Information Processing and Management 59 (2022) 102747

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# A bias–variance evaluation framework for information retrieval systems

Peng Zhang [a,*], Hui Gao [a], Zeting Hu [a], Meng Yang [a], Dawei Song [b,*], Jun Wang [c], Yuexian Hou [a], Bin Hu [d]

[a] College of Intelligence and Computing, Tianjin University, Tianjin, China
[b] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[c] Department of Computer Science, University College London, London, UK
[d] School of Information Science and Engineering, Lanzhou University, Lanzhou, China

ARTICLE INFO

*Keywords:*
Information retrieval
Evaluation metrics
Effectiveness–stability tradeoff

ABSTRACT

In information retrieval (IR), the improvement of the effectiveness often sacrifices the stability of an IR system. To evaluate the stability, many risk-sensitive metrics have been proposed. Since the theoretical limitations, the current works study the effectiveness and stability separately, and have not explored the effectiveness–stability tradeoff. In this paper, we propose a Bias–Variance Tradeoff Evaluation (BV-Test) framework, based on the bias–variance decomposition of the mean squared error, to measure the overall performance (considering both effectiveness and stability) and the tradeoff between effectiveness and stability of a system. In this framework, we define generalized bias–variance metrics, based on the Cranfield-style experiment set-up where the document collection is fixed (across topics) or the set-up where document collection is a sample (per-topic). Compared with risk-sensitive evaluation methods, our work not only measures the effectiveness–stability tradeoff of a system, but also effectively tracks the source of system instability. Experiments on TREC Ad-hoc track (1993–1999) and Web track (2010–2014) show a clear effectiveness–stability tradeoff across topics and per-topic, and topic grouping and max–min normalization can effectively reduce the bias–variance tradeoff. Experimental results on TREC Session track (2010–2012) also show that the query reformulation and increase of user data are beneficial to both effectiveness and stability simultaneously.

## 1. Introduction

In the Information Retrieval (IR) community, the design of evaluation metrics plays an important role for developing advanced IR models and systems (A survey on evaluation of summarization methods, 2019; An in-depth study on diversity evaluation: The importance of intrinsic diversity, 2017; Ranking themes on co-word networks: Exploring the relationships among different metrics, 2018). Generally speaking, the improvement of mean retrieval effectiveness for all topics over a baseline may sacrifice the retrieval stability, meaning that the performance (e.g., Average Precision, AP) for some individual topics may decline, compared with the baseline (Amati, Carpineto, & Romano, 2004; Collins-Thompson, 2009a; Collins-Thompson, Macdonald, Bennett, Diaz, & Voorhees, 2014; Zighelnic & Kurland, 2008). A number of risk metrics (e.g., < *init* Amati et al., 2004 and Robustness Index Collins-Thompson, 2009b) were proposed to evaluate the downside performance or the stability of a system (Amati et al., 2004; Bah & Carterette, 2015; Collins-Thompson, 2009b). These metrics, however, were designed separately from the mean effectiveness metrics. In Zhang,

* Corresponding authors.
*E-mail addresses:* pzhang@tju.edu.cn (P. Zhang), dawei.song2010@gmail.com (D. Song).

https://doi.org/10.1016/j.ipm.2021.102747
Received 25 January 2021; Received in revised form 6 August 2021; Accepted 30 August 2021
Available online 6 October 2021

Song, Wang, and Hou (2013), it has been shown that the overall performance cannot be solely decided by either an effectiveness or a stability metric. Indeed, there is a lack of a systematic evaluation strategy that is able to not only unify the effectiveness and stability, but also analyze the *tradeoff* between them.

Various risk-sensitive evaluation metrics have been proposed to unify effectiveness and stability (Dinçer, Macdonald, & Ounis, 2014; Dinçer, Macdonald, & Ounis, 2016; Wang, Bennett, & Collins-Thompson, 2012). However, they often obtain a unified metric from the component sub-metrics (each representing either effectiveness or stability), which is defined in a bottom-up manner (e.g., $A + B \rightarrow C$). For example, the $U_{Risk}$ (Wang et al., 2012) creates a win–loss distribution over all topics, where $win$ means the current system outperforms a certain baseline and $loss$ means the opposite situation. The total wins and the total losses represent reward and risk, respectively. The result of $U_{Risk}$ is a linear combination of risk and reward. In addition, the latest risk-sensitive metrics $GeoRisk$ (Dinçer et al., 2016) combines the standard mean effectiveness (e.g., MAP) with a robustness metric $Z_{Risk}$ (Dinçer et al., 2016). Such a bottom-up manner in the risk-sensitive evaluation metric is insufficient to explore the effectiveness–stability tradeoff, since the intrinsic relation between component metrics is not modeled in the evaluation framework. Indeed, this bottom-up approach is easy to control. However, in the bottom-up manner, the effectiveness and the stability can only be modeled separately. We need a framework that can diagnose or analyze the effectiveness–stability tradeoff in a unified theoretical framework.

Therefore, it is essential to build a unified metric/framework through a top-down manner, and it is expected that such a unified metric can evaluate the overall performance (considering effectiveness and stability simultaneously). Motivated by the analysis of the learning algorithm's expected error in machine learning (Bishop, 2006; Papo, 2019; Qiu, Hu, & Wu, 2014), we propose a bias–variance Tradeoff Evaluation (BV-Test) framework, to systematically analyze effectiveness and stability as well as the effectiveness–stability tradeoff.

In the top-down manner, the goal is to measure the difference between the current system and the target system. The difference is formalized as an error, which can be decomposed into two aspects, namely bias and variance. These two aspects have an intrinsic relation, i.e., bias–variance tradeoff (Geman, Bienenstock, & Doursat, 1992a; Zhang et al., 2013). We can use bias–variance tradeoff to model the retrieval effectiveness–stability tradeoff. Then, we can use the intrinsic relation of bias and variance to analyze the effectiveness–stability tradeoff.

The relation between bias and variance (also called bias–variance tradeoff Geman, Bienenstock, & Doursat, 1992b) is well-studied in machine learning, where a learning algorithm's expected error can be decomposed into bias and variance. Therefore, we propose to apply bias–variance decomposition theory to IR evaluation and utilize a virtual target system (Dinçer, Ounis, & Macdonald, 2014; Zhang, Hao et al., 2014) as a baseline. The target system represents the best performance for each topic over all the considered retrieval systems. The performance difference between the current system and the target system can be regarded as a kind of overall error to measure the overall performance. Similarly, the overall error can be decomposed into the sum of bias and variance. On the one hand, the bias represents the difference between the expected performance of the current system and that of a target system, which reflects the mean effectiveness of the current system. On the other hand, the variance represents the variation of the performance across different data samples (e.g., different topics or different document collections), which reflects the stability of the current system. The smaller the bias and variance are, the better the system's effectiveness and stability tend to be. Therefore, the bias–variance tradeoff can be naturally utilized to evaluate the retrieval effectiveness–stability tradeoff.

We will study the bias–variance evaluation from two categories, namely across-topic and per-topic. When the document collection is fixed in the Cranfield-style experiment, we define the bias–variance evaluation across topics. Cormack, Lynam (Cormack & Lynam, 2006) and Robertson (Sparck Jones, Walker, & Robertson, 2000) suggested that the document collection in the Cranfield-style experiments was just a sample from a larger document population, so that the score for a system on a topic varies with different samples. In this case, we define the per-topic bias–variance evaluation.

In order to verify BV-Test framework, we carried out a series of experiments. First, the experimental results on Ad-hoc track 1993–1999 and Web track 2010–2014 show that the bias–variance tradeoff often exists in the evaluation process of information retrieval. Second, compared with the risk-sensitive metrics, the bias–variance metrics can be used to identify different causes/factors of system instability. Finally, the query reformatting task is carried out on the Session track 2010–2012. The experimental results show that adding additional data can effectively reduce the bias–variance tradeoff.

## 2. Related work

### 2.1. Risk-sensitive evaluation

Many techniques in IR, such as query expansion (Amati et al., 2004; Carmel, Farchi, Petruschka, & Soffer, 2002; Deveaud, Mothe, Ullah, & Nie, 2018) and learning to rank (Dai, Shokouhi, & Davison, 2011; Dinçer et al., 2016; Macdonald et al., 2013), behave differently across topics, often improving the effectiveness of some topics while hurting the others' effectiveness, resulting in retrieval instability/risk. To measure such risk, there has been an increasing focus on the risk-sensitive evaluation across topics (Collins-Thompson, 2009a, 2009b; Collins-Thompson, Bennett, Diaz, Clarke, & Vorhees, 2014; de Sousa, Canuto, Gonçalves, Rosa, & Martins, 2019; Wang et al., 2012).

In the risk-sensitive evaluation, the basic metric is $U_{Risk}$ (Wang et al., 2012). Its basic idea is to balance the system's reward and risk, and the sum of them only reflects the absolute difference in effectiveness between the current system and baseline. Motivated by the hypothesis testing, Dincer et al. proposed the $T_{Risk}$ (Dinçer et al., 2014) to infer whether a system exhibits a real risk. $T_{Risk}$ was obtained by a linear transformation of $U_{Risk}$.

However, the baseline of above two metrics is a single system, and using a single system as a baseline suffers from that evaluation results vary with baseline (Dinçer et al., 2014; Kharazmi, Scholer, Vallet, & Sanderson, 2016). In order to address this problem, $Z_{Risk}$ (Dinçer et al., 2016) was proposed by combining the idea of $U_{Risk}$ (Wang et al., 2012) with Chi-square test, in which the baseline was composed of multiple systems. However, in (Dinçer et al., 2016), it also showed that $Z_{Risk}$ ignored system's mean effectiveness (Dinçer et al., 2016), the *GeoRisk* was proposed (Dinçer et al., 2016) to consider the mean effectiveness and stability simultaneously.

We systematically compare the bias–variance metric with the existing risk-sensitive metrics from four aspects, i.e., baseline, effectiveness, stability, and tradeoff (see Section 4 for details). In addition to the differences from the first three aspects, we find that the bias and variance decomposed from the overall error have an intrinsic tradeoff, which can be further quantified. Moreover, we investigate the variance in-depth and find that if we decompose the effectiveness score difference between the current system and the target system, we can trace where the instability comes from.

