Bias-variance analysis in estimating true query model for information retrieval

Journal Item

Bias-Variance Analysis in Estimating True Query Model for Information Retrieval

Peng Zhang[a], Dawei Song[a,b,*], Jun Wang[c], Yuexian Hou[a]

[a] *Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, Tianjin, China*
[b] *The Computing Department, The Open University, UK*
[c] *Department of Computer Science, University College London, UK*

Abstract

The estimation of query model is an important task in language modeling (LM) approaches to information retrieval (IR). The ideal estimation is expected to be not only *effective* in terms of high mean retrieval performance over all queries, but also *stable* in terms of low variance of retrieval performance across different queries. In practice, however, improving effectiveness can sacrifice stability, and vice versa. In this paper, we propose to study this *tradeoff* from a new perspective, i.e., the bias-variance *tradeoff*, which is a fundamental theory in statistics. We formulate the notion of bias-variance regarding retrieval performance and estimation quality of query models. We then investigate several estimated query models, by analyzing when and why the bias-variance tradeoff will occur, and how the bias and variance can be reduced simultaneously. A series of experiments on four TREC collections have been conducted to systematically evaluate our bias-variance analysis. Our approach and results will potentially form an analysis framework and a novel evaluation strategy for query language modeling.

Keywords: Information Retrieval, Query Language Model, Bias-Variance

1. Introduction

Estimating query language model is an important task in language modeling (LM) approaches, since the query language model represents the underly-

*Corresponding author: Dawei Song (Email: dawei.song2010@gmail.com)

Table 1: An Example of Retrieval Effectiveness-Stability Tradeoff and Performance Bias-Variance Tradeoff

Method	A		B	
	q_1	q_2	q_1	q_2
AP	0.3	0.1	0.6	0.08
MAP	0.2		0.34	
VAP	0.01		0.0676	
$Bias$	0.25		0.11	
Var	0.01		0.0676	

ing information need and has a significant impact on retrieval performance. Ideally, the estimation should be not only effective in terms of high mean performance over all queries, but also stable in terms of low variance of performance across different individual queries. In practice, however, improving effectiveness can sacrifice stability, and vice versa.

For example, suppose that there are two queries q_1 and q_2, and for each query we use the average precision (AP) to measure the retrieval performance of a query model estimation method. Assume that we have two estimation methods A and B, where A and B can correspond to the original query model and an expanded query model, respectively. In Table 1, the mean average precision (MAP) over all queries for methods A and B are 0.2 and 0.34, respectively, meaning A is less effective than B. On the other hand, we compute the variance of AP across all concerned queries (denoted as VAP). Specifically, for A, VAP is 0.01, and VAP for B is 0.0676. It turns out the VAP_B is greater than VAP_A. The smaller VAP generally reflects the better retrieval stability. Thus, A is more stable than B. This shows a retrieval effectiveness-stability *tradeoff* between methods A and B.

In this paper, we propose to study the *tradeoff* between retrieval effectiveness and stability from a new perspective, i.e., bias-variance *tradeoff*. The bias-variance tradeoff is fundamental in the estimation theory and has been extensively studied in density estimation (Zucchini et al., 2005), linear regression (Geman et al., 1992), classification (Valentini et al., 2004), and other areas (Bishop, 2006). In general, the bias represents the gap between the expectation (i.e., mean) of estimated values and the true target value, while the variance represents the variability over all estimated values.

This motivates us to formulate the *performance bias and variance*, which are related to the retrieval effectiveness and stability, respectively. Specif-

ically, assumes that we have a performance target (in practice, an upper bound performance). The performance bias represents the gap between the actual mean performance and the mean performance target. For the example in Table 1, assumes that the target AP (i.e., the upper-bound AP) can be 0.7 and 0.2 for queries q_1 and q_2, respectively. Then, for A, the bias is $0.45 - 0.2 = 0.25$, where 0.45 is the target MAP and 0.2 is the MAP of A. Similarly, the bias for method B is 0.11 (see Table 1). On the other hand, the performance variance (denoted as Var in Table 1) is the variance of the retrieval performance across different queries, i.e., VAP. In Table 1, there is a bias-variance tradeoff between A and B, and the smaller performance bias and variance generally reflect the better retrieval effectiveness and stability, respectively. Therefore, we can investigate the problem of *improving* the retrieval effectiveness and stability from the perspective of *reducing* performance bias and variance, respectively [1].

