimisation (FARO) maintains the average retrieval effectiveness nearly constant across all α values, as well as the values of the quality and robustness measures, namely the risk-reward ratio and the loss-win ratio.

For SARO, the observed values of the two quality and robustness metrics (risk-reward ratio and loss-win ratio) are better than for U_{Risk} across the α values. For the metric "Loss > 20%", they are comparable between SARO and U_{Risk}, given a topic sample as large as 9685 in size.

Next, for FARO, the observed values of the two quality and robustness metrics are comparable with that of the risk-sensitive optimisation based on U_{Risk} across α values, and for the metric, "Loss > 20%" the observed values for FARO are slightly worse than that of both U_{Risk} and SARO.

To summarise, the empirical evidence in Table 3 suggest that (i) FARO is best suited for retrieval tasks that are not tolerant to any loss in average effectiveness but also require robustness in effectiveness across the topics, and (ii) SARO suits retrieval tasks that require primarily robustness but are tolerant to some loss in the achievable average effectiveness.

[4]All of our code has been integrated to Jforests, available at https://code.google.com/p/jforests/

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

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

[7]Similar measures are reported in [26]. With 9685 topics, all NDCG differences are statistically significant.

7. RELATED WORK

To the best of our knowledge, this paper is the first work examining risk-sensitive evaluation from the perspective of statistical inference. Indeed, while there has been some investigation into measures of robustness in the literature, such as Geometric-Mean Average Precision [24], developed within the context of the TREC 2004 Robust track, this paper advances upon the U_{Risk} measure, first proposed in [26] in 2012. The T_{Risk} measure is the test statistic counterpart of U_{Risk}, which enables hypothesis testing on the level of risk associated with a given IR system. As a result, it facilitates adaptive risk-sensitive optimisation within learning to rank.

Outside of risk-sensitive evaluation, statistical hypothesis testing has a long history within IR. Van Rijsbergen [22] noted that *"there are no known statistical tests applicable to IR"*. However, later, Hull [32] recommended various hypothesis tests for the evaluation of retrieval experiments, including the Student's t test for matched pairs. Zobel [31] was the first to apply re-sampling techniques in IR, by using a leave-one-out technique for assessing the effect of pooling on the effectiveness measurements and the significance of hypothesis tests, including the paired t test and the Wilcoxon signed rank test. Later, Smucker et al. [19, 20] and also Urbano et al. [27] investigated nonparametric re-sampling techniques, such as the bootstrapping and permutation tests, for the purposes of the evaluation of retrieval experiments.

Finally, much work in developing effective learning to rank techniques has occurred in the last few years, as reviewed by Liu [16]. Macdonald et al. [17] examined how the choice of evaluation measure encoded within their loss functions impacted upon the effectiveness of various learning to rank techniques. In particular, it is notable that the AdaRank technique [16, Ch. 4] focuses on hard queries using boosting. Taking a different approach, Wang et al. [26] proposed a risk-sensitive optimisation for the state-of-the-art LambdaMART technique, based on their U_{Risk} measure. We further extend U_{Risk} to the new T_{Risk} measure within this paper, which is both theoretically founded, and results in more effective and less risky learning to rank.

8. CONCLUSIONS

This paper proposed the new T_{Risk} measure for risk-sensitive evaluation, which is theoretically grounded within hypothesis testing. It easily allows inferential hypothesis testing of risk, as well as the exploratory identification of topics that commit significant levels of risk. In particular, we showed how T_{Risk} could be integrated within the state-of-the-art LambdaMART learning to rank technique, to permit effective yet risk-averse retrieval. Indeed, compared to the existing U_{Risk} measure, we attain higher effectiveness with comparable or better risk/reward tradeoffs. For future work, we believe that there is a huge scope to build further effective and risk-averse adaptations for learning to rank upon T_{Risk}, other than SARO and FARO, and beyond LambdaMART.

9. REFERENCES

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

[2] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. SIGIR*, 2000.

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

[4] C. J. C. Burges. From RankNet to Lambdarank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, 2010.

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

[6] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, 2009.

[7] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2^{nd} edition, 1988.

[8] K. Collins-Thompson. Accounting for stability of retrieval algorithms using risk-reward curves. In *Proc. Future of Evaluation in IR Workshop*. SIGIR. 2009.

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

[10] K. Collins-Thompson, P. N. Bennett, F. Diaz, C. Clarke, and E. Voorhees. TREC 2013 Web Track Guidelines.

[11] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, (7):1–26, 1979.

[12] H. Fisher. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, 2010.

[13] Y. Ganjisaffar, R. Caruana and C. Lopes. Bagging Gradient-Boosted Trees for High Precision, Low Variance Ranking Models. In *Proc. SIGIR*, 2011.

[14] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors. *Understanding robust and exploratory data analysis*. 1983.

[15] R. V. Hogg, A. T. Craig, and J. W. McKean. *Introduction to Mathematical Statistics*. 6^{th} edition, 2004.

[16] T.-Y. Liu. Learning to rank for information retrieval. *Foundations & Trends in IR*, 3(3):225–331, 2009.

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

[18] M. Quenouille. Approximate tests of correlation in time series. *J. Royal Statistical Society*, (11):18–84, 1949.

[19] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. CIKM*, 2007.

[20] M. D. Smucker, J. Allan, and B. Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proc. SIGIR*, 2009.

[21] J. W. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29(2):614, 1958.

[22] C. van Rijsbergen. *Information Retrieval*. 2^{nd} edition, 1979.

[23] E. M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proc. TREC*, 2003.

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

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

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

[27] J. Urbano, M. Marrero, and D. Martín. On the measurement of test collection reliability. In *Proc. SIGIR*, 2013.

[28] W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In *Proc. SIGIR*, 2008.

[29] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Annuals of Statistics*, 14:1261–1350, 1986.

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

[31] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, 1998.

[32] J. Zobel. Using statistical testing in the evaluation of retrieval experiments. In *Proc. SIGIR*, 1993.