### 2.2. Previous bias–variance work

We first summarize different bias in different scenarios. Generally speaking, the "bias" often occurs at the *data* level, the *model* level and, the *estimation or evaluation* level of a task (Mitchell et al., 1997; Robertson, 1981; Voorhees et al., 2005).

At the data level, the most typical data bias is *selection bias* (Cortes, Mohri, Riley, & Rostamizadeh, 2008; Phillips et al., 2009; Williamson, Gamble, Altman, & Hutton, 2005), which refers to the bias caused by the non-randomness of sample selection in the research process. Selection bias in IR tasks may lead to better performance for some topics, and bad performance for others. Theoretically, the more topics the dataset contains, the less influence of selection bias, and the more reliable the results of the model prediction (Voorhees & Buckley, 2002a).

At the model level, the bias usually refers to *inductive bias* in machine learning (Mitchell et al., 1997). By assuming a proper inductive bias, a model can learn a better objective function through some assumptions corresponding to the inductive bias, so as to get a learner with better accuracy or generalization performance.

In the estimation theory (Geman et al., 1992a; Lerman, Veshchikov, Markowitch, & Standaert, 2018), *the estimation bias* describes the difference between the predicted value of the current estimator and the target true value. In machine learning, the purpose of bias and variance is to evaluate the difference between the predicted value of a learned function and the real value, so as to select the best model, under the given training data.

At the evaluation level, our work aims to propose an evaluation framework based on the bias–variance decomposition. The bias in this framework is used to evaluate the performance gap between the current model and the target model under a set of given topics.

In Zhang et al. (2013), Zhang, Song, Wang and Hou (2014), Zhang et al. studied the bias–variance tradeoff of query models and investigated when and why the tradeoff occurs and how to avoid the tradeoff using bias–variance evaluation. In Zhang, Hao et al. (2014), the bias–variance metric applied to evaluate TREC participated systems. The experimental results show that the evaluated systems have obvious tradeoff between effectiveness and stability. At the same time, a strong and unbiased "baseline" is applied to the work as a virtual target model.

Compared with the earlier work (Zhang, Hao et al., 2014), the main extensions of this paper are as follows:

- We further generalize the bias–variance formulation, and propose a generalized framework to evaluate the bias and variance of IR systems.
- In addition to the bias–variance evaluation across topics, we also propose to study the per-topic effectiveness and stability and the relation between them, based on the per-topic bias and variance.
- We systematically compare bias–variance evaluation framework with risk-sensitive evaluation metrics.
- We study the effects of query reformulation on improving system performance in the interactive search scenario.

## 3. A Bias–Variance Evaluation (BV-test) framework

In this section, we first introduce the background knowledge of the bias–variance decomposition. Then, we propose a bias–variance evaluation (BV-Test) framework to evaluate the effectiveness and stability of a retrieval system.

### 3.1. Introduction to the bias–variance decomposition theory

The bias–variance decomposition theory has been applied into many areas, such as density estimation, regression, classification (Domingos, 2000; Heskes, 1998; Suen, Melville, & Mooney, 2005; Valentini & Dietterich, 2004). In machine learning, the bias–variance decomposition is imposed on the analysis of the learning algorithm's expected error. Such errors can be decomposed into two main components, bias and variance, which will be described in the following.

For a data point $\mathbf{x}$, $y$ is its label and $f(\cdot)$ is the ideal model that can output the true label, i.e., $y = f(\mathbf{x})$. $\hat{f}(\mathbf{x}; D)$ is the output of the actual model $\hat{f}(x)$ learned from the data sample $D$. $\mathbb{E}_D\hat{f}(\mathbf{x}; D)$ is the expectation of $\hat{f}(\mathbf{x}; D)$ over different samples $D$. We denote such an expectation is as follows:

$$\bar{f}(\mathbf{x}) = \mathbb{E}_D\hat{f}(\mathbf{x}; D) \tag{1}$$

The bias is the squared error between the expectation value $\bar{f}(\mathbf{x})$ and the true value $f(\mathbf{x})$:

$$Bias^2(\hat{f}) = (\bar{f}(\mathbf{x}) - f(\mathbf{x}))^2 \tag{2}$$

The smaller $Bias^2$ signifies that model's expected output value $\bar{f}(\mathbf{x})$ is closer to the true label $f(\mathbf{x})$. The variance of the output value can be defined as follows:

$$Var(\hat{f}) = \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \tag{3}$$

The above formula measures the variation of learning performance caused by changes on different data sample $D$. The expected error of learning algorithm is formulated as follows:

$$\begin{aligned} E(f; D) &= \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - f(\mathbf{x}))^2 \\ &= \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - f(\mathbf{x}))^2 \\ &= \mathbb{E}_D(\bar{f}(\mathbf{x}) - f(\mathbf{x}))^2 + \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \\ &+ 2 * \mathbb{E}_D(\bar{f}(\mathbf{x}) - f(\mathbf{x})) * \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x})) \end{aligned} \tag{4}$$

In the last line, since $\mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x})) = \mathbb{E}_D\hat{f}(\mathbf{x}; D) - \mathbb{E}_D\bar{f}(\mathbf{x})$, $\mathbb{E}_D\hat{f}(\mathbf{x}; D) = \bar{f}(\mathbf{x})$ and $\mathbb{E}_D\bar{f}(\mathbf{x}) = \bar{f}(\mathbf{x})$, we can get $\mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x})) = \bar{f}(\mathbf{x}) - \bar{f}(\mathbf{x}) = 0$. It turns out that $2 * \mathbb{E}_D(\bar{f}(\mathbf{x}) - f(\mathbf{x})) * \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x})) = 0$. Then, we have:

$$\begin{aligned} E(f; D) &= \mathbb{E}_D(\bar{f}(\mathbf{x}) - f(\mathbf{x}))^2 + \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \\ &= (\bar{f}(\mathbf{x}) - f(\mathbf{x}))^2 + \mathbb{E}_D(\hat{f}(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \\ &= Bias^2(\hat{f}) + Var(\hat{f}) \end{aligned} \tag{5}$$

From the first line to the second line in Eq. (5), as the true value $f(\mathbf{x})$ is independent of different data samples ($D$), and expected value $\bar{f}(\mathbf{x})$ is fixed for all the data samples ($D$) (see Eq. (1)), we can derive $\mathbb{E}_D(\bar{f}(\mathbf{x}) - f(\mathbf{x}))^2 = (\bar{f}(\mathbf{x}) - f(\mathbf{x}))^2$.

Eq. (5) demonstrates that the expected error can be decomposed into the sum of $Bias^2$ and $Var$. The bias is regarded as the difference between the expected/average output of the model and the correct value (see Eq. (2)), and the variance is considered as the variability of the algorithm on different data samples (see Eq. (3)). Generally speaking, there is a bias–variance dilemma between $Bias^2$ and $Var$ (Domingos, 2000; Geman et al., 1992b).

### 3.2. An IR evaluation framework based on bias–variance decomposition

Now, we will show how to formulate a similar bias–variance decomposition in the IR evaluation. In information retrieval, an evaluation metric, e.g., Average Precision (AP), is needed to measure the performance of a retrieval system. Let $M$ denote an evaluation metric, $f$ denote a target system and which is an ideal system that correctly computes the relevance status of each document, and $\hat{f}$ be the current system that needs to be evaluated. Let $S$ be a data sample that system is performed on and the current system can be evaluated by the evaluation metric $M$ on different samples $S$. Now, we first define the general bias–variance formulation given any sample $S$.

Let $M(f)$ be the performance of the target system $f$, and $M(\hat{f}; S)$ be the evaluation result of the current system $\hat{f}$ on data sample $S$, where $M$ denotes an effectiveness evaluation metric (e.g., AP). Based on Eqs. (4) and (5), the corresponding form in IR evaluation can be defined as follows:

$$\begin{aligned} &\mathbb{E}_S(M(\hat{f}; S) - M(f))^2 \\ =&\mathbb{E}_S(M(\hat{f}; S) - \overline{M}(\hat{f}; S) + \overline{M}(\hat{f}; S) - M(f))^2 \\ =&\mathbb{E}_S(\overline{M}(\hat{f}; S) - M(f))^2 + \mathbb{E}_S(M(\hat{f}; S) - \overline{M}(\hat{f}; S))^2 \\ =&Bias_S^2(M(\hat{f})) + Var_S(M(\hat{f})) \end{aligned} \tag{6}$$

where $\overline{M}(\hat{f}, S) = \mathbb{E}_S(M(\hat{f}, S))$, and it reflects the mean performance of a system over samples $S$. For the bias, as the $\overline{M}(\hat{f}; S)$ and $M(f)$ represents the mean effectiveness of $\hat{f}$ and $f$, $Bias_S^2(M(\hat{f}))$ can reflect the effectiveness difference between $f$ and $\hat{f}$. For the variance, as it calculates the sum of total difference between the different effectiveness scores $M(\hat{f}; S)$ and $\overline{M}(\hat{f}; S)$, $Var_S(M(\hat{f}))$ reflects the stability of $\hat{f}$ across all the data samples. The total error $\mathbb{E}_S(M(\hat{f}, S) - M(f))^2$ represents the gap between current system $\hat{f}$ and target system $f$, which indicates a system's overall performance on samples $S$. The smaller the bias and variance are, the better the system's effectiveness and stability will be. In addition, the relation (e.g., the tradeoff) between bias and variance can be explored.

In information retrieval, $S$ can be a topic (or a query). The evaluation should be carried out on a set of topics and a given document collection. This is a typical Cranfield-style evaluation set-up, e.g., Text REtrieval Conference (TREC) evaluation task (Donna, 1994). Recently, Cormack and Lynam postulated that the document collection is a sample from a large document population (Cormack & Lynam, 2006). Under this assumption, we can have many simulated document collections as samples, yielding the per-topic variance (Robertson & Kanoulas, 2012). Therefore, $S$ can also be a document collection or a simulated one.

Based on different sample sources, bias–variance interpretation has different meanings. If we consider a topic as each sample to evaluate the retrieval system, we can define the bias–variance formulation across topics (see Section 3.3). On the other hand, if we regard a document collection as a sample, we can define the per-topic bias–variance formulation (see Section 3.4).

**Table 1**
The score matrix of an IR evaluation,when each sample is a topic.

| systemtopic | $t_1$ | $t_2$ | ... | $t_j$ | ... | $t_n$ | Mean |
|---|---|---|---|---|---|---|---|
| $\hat{f}_1$ | $x_{1,1}$ | $x_{1,2}$ | ... | $x_{1,j}$ | ... | $x_{1,n}$ | $\overline{x_1}$ |
| $\hat{f}_2$ | $x_{2,1}$ | $x_{2,2}$ | ... | $x_{2,j}$ | ... | $x_{2,n}$ | $\overline{x_2}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $\hat{f}_i$ | $x_{i,1}$ | $x_{i,2}$ | ... | $x_{i,j}$ | ... | $x_{i,n}$ | $\overline{x_i}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $\hat{f}_m$ | $x_{m,1}$ | $x_{m,2}$ | ... | $x_{m,j}$ | ... | $x_{m,n}$ | $\overline{x_m}$ |
| target system($f$) | $x_{T,1}$ | $x_{T,2}$ | ... | $x_{T,j}$ | ... | $x_{T,n}$ | $\overline{x_T}$ |

### 3.3. Bias–variance evaluation across topics

In IR evaluation, the main task is to evaluate the performance of an IR system given an evaluation metric (e.g., $AP$) and a test collection (with a set of test queries/topics and a document collection). In this condition, the evaluation is based on different topics, i.e., the bias–variance evaluation across topics.

To illustrate the bias–variance evaluation across topics, Table 1 presents an example data matrix composed of $m$ systems and $n$ topic, where $\hat{f}_i$ and $t_j$ are the $i$th system under test and the $j$th topic sample, respectively. Given an effectiveness metric (e.g., $AP$ or $ERR@20$) and a retrieval system $\hat{f}_i$, the variable $x_i$ represents the effectiveness scores (see the $i$th row of the data matrix) on all the topics, where $x_{i,j}$ is the effectiveness score of $f_i$ on the topic $t_j$. Meanwhile, the mean value $\overline{x}_i = Sum(x_i)/n$, where $Sum(x_i) = x_{i,1} + x_{i,2} + \cdots + x_{i,j} + \cdots + x_{i,n}$. For the target system $f$, its performance on each topic $t_j$ is denoted as $x_{T,j}$.

According to Eq. (6), when each sample ($S$) is a topic and the evaluation metric ($M$) is an effectiveness metric, $x_i$ corresponds to $M(\hat{f}_i; S)$. Meanwhile, let the fixed constant value $M(f)$ be a constant $c$. Then, based on Eq. (6) and the above notations, we can define the bias–variance decomposition across topics as follows:

$$\begin{aligned} E(x_i - c)^2 &= E(x_i - \overline{x_i} + \overline{x_i} - c)^2 \\ &= E(\overline{x_i} - c)^2 + E(x_i - \overline{x_i})^2 \\ &= Bias^2(x_i) + Var(x_i) \end{aligned} \tag{7}$$

where $\overline{x_i}$, $Bias^2(x_i)$ and $Var(x_i)$ correspond to the $\overline{M}(\hat{f}; S)$, $Bias^2_S(M(\hat{f}))$ and $Var_S(M(\hat{f}))$ in Eq. (6), respectively.

For the bias, as the value of $c$ represents the performance of target system, $Bias^2(x_i)$ measures the squared difference between the expected effectiveness of the system $\hat{f}_i$ across all topics and the target value $c$. The larger the bias is, the worse the effectiveness of the current system will be. For the variance, it measures the retrieval stability of $\hat{f}_i$. The larger the variance is, the worse the stability of a system is. For the overall error, $E(x_i - c)^2$ reflects the overall difference between the current system $\hat{f}_i$ and the target system $f$, and it can be directly decomposed to the bias and variance, corresponding to retrieval effectiveness and stability, respectively.

In Eq. (7), the value of $c$ is a constant value and it corresponds to the performance of the target system. In the following, we discuss the bias–variance evaluation under different choices of $c$.

First, $c$ can be set as the maximum value of an effectiveness metric. Suppose we use $AP$ as the effectiveness metric, $M(f) = c = 1$. For the bias, $Bias^2(x_i) = (\overline{x_i} - 1)^2$, the mean value $\overline{x}_i$ corresponds to $MAP$. Thus $Bias^2(x_i) = (1 - MAP)^2$. Since the actual $MAP$ value is always not greater than 1, the mean effectiveness score ($\overline{x_i}$) will be proportional to $Bias^2(x_i)$ in a reverse direction. In other words, the larger the mean value $\overline{x}_i$ (e.g., $MAP$) is, the smaller the $Bias^2(x_i)$ will be. For the variance, $Var(x_i) = E(x_i - \overline{x_i})^2$, which represents the stability of a system over all the topic samples and is independent of the value of $c$.

Second, in practice, $c = 1$ is not achievable for any retrieval system. Therefore, in our previous work (Zhang et al., 2013), we designed a target system $f$ with the best performance for each topic set among all the considered systems, also including the current system under test. As Table 1 shows, we use a variable $x_T$ to denote the effectiveness of the target system, and the effectiveness score of this system on the topic $t_j$ is $x_{T,j} = max(x_{1,j}, x_{2,j}, \ldots, x_{i,j}, \ldots, x_{m,j})$. Since $c$ should be a constant value, we can let the value of $c$ equal to the mean value of $x_T$, i.e., $c = \overline{x_T}$.

The above analysis can be proved by the following process. Referring to Table 1 and Eq. (7), Proof 1 proves that the value of $c$ in a certain interval $[\overline{x_T}, 1]$ has the same effect on the result of the BV-Test framework.

**Lemma 1.** *As long as $c \in [\overline{x_T}, 1]$, it can ensure that both the bias and variance values are consistent for different $c$.*

**Proof.** For bias metric, it can be obtained from Eq. (7):

$$Bias^2(x_i) = E(\overline{x_i} - c)^2 = (\overline{x_i} - c)^2 \tag{8}$$

Derivation of $\overline{x_i}$:

$$\frac{d(Bias^2(x_i))}{d(\overline{x_i})} = 2(\overline{x_i} - c) \tag{9}$$

In any case $c \geq \overline{x_T} \geq \overline{x_i} \geq 0$, $(\overline{x_i} - c) \leq 0$. It shows that $Bias^2(x_i)$ decreases monotonously with $\overline{x_i}$. When $\overline{x_i}$ monotonically increases, $(\overline{x_i} - c)$ monotonically increases, $(\overline{x_i} - c)^2$ monotonically decreases, that is, $Bias^2(x_i)$ monotonically decreases. The larger the $\overline{x_i}$, the

**Table 2**
An example to illustrate the effect of target system on bias–variance evaluation.

| systemtopic | $t_1$ | $t_2$ | $t_3$ | Mean |
|---|---|---|---|---|
| $\hat{f}_1$ | 0.8 | 0.9 | 0.4 | 0.7 |
| $\hat{f}_2$ | 0.5 | 0.6 | 0.7 | 0.6 |
| $\hat{f}_3$ | 0.3 | 0.6 | 0.3 | 0.4 |
| target system ($f$) | 0.8 | 0.9 | 0.7 | 0.8 |

smaller the $Bias^2(x_i)$, and the better the effectiveness of the model. It can be deduced that $\overline{x_i}$ is inversely proportional to $Bias^2(x_i)$, for which its changing trend is independent of the value of $c \in [\overline{x_T}, 1]$.

For variance metric, it can be obtained from Eq. (7):

$$Var(x_i) = E(x_i - \overline{x_i})^2 \tag{10}$$

where the calculation process of variance does not include $c$, so the choice of $c$ does not affect the variance metric. □

We will show that the bias–variance evaluation based on $c = \overline{x_T}$ has the same effect with that based on the previous setting $c = 1$ through examples. First, for the variance, as we showed in Eq. (7) , the variance term is independent of the value of $c$, so that the variances in the two cases are the same. Second, for the bias, no matter whether the value of $c$ is 1 or $\overline{x_T}$, $c$ is a constant and fixed value for all the considered systems and the systems' ranking based $Bias^2(x_i)$ is the same. For example, in Table 2, the mean effectiveness of $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ are $\overline{x}_1 = 0.7$, $\overline{x}_2 = 0.6$ and $\overline{x}_3 = 0.4$, respectively. When $c = 1$, the corresponding bias of $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ are $Bias^2(x_1) = 0.09$, $Bias^2(x_2) = 0.16$, and $Bias^2(x_3) = 0.36$, respectively. We can infer that the systems' ranking based on $Bias^2(\hat{f}_i)$ is $\hat{f}_1 > \hat{f}_2 > \hat{f}_3$. When $c = \overline{x_T} = 0.8$, $Bias^2(x_1) = 0.01$, $Bias^2(x_2) = 0.04$, and $Bias^2(x_3) = 0.16$, so the ranking of $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ has not changed. Please also note that we can also adopt a normalization method to enforce $c$ to be 1, as described later.

With reference to Moffat's work (Moffat, 2013), we analyze the numeric properties of bias from the aspects of boundedness and monotonicity. In the BV-Test framework, the choice of effectiveness metric will have an impact on the numerical properties of the bias metric.

Considering the **boundedness** of the bias, if the effectiveness metric (such as $AP$ and $NDCG$) is boundedness (Moffat, 2013), and the score of the target system is higher than that of the current system under test and $\leq 1$, the gap between the current system and the target system (refer to the Eq. (6)) is also in the range $[0, 1]$, so the bias conforms to the boundedness.

For **monotonicity**, if the effectiveness metric satisfies monotonicity (such as $AP$), then the bias also satisfies monotonicity. On the other hand, if the effectiveness metric does not satisfy monotonicity (such as $NDCG$), then the bias does not satisfy monotonicity either.

*The Method of Topic Grouping* Now, we discuss the situation when each sample ($S$) contains a *group* of topics. In this case, $M$ in Eq. (6) then denotes the mean effectiveness metric, such as $MAP$. Accordingly, $x_{i,j}$ of Table 1 represents the mean effectiveness scores of the $j$th topic group of the $i$th system under test. Compared with the aforementioned single topic sampling, the topic group sampling can avoid adverse influence of some too big/small values in the computation of variance. In practice, all the topics are partitioned into several subsets, which has one or more topics. The topic set containing a group of topics is constructed by the random partitioning or by the topic difficulty (see the experiment section for details).

*The Method of Normalization* There are two main factors that cause the retrieval instability (Ferro, Kim, & Sanderson, 2019; Voorhees & Buckley, 2002b). One is about the topic effect, i.e., different topics have different inherent retrieval difficulties, which causes the variance of retrieval performance over topics. The other is about the retrieval system effect, i.e., different retrieval systems perform differently, which causes the retrieval instability over topics. If one wants to focus on evaluation of the second effect caused by a retrieval system, the topic effect should be normalized before evaluation.

In this paper, we choose the max–min normalization (Jain, Nandakumar, & Ross, 2005; Patro & Sahu, 2015; Zhao, Kleinhans, Sandhu, Patel, & Unnikrishnan, 2019), also known as dispersion normalization, which is a linear transformation of the original data, mapping the resulting values between 0 and 1. According to Table 1, for the given topic set sample, the max of observed scores for all systems is denoted as $Max$ ($Max = max(x_{1,j}, x_{2,j}, \ldots, x_{i,j}, \ldots, x_{m,j})$) and the corresponding min is denoted as $Min$ ($Min = min(x_{1,j}, x_{2,j}, \ldots, x_{i,j}, \ldots, x_{m,j})$). Given the effectiveness score $x_{i,j}$, the max–min score for that topic is:

$$x'_{i,j} = \frac{x_{i,j} - Min}{Max - Min} \tag{11}$$

For the corresponding $x'_{i,j}$ in Eq. (11), there are two advantages. One is that, the value of $c$ in Eq. (7) are forced to 1, which means that the variance of the target system is 0. Therefore, we can focus on the performance variability caused by different systems rather than topic difficulties. The other is that the scores of every topic for each system become smoother, which could alleviate the bias–variance tradeoff. We will further explore whether the tradeoff will decrease in Section 5.3.

### 3.4. The per-topic bias–variance evaluation

Researchers (Cormack & Lynam, 2006) proposed that the document collection in Cranfield-style experiment is a sample from a large document population. Under this notion, the per-topic metric value is considered as an *estimation* of the metric's *true value* for

**Table 3**
The data matrix for the results of an IR experiment, when each sample is a document collection.

| systemsample | $dc_1$ | $dc_2$ | ... | $dc_j$ | ... | $dc_n$ | Mean |
|---|---|---|---|---|---|---|---|
| $\hat{f}_1$ | $y_{1,1}$ | $y_{1,2}$ | ... | $y_{1,j}$ | ... | $y_{1,n}$ | $\overline{y_1}$ |
| $\hat{f}_2$ | $y_{2,1}$ | $y_{2,2}$ | ... | $y_{2,j}$ | ... | $y_{2,n}$ | $\overline{y_2}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $\hat{f}_i$ | $y_{i,1}$ | $y_{i,2}$ | ... | $y_{i,j}$ | ... | $y_{i,n}$ | $\overline{y_i}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $\hat{f}_m$ | $y_{m,1}$ | $y_{m,2}$ | ... | $y_{m,j}$ | ... | $y_{m,n}$ | $\overline{y_m}$ |
| target system($f$) | $y_{T,1}$ | $y_{T,2}$ | ... | $y_{T,j}$ | ... | $y_{T,n}$ | $\overline{y_T}$ |

the corresponding topic in the whole document population. For different document collection samples, the per-topic metric values are different. As a result, the per-topic variance is generated (Robertson & Kanoulas, 2012). In addition to the per-topic variance, we will also study the per-topic bias, as well as their tradeoff, in this paper.

The implementation of document collection being a sample is through the method of simulation, such as bootstrap (Ferro & Sanderson, 2019). Given a system-topic pair, suppose there are 1000 retrieved documents with 1000 scores correspondingly, and a number (denoted as $r$) of them are relevant. Following (Robertson & Kanoulas, 2012), using Poisson distribution with a mean $r$, we take a sample $r_s$ from it. Specially, $r_s$ simulated scores are taken from the distribution of relevant scores and $1000-r_s$ simulated scores are from the score distribution of non-relevant ones. These 1000 simulated scores are sorted in a descending order. Then, a binary IR metric can be applied because each score is labeled as relevant or non-relevant. The bootstrap takes samples from the relevant scores and non-relevant scores with replacement, respectively.

For Eq. (6), when each sample ($S$) is a document collection, we can get a data matrix as shown in Table 3, where $\hat{f}_i$ and $dc_j$ are the $i$th system and the $j$th document collection sample, respectively, $y_i$ represents the effectiveness score variable of the $\hat{f}_i$ on document collections for a given topic, and $y_{i,j}$ (e.g., the value of $AP$ or $ERR@20$) is the effectiveness score of $j$th document collection on $\hat{f}_i$. The effectiveness score of $j$th document collection of the target system is $y_{T,j}$ ($j=(1,2,\dots,m)$), where $y_{T,j}=max(y_{1,j},y_{2,j},\dots,y_{i,j},\dots,y_{m,j})$.

Similar with the analysis the Eq. (7) in Section 3.3, the per-topic bias–variance evaluation is formulated as follows:

$$\begin{aligned} E(y_i-c)^2 &= E(y_i-\overline{y_i}+\overline{y_i}-c)^2 \\ &= E(\overline{y_i}-c)^2+E(y_i-\overline{y_i})^2 \\ &= Bias^2(y_i)+Var(y_i) \end{aligned} \tag{12}$$

Just like the situation where the document collection is fixed, the target model ($f$) is devised with the best performance for each document collection among all the considered systems. As Table 3 shows, $y_{T,j}=max(y_{1,j},y_{2,j},\dots,y_{i,j},\dots,y_{m,j})$ and the performance of target system $f$ is the value of $c$ ($c=\overline{y_T}$).

$E(y_i-c)^2$ shows the system's overall performance for a given topic on all the different document collection samples through the method of bootstrap. Similar to Eq. (7), the smaller $Bias(y_i)$ and $Var(y_i)$ reflect the better effectiveness and stability on a given topic. Naturally, the relation between per-topic bias and variance can be studied.

Considering the per-topic bias and variance for all topics, a direct way is to average per-topic bias and variance across topics, which helps us to understand the averaged per-topic effectiveness and stability on all topics. The averaged per-topic bias ($\overline{Bias^2}$) and variance ($\overline{Var}$) can be computed as follows:

$$\overline{Bias^2}=\frac{1}{n}\sum_{j=1}^{n}(Bias^2(y_i))_j \tag{13}$$

$$\overline{Var}=\frac{1}{n}\sum_{j=1}^{n}(Var(y_i))_j \tag{14}$$

where $(Bias^2(y_i))_j$ and $(Var(y_i))_j$ represent the squared bias and variance of the $i$th system of the $j$th topic, respectively.

## 4. The comparison between risk-sensitive evaluation and bias–variance evaluation

In this section, our goal is to compare the risk-sensitive evaluation and bias–variance evaluation. As Table 4 shows, we will compare their differences from four aspects, i.e., baseline, effectiveness, stability and tradeoff in Section 4.1. Then, in Section 4.2, we investigate the instability issue in-depth by defining a variable of the effectiveness difference.

### 4.1. Risk-sensitive evaluation metrics

#### 4.1.1. Baseline

Firstly, the baseline of $U_{Risk}$ is a single baseline, which could be a weak baseline. Different single baselines also leads to different risk values. Secondly, the baseline of $Z_{Risk}$ or $GeoRisk$ is that the average performance of all systems except for the system under

**Table 4**
The comparison between the bias–variance evaluation metric with other related metrics.

| Metric | Baseline | Effectiveness | Stability | Tradeoff |
|---|---|---|---|---|
| $U_{Risk}$ | Single | Yes | Risk | $\alpha$ |
| $Z_{Risk}$ | $Multiple^1$ | None | Risk | $\alpha$ |
| $GeoRisk$ | $Multiple^1$ | Yes | Risk | $\alpha$ |
| $Bias^2(x_i) + Var(x_i)$ | $Multiple^2$ | Yes | Variance | Correlation |

test, i.e., $Multiple^1$ in Table 4. However, it still change with the system to be tested. The baseline of our framework is the target system with the best performance over all the systems for each topic, and we denote such a baseline as $Multiple^2$ in Table 4. It should be noted that our target system will not remove the system under test to ensure that the baseline will not affect the evaluation process of the model.

#### 4.1.2. Effectiveness

In Appendix, we derive that when $\alpha = 0$, $U_{Risk}$ calculates the effectiveness score of the current system, which corresponds to the bias in our framework. For $Z_{Risk}$, since the statistics have been normalized, $Z_{Risk}$ cannot reflect the effectiveness of the retrieval system. For $GeoRisk$, it can reflect the system effectiveness score. However, one cannot derive the mean effectiveness score based on a particular setting or parameter in $GeoRisk$, as we can do so in $U_{Risk}$ and in bias–variance framework.

#### 4.1.3. Stability

For $U_{Risk}$, $Z_{Risk}$ and $GeoRisk$, the risk captures a related negative aspect of retrieval performance across topics and the reward captures the corresponding positive aspect, and the sum of risk and reward reflect the volatility of $\hat{f}_i$ relative to the baseline over all topics (Collins-Thompson, 2009a). However, in bias–variance framework, the stability of the system is reflected by the variance of the effectiveness scores.

#### 4.1.4. Tradeoff

The BV-Test framework is a top-down design, which explores the inherent relation between bias and variance based on the bias–variance decomposition theory. However, $U_{Risk}$, $Z_{Risk}$ and $GeoRisk$ are risk-sensitive evaluation metrics designed from the bottom-up manner. It is a combination metric from the component sub-metrics.

Meanwhile, in our work, the effectiveness–stability tradeoff is quantified by the correlation between bias and variance, and we test it in Section 5.3. For $U_{Risk}$, $Z_{Risk}$ and $GeoRisk$, the tradeoff exists between the risk and reward, which is controlled artificially by the parameter $\alpha$ (see Appendix for details).

In addition, we will explore how to reduce the bias–variance tradeoff. In Section 5.4, we find that when additional data is used for query reformulation, bot bias and variance can be reduced simultaneously, indicating a better overall performance.

### 4.2. Investigation of the instability of risk-sensitive evaluation and bias–variance evaluation

Now we study the instability issue in-depth by comparing risk-sensitive evaluation and our bias–variance evaluation. We will show the advantage of bias–variance metrics in terms of the traceability of the instability.

For risk-sensitive metrics $U_{Risk}$, $Z_{Risk}$ and $GeoRisk$ (detailed in Appendix), a very basic statistics corresponds to the effectiveness difference between the current effectiveness score and the baseline (or a normalized baseline), i.e., $x_{i,q} - b_{i,q}$ in $U_{Risk}$, or $x_{i,j} - e_{i,j}$ in $Z_{Risk}$ (and $GeoRisk$). The instability in these risk-sensitive measures are corresponding to the positive/negative distribution of such a statistic, which can be reflected by the variance of such a statistic. However, an explicit variance of this statistic has not been defined before. In this section, we are going to study in-depth such a variance by decomposing it into several parts, through which we can clearly observe which kinds of variance/instability are considered. In addition, we can measure which kinds of variance are the main reason/source of the instability of a retrieval system.

To be consistent with the baseline system in our BV-Test framework, we use the target system as the baseline system, and then define:

$$\rho_{i,j} = x_{T,j} - x_{i,j} \tag{15}$$

where $\rho_{i,j}$ reflects the effectiveness difference between the target system $f$ and current system $\hat{f}_i$ for the topic $t_j$. We then have a variable $\rho_i$ that represents all the effectiveness score differences for all the topics. The corresponding variance based on such a

**Table 5**
Datasets and topics used for Ad-hoc Task.

| | Datasets | Topics |
|---|---|---|
| 1993(TREC-2) | disk1&2 | 101-150 |
| 1994(TREC-3) | disk1&2 | 151-200 |
| 1995(TREC-4) | disk2&3 | 201-250 |
| 1996(TREC-5) | disk2&4 | 251-300 |
| 1997(TREC-6) | disk4&5 | 301-350 |
| 1998(TREC-7) | disk4&5 | 351-400 |
| 1999(TREC-8) | disk4&5 | 401-450 |

**Table 6**
Datasets and topics used for Web Track.

| | Datasets | Topics |
|---|---|---|
| WebTrack2010 | ClueWeb09 | 51-100 |
| WebTrack2011 | ClueWeb09 | 101-150 |
| WebTrack2012 | ClueWeb09 | 151-200 |
| WebTrack2013 | ClueWeb12 | 201-250 |
| WebTrack2014 | ClueWeb12 | 251-300 |

**Table 7**
Datasets and the number of sessions used for Session Track.

| | Datasets | The number of sessions |
|---|---|---|
| SessionTrack2010 | ClueWeb09 | 150 |
| SessionTrack2011 | ClueWeb09 | 76 |
| SessionTrack2012 | ClueWeb09 | 78 |

variable $\rho_i$ is as follows:

$$\begin{aligned} & Var(\rho_i) \\ =& \frac{1}{n} * \sum_{j=1}^{n} (\rho_{i,j} - \overline{\rho_i})^2 \\ =& \frac{1}{n} * \sum_{j=1}^{n} [(x_{T,j} - x_{i,j}) - \frac{1}{n}\sum_{j=1}^{n} x_{T,j} + \frac{1}{n}\sum_{j=1} x_{i,j}]^2 \\ =& \frac{1}{n} * \sum_{j=1}^{n} [(x_{T,j} - \overline{x_T}) - (x_{i,j} - \overline{x_i})]^2 \\ =& Var(x_T) + Var(x_i) - 2 * Cov(x_T, x_i) \end{aligned} \tag{16}$$

where $Var(x_i)$ and $Var(x_T)$ measure the stability of the effectiveness scores across topics for the current system $\hat{f}_i$ and target system $f$, respectively. Meanwhile, $Cov(x_T, x_i)$ represents the score covariance of $f$ and $\hat{f}_i$. As shown in Eq. (16), we can find that if $f$ and $\hat{f}_i$ have the same score distribution on each topic, the $Cov(x_T, x_i) = Var(x_i) = Var(x_T)$. In other words, the covariance reflects the inconsistency of two score distributions of $f$ and $\hat{f}_i$, respectively. By decomposing $\rho_i$ into three parts $Var(x_i)$, $Var(x_T)$ and $Cov(x_T, x_i)$, we can find the main factor that cause the system instability according to their values. The empirical evidence will be reported in Section 5.2.

## 5. Experiments

In this section, we verify our analysis of bias–variance evaluation metric through experiments. In Section 5.1, we introduce the database information needed for the experiments. In Section 5.2, compared with the risk-sensitive metric, we study the advantages of bias–variance evaluation metric in the traceability of instability. In Section 5.3, we systematically study the tradeoff between the $Bias^2$ and $Var$ across topics and on per-topic. In Section 5.4, the Session track are used to measure the effect of query reformulation on $Bias^2$ and $Var$ and the relation between them.

### 5.1. Datasets

We carry out our bias–variance evaluation on three TREC tracks, i.e., Ad-hoc, Web and Session track. For each track, we evaluate all the systems/runs submitted to TREC for several years, where the runs/systems represent the results of the participants' submission in their raw form. For document retrieval tasks, each result file is a list of the 1000 documents retrieved for each topic in order of decreasing similarity with the query. The data information for these three tracks are listed in Tables 5–7, respectively. The metrics used in Ad-hoc, Web and Session tracks are AP, ERR@20 and NDCG@10, separately.

**Table 8**
The comparison of the instability of risk-sensitive metric and the variance.

| systemmetric | $\|Z_{Risk}\| \downarrow$ | $Var(\rho) \downarrow$ | $Var(x_i) \downarrow$ | $Cov(x_T, x_i) \uparrow$ |
|---|---|---|---|---|
| uogTrA44xu | 1.7366 | 0.0826 | 0.1065 | 0.0747 |
| srchvrs12c00 | 2.6686 | 0.1113 | 0.1316 | 0.0729 |
| DFalah121A | 2.3990 | 0.0829 | 0.1066 | 0.0746 |
| QUTparaBline | 2.1774 | 0.0811 | 0.0915 | 0.0679 |
| utw2012c1 | 1.8495 | 0.1003 | 0.0738 | 0.0495 |
| ICTNET12ADR2 | 2.0159 | 0.0988 | 0.0672 | 0.0469 |
| autoSTA | 2.3491 | 0.1088 | 0.0618 | 0.0381 |
| irra12c | 1.5193 | 0.0999 | 0.0490 | 0.0372 |

### 5.2. An in-depth study of the tracebility of instability

*Experiment Setting:* To verify our analysis in Section 4.2, we have selected the top 8 systems of WebTrack2012, according to the mean effectiveness score based on $ERR@20$. Meanwhile, setting the risk parameter of $Z_{Risk}$ to 0 and baseline to the target system ($f$). The experimental hypothesis is as follow:

- $H_1$: Based on the BV-Test framework, the variance of the statistical value $\rho$ can be decomposed to track the source of system instability.

Table 8 presents the results of instability of risk-sensitive metric and the variance, where the smaller value of $|Z_{Risk}|$ (or $Var(\rho)$) is, the lower instability of current system will be (Dinçer et al., 2016). Based on $|Z_{Risk}|$ or $Var(\rho)$, we can observe that srchvrs12c00 is the system with the highest risk among uogTrA44xu, srchvrs12c00 and DFalah121 A. However, for $Z_{Risk}$, it cannot further explore the source of instability, but the decomposition of our variance ($Var(\rho)$) can trace the instability. Specifically, the $Var(x_i)$ of the srchvrs12c00 is 0.1316, which is larger than the variance of uogTrA44xu and DFalah121 A. Meanwhile, the $Cov(x_T, x_i)$ of uogTrA44xu, srchvrs12c00 and DFalah121 A is almost the same. Therefore, compared with uogTrA44xu and DFalah121 A, we can observe that the instability of srchvrs12c00 mainly comes from the variance of the current system (i.e., $Var(x_i)$).

For ICTNET12ADR2 and autoSTA, according to $Z_{Risk}$ and $Var(\rho)$, ICTNET12ADR2 has the lower risk than autoSTA. However, for $Var(x_i)$, the variance of ICTNET12ADR2 is larger than that of autoSTA. Therefore, $Var(x_i)$ cannot account for the overall instability comparison between ICTNET12ADR2 and autoSTA, as evidenced by $Z_{Risk}$ and $Var(\rho)$. On the other hand, for the covariance of them, the covariance of autoSTA is much smaller than that of ICTNET12ADR2, indicating that the inconsistency of score distributions (between the current system and the target system) of autoSTA is higher than that of ICTNET12ADR2. In this perspective, such an inconsistency leads to the instability of autoSTA based on $Var(\rho)$. Recall that $\rho$ stands for the effectiveness difference between the current system and the target system. Therefore, compared with ICTNET12ADR2, the instability of autoSTA mainly comes from the $Cov(x_T, x_i)$). The above results support the hypothesis $H_1$.

### 5.3. Bias–variance evaluation

In this section, we study $Bias^2$ and $Var$ across topics and on per-topic. Firstly, in Section 5.3.1, we study the relation between $Bias^2$ and $Var$ across topics. Then, in Section 5.3.2, we will not only study the effect of the topic grouping but also the score normalization on the bias–variance evaluation. After that, in Section 5.3.3, we analyze $Bias^2$ and $Var$ on per-topic. The experimental hypotheses are as follows:

- $H_2$: Bias–variance tradeoff exists in both across topic and on per-topic.
- $H_3$: After the topic grouping, the bias–variance tradeoff will be less obvious.
- $H_4$: After the max–min normalization, the bias–variance tradeoff will be alleviated.

#### 5.3.1. Bias–variance evaluation across topics

*Experiment Setting:* This section's experiment is based on the bias–variance evaluation across topics in Section 3.3. The datasets are the Ad-hoc Track(1993–1999) and the WebTrack (2010–2014). The baseline is the target system, which is composed of the achievable best performance of every topic. We choose the Pearson correlation coefficient to quantify the relation between $Bias^2$ and $Var$.

By observing the first column of Ad-hoc Table 9, we can see that almost all the correlation coefficients are strongly negative when the metric AP is used before the score normalization, which indicates that there is a noticeable tradeoff between $Bias^2$ (indicating the effectiveness of a system) and $Var$ (signifying the stability of a system).

To visualize this correlation, the data on Ad-hoc 1995 is taken as an example. In Fig. 1, it shows that as $Bias^2$ increases, $Var$ is decreasing, and we observe a clear tradeoff between $Bias^2$ and $Var$. The above results support the hypothesis $H_2$.

**Table 9**
The Pearson correlation between $Bias^2$ and $Var$ on Ad-Hoc and Web Track. The number is **bold** when the absolute value is larger than 0.8.

| | Original Data | | Random Partition | | Query Difficulty | |
|---|---|---|---|---|---|---|
| | $AP$ | $AP'$ | $MAP$ | $MAP'$ | $MAP$ | $MAP'$ |
| Ad-Hoc1993 | **−0.8732** | −0.3941 | −0.7378 | 0.0324 | **−0.8963** | −0.5032 |
| Ad-Hoc1994 | **−0.8640** | −0.5643 | −0.5044 | −0.4159 | **−0.8283** | −0.3980 |
| Ad-Hoc1995 | **−0.9376** | **−0.8490** | −0.7202 | −0.6082 | **−0.8897** | −0.7841 |
| Ad-Hoc1996 | **−0.8792** | **−0.8141** | −0.5922 | −0.5843 | −0.7875 | −0.4884 |
| Ad-Hoc1997 | **−0.9139** | **−0.8168** | −0.6727 | −0.5082 | **−0.8565** | −0.6157 |
| Ad-Hoc1998 | **−0.8981** | −0.7883 | −0.7012 | −0.4967 | **−0.8426** | −0.6317 |
| Ad-Hoc1999 | **−0.9162** | **−0.9376** | **−0.8256** | −0.7035 | **−0.8823** | −0.7048 |
| | $ERR@20$ | $ERR@20'$ | $(M)ERR@20$ | $(M)ERR@20'$ | $(M)ERR@20$ | $(M)ERR@20'$ |
| WebTrack2010 | −0.7981 | **−0.8083** | −0.5218 | −0.2538 | −0.7389 | −0.4124 |
| WebTrack2011 | −0.7687 | −0.7258 | −0.7785 | 0.0095 | −0.7918 | 0.2621 |
| WebTrack2012 | **−0.9509** | **−0.8772** | −0.4019 | −0.1866 | **−0.9239** | −0.7849 |
| WebTrack2013 | −0.7905 | **−0.8022** | −0.4982 | −0.1979 | −0.6976 | −0.2790 |
| WebTrack2014 | −0.6546 | −0.6654 | −0.3870 | −0.5921 | −0.5814 | −0.0535 |

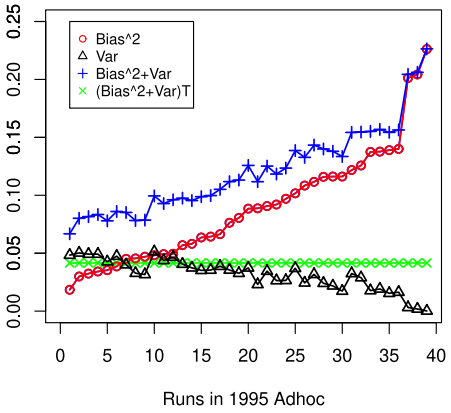


**Fig. 1.** $Bias^2$, $Var$ and $Bias^2 + Var$ of $AP$ on Ad-hoc 1995, where all the systems are sorted in an ascending order according to $Bias^2$.

### 5.3.2. Impact on tradeoff by topic grouping and normalization

*Experiment Setting:* In this section, we are going to study the effect of topic group and normalization on system evaluation by comparing the tradeoff with a number of hypotheses, drawing on the analysis in Section 3.3. In Table 9, $AP/MAP$ is the original data and $AP'/MAP'$ is the normalized scores by the max–min normalization, and the random partition and query difficulty are two topic grouping methods detailed later.

*The results of before and after topic grouping.* Two strategies are adopted for grouping topics. The first is the random grouping. 10 topics are randomly extracted from the entire topic as a group and 50 topic groups are used. Then, $Bias^2$ and $Var$ are computed for each group. In this way, the results are uncertain because of the random partition of topic set. In order to solve this issue, this process is repeated for 1000 times. Because of the uncertainty of random grouping partition, we use another way of grouping, which is based on the topic difficulty. In our experiment, the topic difficulty is measured by the best metric value (e.g., the maximum AP value) of a given topic across all the participated systems in a track task. The lower the best metric value is, the more difficult the corresponding topic is. We rank all the topics based on the topic difficulty degree and group them into several subsets. Each subset includes 5 topics. Since the partition is fixed, the results will not be random.

For Ad-hoc1995 in Table 9, the absolute value of the coefficient becomes less strong after the random partition, i.e., $|-0.7202| < |-0.9376|$ and $|-0.6082| < |-.0.8490|$. Meanwhile, the tradeoff becomes small for the data after topic grouping based on query difficulty , i.e., $|-0.8897| < |-0.9376|$ and $|-0.7841| < |-0.8490|$. Similarly, we can find the same phenomenons exist in the remaining row data, which supports the hypothesis $H_3$.

*The results of before and after max–min normalization.* In Table 9, $AP/MAP$ is the original data and $AP'/MAP'$ is the normalized scores by the max–min normalization. As we described in Lemma 1, the normalization is designed to make the system mainly effect performance variability rather than topic difficulty. As observed in Table 9, for the Ad-hoc1995, we can find that the absolute value of correlation after max–min normalization become smaller (e.g., $|-0.8490| < |-0.9376|$, $|-0.6082| < |-0.7202|$, $|-0.8897| < |-0.7841|$). Similarly, the same phenomenons happen to the remaining data, which supports the hypothesis $H_4$.

**Table 10**
The Pearson correlation coefficients of averaged $Bias^2$ and $Var$ of all the runs on Ad-Hoc 1993~1999 across topics. The number is **bold** when the absolute value is larger than 0.8.

| year | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 |
|---|---|---|---|---|---|---|---|
| Pearson | **−0.9809** | **−0.9687** | **−0.8762** | −0.7826 | **−0.8939** | **−0.9591** | **−0.9370** |

**Table 11**
The Pearson correlation coefficients of $Bias^2$ and $Var$ for each topic on Ad-hoc 1995. The number is **bold** when the absolute value is larger than 0.8.

| Topic | 202 | 203 | 204 | 205 | 206 | 207 | 208 |
|---|---|---|---|---|---|---|---|
| Pearson | **−0.9562** | **−0.9587** | **−0.9602** | **−0.9574** | **−0.9506** | **−0.9018** | **−0.9687** |
| Topic | 209 | 210 | 211 | 212 | 213 | 214 | 215 |
| Pearson | **−0.9515** | −0.7874 | **−0.9671** | **−0.9477** | **−0.9555** | **−0.9500** | **−0.9317** |
| Topic | 216 | 217 | 218 | 219 | 220 | 221 | 222 |
| Pearson | **−0.9516** | **−0.9658** | **−0.9613** | **−0.9612** | −0.7796 | **−0.9531** | **−0.8111** |
| Topic | 223 | 224 | 225 | 226 | 227 | 228 | 229 |
| Pearson | **−0.9684** | **−0.9481** | **−0.8940** | **−0.9687** | **−0.9638** | **−0.9316** | **−0.9087** |
| Topic | 230 | 231 | 232 | 233 | 234 | 235 | 236 |
| Pearson | **−0.9507** | **−0.9250** | **−0.9680** | **−0.9540** | **−0.8984** | **−0.9074** | **−0.8796** |
| Topic | 237 | 238 | 239 | 240 | 241 | 242 | 243 |
| Pearson | **−0.9685** | **−0.9738** | **−0.9522** | **−0.9865** | **−0.9769** | **−0.9330** | **−0.9589** |
| Topic | 244 | 245 | 246 | 247 | 248 | 249 | 250 |
| Pearson | **−0.9141** | **−0.9612** | **−0.9693** | **−0.9237** | **−0.9809** | **−0.9426** | **−0.9295** |

#### 5.3.3. Bias–variance evaluation on per-topic

*Experiment Setting:* This section's experiment is based on the per-topic bias–variance evaluation in Section 3.4. The dataset uses the Ad-hoc Track(1993–1999), for Ad-hoc track in one year, there are $m$ runs/systems and $n$ topics. There is one metric ($M$) value if there is only one document collection for each topic. As a result, there are $k$ values if there are $k$ document collection samples. Totally, these $k \times m \times n$ values are utilized to explore the per-topic quality. In Robertson and Kanoulas (2012), the range of $k$ is $[10, 1000]$. We find that the experimental results are unstable when $k = 10$, while the experimental results tend to be stable when $k > 100$. In the following experiment, in order to ensure the stability of the experimental results, we set $k = 100$. the evaluation metric $M$ is AP and the target system is composed of the achievable best performance.

The correlation coefficients between the $\overline{Var}$ in Eq. (13) and the $\overline{Bias^2}$ in Eq. (14) are given in Table 10. The tradeoff between the averaged per-topic effectiveness and variance still exists when the document collection is a sample evidenced by the numbers in this Table.

In addition to the averaged $Bias^2$ and $Var$ over topics, the $Bias^2$ and $Var$ for each topic are also analyzed. There are only 49 useful topics used in Ad-hoc 1995 because the topic 201 retrieved no relevant documents. Table 11 shows Pearson correlation coefficients of $Bias^2$ and $Var$ for each topic on Ad-hoc 1995. Except for the correlation coefficients on topic 210 and 220, all the other coefficients are strongly negative signifying that there is a obvious tradeoff on effectiveness and stability for a given topic. If the $Bias^2$ of a system on a topic is smaller (the better effectiveness), the $Var$ then is larger (the worse stability). The above results support the hypothesis $H_2$.

### 5.4. The impact of relevance feature on effectiveness and stability

In this section, we study the query reformulation's effect on the effectiveness and stability of a system and the relation between them on the Session track task. The goal of this experiment is to test whether we can improve *both the effectiveness and stability* by using previous queries and user interactions with a retrieval system. The primary evaluation measurement used in Session track is normalized discounted cumulative gain at 10 documents (denoted as NDCG@10).

In machine learning, with the increase of external data, the performance of the model will gradually approach the target system, and its bias and variance will gradually decrease simultaneously (Belkin, Hsu, Ma, & Mandal, 2018; Briscoe & Feldman, 2011; Geman et al., 1992a). In information retrieval, the query reformatting involves more user information, which can be regarded as the external data (Kanoulas, Carterette, Clough, & Sanderson, 2010). We assume that by adding more user data, the values of bias and variance metrics can be reduced simultaneously. We will conduct a series of experiments on Session track to verify such a hypothesis, which is stated formally as follows.

- $H_5$: In the query reformatting task, by adding more external user data, the bias and variance can be reduced simultaneously.

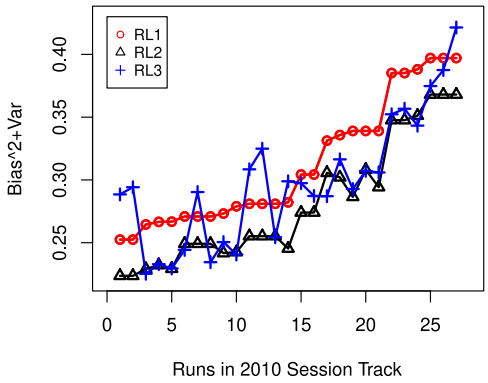


**Fig. 2.** The $Bias^2 + Var$ on Session track 2010.

**Table 12**
The Pearson correlation coefficients between $Bias^2$ and $Var$ on Session track 2010, 2011 and 2012. The number is **bold** when its absolute value is larger than 0.8.

| | 2010 | | | 2011 | | | | 2012 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | RL1 | RL2 | RL3 | RL1' | RL2' | RL3' | RL4' | RL1' | RL2' | RL3' | RL4' |
| $NDCG@10$ | **−0.9706** | **−0.8835** | **−0.9181** | **−0.8443** | **−0.8160** | **−0.9042** | **−0.8703** | **−0.8374** | **−0.9184** | **−0.8869** | −0.7998 |
| $NDCG@10'$ | **−0.9743** | **−0.9136** | **−0.9486** | −0.5416 | −0.6883 | **−0.8522** | **−0.8350** | **−0.8409** | **−0.9125** | **−0.9038** | −0.7514 |

### 5.4.1. Bias–variance analysis of session track 2010

*Experiment Setting:* In Session track 2010, each participant submitted three ranked lists of documents for three experimental conditions. In Table 12, RL1 refers to the ranked lists over the initial query. RL2 refers to the ranked lists over the query reformulation, ignoring the initial query. RL3 refers to the ranked list over the initial query and query reformation (Kanoulas et al., 2010). Here we set the target model as a virtual model which has the maximum $NDCG@10$ value on each session/query.

In Fig. 2, there is no systems performing best on condition RL1; 19 of 27 systems perform best on condition RL2; 8 on condition RL3. For $NDCG@10'$, 20 of 27 systems perform best on condition RL2; 7 on RL3, 0 on RL1. The evaluation metric here uses $NDCG@10$, and $NDCG@10'$ is the result of max–min normalization. That illustrates the query reformulation benefits the performance of system, although some systems are even worse after using the query reformulation.

In the first column of Table 12, the relation between $Bias^2$ and $Var$ is analyzed in different conditions on Session track 2010. There are negative coefficients between $Bias^2$ and $Var$ indicates that there is the tradeoff between them. The coefficients are stronger in RL1, and weaker in RL3 and RL2. The reasons could be that a part of systems achieve both the improvement of effectiveness and stability, which weakens this kind of negative correlation between $Bias^2$ and $Var$. The tradeoff can be relieved by query reformulation.

### 5.4.2. Bias–variance analysis of session track 2011 and 2012

*Experiment Setting:* In Session track 2011 and 2012, each participant submitted four ranked lists of documents over four experimental conditions. RL1' refers to the ranked lists on the current query. RL2' is the ranked lists when using the set of past queries in the session. RL3' is the ranked list when making use of more user data including the historical queries (prior to the current one) along with the ranked lists of URLs and the corresponding web pages. RL4' is the ranked list when one makes full use of the user data (i.e., the historical queries, the ranked lists of URLs and the corresponding web pages and the clicked URLs and the time spent on the corresponding web pages) (Kanoulas, Carterette, Hall, Clough, & Sanderson, 2011).

In Fig. 3, 22 of 34 systems perform better on RL2'∼RL4' compared with RL1'. It is apparent that a certain number of systems perform best on condition RL2'∼RL4', which contains the full of user information. This does indicate system can use the increasing amounts of information prior to a query to improve performance.

In Table 12, $Bias^2$ and $Var$ have the negative relation in Session track 2011 and 2012. It shows that the tradeoff between $Bias^2$ and $Var$. Compared with RL1', RL2', and RL3', the tradeoff on RL4' is reduced to varying degrees. We can expect that the full use of information can improves the overall performance of the system and alleviates the tradeoff.

## 6. Conclusions and future work

In this paper, we have proposed a unified framework to systematically analyze the effectiveness and stability of IR systems as well as the relation between them. It takes inspiration from the mean squared error in machine learning, in which the error can naturally be decomposed into bias and variance. We apply the above evaluation framework to information retrieval, and the performance of a current system can be measured by the error between it and the target system, and the effectiveness and stability can be mapped

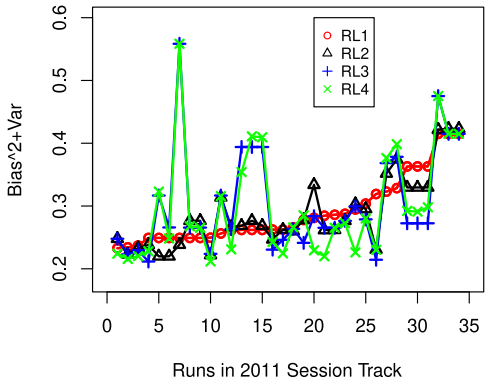


**Fig. 3.** The $Bias^2 + Var$ on Session track 2011.

to bias and variance respectively. During the process of decomposition, an intrinsic relation between bias and variance, namely bias–variance tradeoff, is modeled to study the tradeoff between retrieval effectiveness and stability. Experimental results show that there is a clear effectiveness–stability tradeoff across topics and per-topic.

In addition, we explored the factors that affect the bias–variance evaluation. Based on the experimental results, we found that topic grouping and max–min normalization can reduce the effectiveness–stability tradeoff.

Meanwhile, we systematically compared risk-sensitive metric and bias–variance evaluation metric from four aspects: baseline, effectiveness, stability, and tradeoff, illustrating the advantages of our work. We define a statistic which computes the effectiveness score difference between the current system and the target system. The variance of this statistic can be decomposed into three parts: the instability of the current system, the instability of the target system, and the instability measured by the score covariance (see Eq. (16)). Each of these three parts corresponds to a factor of system instability. The experimental results show that based on such a decomposition, we can find the causes/factors of system instability.

Finally, we apply our framework to evaluate the participating systems in Session track (2010–2012), aiming to study the effect of query reformulation on a system's overall performance and the relation between $Bias^2$ and Var.

In the future, we will further explore the applicability of our framework in other retrieval tasks (Emadi, Tanha, Shiri, & Aghdam, 2021; Neural embedding-based specificity metrics for pre-retrieval query performance prediction, 2020). Meanwhile, that hypothesis testing (paired $t$-test, etc.) has been widely used in the field of IR (Dinçer et al., 2014), and its purpose is to study whether the improvement of the current system (over the baseline) is stable. The variance in our work studies whether the performance of the current system itself is stable across different topics. In the future work, we will study the detailed connection between our BV-Test framework and traditional hypothesis test (such as paired $t$-test).

## CRediT authorship contribution statement

**Peng Zhang:** Methodology, Conceptualization, Writing – review & editing, Writing – original draft. **Hui Gao:** Investigation, Software, Writing – original draft, Writing – review & editing. **Zeting Hu:** Experiment, Data curation, Writing – original draft. **Meng Yang:** Experiment, Data curation, Writing – original draft. **Dawei Song:** Supervision, Writing – original draft, Conceptualization. **Jun Wang:** Supervision, Conceptualization. **Yuexian Hou:** Supervision, Conceptualization. **Bin Hu:** Supervision.

## Acknowledgments

This work is supported in part by the state key development program of China (grant No. 2018YFC0831704, 2017YFE0111900) and the European Unions Horizon 2020 research and innovation programe under the Marie SkodowskaCurie grant agreement No. 721321.

## Appendix. Risk-sensitive evaluation metrics

To illustrate the $U_{Risk}$, $Z_{Risk}$ and $GeoRisk$, we use the notation in Table 1. Let a row vector $X_{i\cdot} = [x_{i,1}, x_{i,2}, \dots, x_{i,j}, \dots, x_{i,n}]$, where $x_{i,j}$ is the effectiveness score of $f_i$ on the topic $t_j$. Let $Sum(X_{i\cdot}) = x_{i,1} + x_{i,2} + \cdots + x_{i,j} + \cdots + x_{i,n}$, which means the sum effectiveness scores of $\hat{f}_i$ over all topics.

### A.1. $U_{Risk}$ (Dinçer et al., 2014; Wang et al., 2012)

$U_{Risk}$ is a risk function which can reflect the information retrieval performance over a set of topics $Q$, composed of $n$ topics. $U_{Risk}$ is defined as follows:

$$U_{Risk} = \frac{1}{n} * (\sum_{q \in Q_+} \delta_q + \sum_{q \in Q_-} \delta_q) + \frac{1}{n} * \alpha * \sum_{q \in Q_-} \delta_q \tag{17}$$

where $n = |Q|$, $\delta_q = x_{i,q} - b_{i,q}$, $x_{i,q}$ and $b_{i,q}$ are the effectiveness score of the current system $\hat{f}_i$ and the score of the baseline system $b$ on topic $q$, respectively. $x_{i,q}$ corresponds to $x_{i,j}$ in Table 1. Therefore, $\delta_q$ represents the effectiveness difference between $\hat{f}_i$ and $b$ for a given topic $q$. The set of topic $Q$ can be divided into two sets ($Q_+$ and $Q_-$) according to the value of $\delta_q$. For a topic $q$, if $\delta_q > 0$, then $q \in Q_+$, meaning that there is a win, which represents that the current system $\hat{f}_i$ performs better than the baseline $b$, i.e., $x_{i,q} > b_{i,q}$. For a topic $q \in Q_-$, the loss (i.e., $\delta_q < 0$) means the opposite situation, i.e., $x_{i,q} \leq b_{i,q}$. The total wins for all the topics $q \in Q_+$ represent the reward and the total losses for $q \in Q_-$ represent the risk. $U_{Risk}$ adjusts the tradeoff between risk and reward by controlling the risk parameter $\alpha$.

When $\alpha = 0$, $U_{Risk}$ is equal to $\frac{1}{n} * (\sum_{q \in Q_+} \delta_q + \sum_{q \in Q_-} \delta_q) = \frac{1}{n} * \sum_{q \in Q} \delta_q = \frac{1}{n} * \sum_{q \in Q} (x_{i,q} - b_{i,q})$. If the effectiveness metric is $AP$, $U_{Risk}$ is actually the $MAP$ difference between the current system and the baseline. In this case, $U_{Risk}$ reflects the effectiveness of the current system, which corresponds to the bias in our framework. However, if $\alpha = 0$, there is no variance in $U_{Risk}$, so that the relation between bias and variance cannot be studied.

### A.2. $Z_{Risk}$ (Dinçer et al., 2016)

$Z_{Risk}$ was proposed by combining the idea of $U_{Risk}$ with Chi-square test, which is defined as follow:

$$Z_{Risk} = [\sum_{q \in Q_+} z_{i,q} + (1 + \alpha) * \sum_{q \in Q_-} z_{i,q}] \tag{18}$$

when $z_{i,j} \leq 0$, $q \in Q_-$. Conversely, $q \in Q_+$. The statistic $z_{i,j}$ is defined as:

$$z_{i,j} = \frac{x_{i,j} - e_{i,j}}{\sqrt{e_{i,j}}} \tag{19}$$

where $e_{i,j}$ is the expectation value of $x_{i,j}$. In Eq. (18), the distribution of $z_{i,q}$ values on the population can be approximated by the standard normal distribution with zero mean and unit variance (Dinçer et al., 2016). As $z_{i,j}$ has been normalized, $Z_{Risk}$ cannot reflect the effectiveness of a retrieval system $\hat{f}_i$

### A.3. GeoRisk (Dinçer et al., 2016)

$GeoRisk$ was proposed as follows:

$$GeoRisk = \sqrt{(Sum(X_{i\cdot})/n) * \Phi(Z_{Risk}/n)} \tag{20}$$

where $Sum(X_{i\cdot})/n$ in Eq. (20) is equal to the mean effectiveness score of $\hat{f}_i$, which can reflect the effectiveness of a system. However, we cannot derive the mean effectiveness score based on a particular setting or parameter in $GeoRisk$, as we can do so in $U_{Risk}$ and in bias–variance framework.

## References


(2019). A survey on evaluation of summarization methods. *Information Processing & Management*, *56*(5), 1794–1814.

Amati, G., Carpineto, C., & Romano, G. (2004). Query difficulty, robustness, and selective application of query expansion. *Lecture Notes in Computer Science*, *2997*, 127–137.

(2017). An in-depth study on diversity evaluation: The importance of intrinsic diversity. *Information Processing & Management*, *53*(4), 799–813.

Bah, A., & Carterette, B. (2015). Improving ranking and robustness of search systems by exploiting the popularity of documents. In *AIRS* (pp. 174–187).

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2018). Reconciling modern machine learning and the biasvariance trade-off. ArXiv e-prints.

Bishop, C. M. (2006). *Pattern recognition and machine learning (Information science and statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc..

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*(1), 2–16.

Carmel, D., Farchi, E., Petruschka, Y., & Soffer, A. (2002). Automatic query wefinement using lexical affinities with maximal information gain. In *Proc. of ACM SIGIR* (pp. 283–290).

Collins-Thompson, K. (2009). Accounting for stability of retrieval algorithms using risk-reward curves. In *Proc. of SIGIR* (pp. 27–28).

Collins-Thompson, K. (2009). Reducing the risk of query expansion via robust constrained optimization. In *Proc. of ACM CIKM* (pp. 837–846).

Collins-Thompson, K., Bennett, P., Diaz, F., Clarke, C. L. A., & Vorhees, E. M. (2014). TREC 2013 web track overvie. In *Proc. of TREC*.

Collins-Thompson, K., Macdonald, C., Bennett, P. N., Diaz, F., & Voorhees, E. M. (2014). TREC 2014 web track overview. In *NIST Special Publication: 500–308, Proc. of TREC*. National Institute of Standards and Technology (NIST).

Cormack, G. V., & Lynam, T. R. (2006). Statistical precision of information retrieval evaluation. In *Proc. of ACM SIGIR* (pp. 533–540).

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample selection bias correction theory. In *Proceedings of the 19th international conference on algorithmic learning theory*.

Dai, N., Shokouhi, M., & Davison, B. D. (2011). Learning to rank for freshness and relevance. In *Proc. of ACM SIGIR* (pp. 95–104).

Deveaud, R., Mothe, J., Ullah, M. Z., & Nie, J. Y. (2018). Learning to adaptively rank document retrieval system configurations. *ACM Transactions on Information Systems*, *37*(1), 1–41.

Dinçer, B. T., Macdonald, C., & Ounis, I. (2014). Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proc. of ACM SIGIR* (pp. 23–32).
Dinçer, B. T., Macdonald, C., & Ounis, I. (2016). Risk-sensitive evaluation and learning to rank using multiple baselines. In *Proc. of ACM SIGIR* (pp. 483–492).
Dinçer, B. T., Ounis, I., & Macdonald, C. (2014). Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In *ECIR* (pp. 26–38).
Domingos, P. (2000). A unifeid bias-variance decomposition and its applications. In *Proc. of ICML* (pp. 231–238).
Donna (1994). Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, *31*(3), 271–289.
Emadi, M., Tanha, J., Shiri, M. E., & Aghdam, M. H. (2021). A selection metric for semi-supervised learning based on neighborhood construction. *Information Processing & Management*, *58*(2), Article 102444.
Ferro, N., Kim, Y., & Sanderson, M. (2019). Using collection shards to study retrieval performance effect sizes. *ACM Transactions on Information Systems (TOIS)*, *37*(3), 30.
Ferro, N., & Sanderson, M. (2019). Improving the accuracy of system performance estimation by using shards. In *Proc. of ACM SIGIR* (pp. 805–814).
Geman, S., Bienenstock, E., & Doursat, R. (1992a). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.
Geman, S., Bienenstock, E., & Doursat, R. (1992b). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.
Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, *10*(6), 1425–1433.
Jain, A. K., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, *38*(12), 2270–2285.
Kanoulas, E., Carterette, B., Clough, P., & Sanderson, M. (2010). Session track 2010 overview. In *Proc. of TREC* (pp. 11).
Kanoulas, E., Carterette, B., Hall, M., Clough, P., & Sanderson, M. (2011). Session track 2011 overview. In *Proc. of TREC*.
Kharazmi, S., Scholer, F., Vallet, D., & Sanderson, M. (2016). Examining additivity and weak baselines. *ACM Transactions on Information Systems*, *34*(4), 23:1–23:18.
Lerman, L., Veshchikov, N., Markowitch, O., & Standaert, F. X. (2018). Start simple and then refine: Bias-variance decomposition as a diagnosis tool for leakage profiling. *IEEE Transactions on Computers*, 1.
Macdonald, Craig, Santos, Rodrygo, L., Ounis, & Iadh (2013). The whens and hows of learning to rank for web search. *Information Retrieval*, *16*(5), 584–628.
Mitchell, T. M., et al. (1997). *Machine learning*. McGraw-hill New York.
Moffat, A. (2013). *Seven numeric properties of effectiveness metrics*.
(2020). Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing & Management*, *57*(4), Article 102248.
Papo, Y. C. (2019). Bias-variance tradeoff in a sliding window implementation of the stochastic gradient algorithm. *arXiv: Machine Learning*.
Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *CoRR*,.
Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, *19*(1).
Qiu, X., Hu, R., & Wu, Z. (2014). Evaluation of bias-variance trade-off for commonly used post-summarizing normalization procedures in large-scale gene expression studies. *PLoS One*, *9*(6).
(2018). Ranking themes on co-word networks: Exploring the relationships among different metrics. *Information Processing & Management*, *54*(2), 203–218.
Robertson, S. E. (1981). The methodology of information retrieval experiment. *Information Retrieval Experiment*, *1*, 9–31.
Robertson, S. E., & Kanoulas, E. (2012). On per-topic variance in IR evaluation. In *Proc. of ACM SIGIR* (pp. 891–900).
de Sousa, D. X., Canuto, S. D., Gonçalves, M. A., Rosa, T. C., & Martins, W. S. (2019). Risk-sensitive learning to rank with evolutionary multi-objective feature selection. *ACM Transactions on Information Systems*, *37*(2), 24:1–24:34.
Sparck Jones, K., Walker, S., & Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*.
Suen, Y. L., Melville, P., & Mooney, R. J. (2005). Combining bias and variance reduction techniques for regression trees. In *ECML* (pp. 741–749).
Valentini, G., & Dietterich, T. G. (2004). Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *Journal of Machine Learning Research*, *5*, 725–775.
Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 316–323).
Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proc. of ACM SIGIR* (pp. 316–323).
Voorhees, E. M., Harman, D. K., et al. (2005). *vol. 63*, *TREC: Experiment and evaluation in information retrieval*. MIT press Cambridge, MA.
Wang, L., Bennett, P. N., & Collins-Thompson, K. (2012). Robust ranking models via risk-sensitive optimization. In *Proc. of ACM SIGIR* (pp. 761–770).
Williamson, P., Gamble, C., Altman, D., & Hutton, J. (2005). Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research*, *14*(5), 515.
Zhang, P., Hao, L., Song, D., Wang, J., Hou, Y., & Hu, B. (2014). Generalized bias-variance evaluation of TREC participated systems. In *Proc. of ACM CIKM* (pp. 1911–1914).
Zhang, P., Song, D., Wang, J., & Hou, Y. (2013). Bias-variance decomposition of Ir evaluation. In *Proc. ACM SIGIR* (pp. 1021–1024).
Zhang, P., Song, D., Wang, J., & Hou, Y. (2014). Bias-variance analysis in estimating true query model for information retrieval. *Information Processing & Management*, *50*(1), 199–217.
Zhao, Z., Kleinhans, A., Sandhu, G., Patel, I., & Unnikrishnan, K. P. (2019). Capsule networks with max-min normalization. *CoRR*,.
Zighelnic, L., & Kurland, O. (2008). Query-drift prevention for robust query expansion. In *Proc. of ACM SIGIR* (pp. 825–826). Association for Computing Machinery.