In addition to the performance bias-variance, we also formulate the estimation bias-variance to measure the estimation quality of an estimated query model with respect to the *true* query model. Assume that the true information need can be represented by a set of truly relevant documents. Then, the true query model can be generated from truly relevant documents. Such a true query model is expected to give the upper-bound retrieval performance. The estimation error of an estimated model can be measured by the KL-divergence between the estimated model and the true model. The estimation bias is the expected estimation error over all queries, while the estimation variance is the variance of the estimation error across different individual queries. The sum of bias and variance (see Section 3.3) can yield the total estimation error which directly indicates the total estimation quality. The estimation bias-variance is important, in that it gives finer-grained insights on the estimated query model itself (i.e., its estimation quality).

Our bias-variance analysis is based on general principles of bias-variance tradeoff and four query modeling factors (i.e., query model complexity, query model combination, document weight smoothness, non-relevant document removal). We investigate a series of estimated query models corresponding to the above factors, and analyze when and why the bias-variance tradeoff will occur and how the bias and variance can be reduced simultaneously. Based on the analysis, a set of hypotheses is formed. We then carry out

[1] We will also define additional performance bias-variance in Section 3.2.2.

extensive experiments based on TREC datasets to systematically evaluate the hypotheses based on the bias-variance analysis. Experimental results on performance bias-variance can generally verify the hypotheses. This shows that the retrieval effectiveness and stability can be studied via the performance bias-variance formulation and the general principles of bias-variance analysis. The experimental results on estimation bias-variance can verify the hypotheses on the occurrence of bias-variance tradeoff, but do not fully support hypotheses regarding the simultaneously reduction of the bias and variance. It is an interesting result though, since we find that the corresponding estimated query model may over-fit the relevant feedback documents, but may not fit the relevant documents that do not appear in the feedback document set. It can demonstrate that the improved retrieval performance can not always guarantee the improvement of the estimation quality.

The proposed bias-variance analysis is expected to form an analysis framework and potentially a novel evaluation strategy for the query language modeling. First, for a query modeling approach (or in general other IR models), we can analyze its modeling factors (e.g., model complexity or model combination) and propose hypotheses on the bias-variance tradeoff or even predict the bias-variance trends of the retrieval performance or estimation error/quality. Second, with respect to the evaluation strategy, the estimation bias-variance formulation can provide novel metrics (e.g., estimation bias, estimation variance, and the sum of them) to evaluate the estimation quality of an estimated model. In addition, the summed quantity of performance bias and variance (see Eq. 6 in Section 3.2.1), can naturally be a unified retrieval robustness metric combining retrieval effectiveness and stability.

The rest of the paper is organized as follows. In the next section, we present a literature review on the query language model estimation. Then, in Section 3, we formulate the performance bias-variance as well as the estimation bias-variance. Section 4 presents and analyzes various estimated query models in relation to the bias-variance tradeoff. In Section 5, we move on to the evaluation of the hypotheses of the bias-variance analysis for the concerned query language models. Finally, in Section 6, we conclude our paper by summarizing the main contributions and highlighting the potential impact and future research directions.

2. Literature Review

Over decades, various probabilistic IR models have been developed (Maron and Kuhns, 1960; Lafferty and Zhai, 2003; Zhai, 2007; Robertson and Zaragoza, 2009) to estimate document relevance with respect to an information need (often represented as a query). One way is from the document-generation point of view, leading to the classical probabilistic model (Robertson and Zaragoza, 2009), while another way is from the query-generation perspective, leading to the language modeling approach (Lafferty and Zhai, 2003). Lafferty and Zhai (2003) considered the above two directions into a unified generative relevance model. Indeed, there are other kinds of probabilistic retrieval models (Fuhr, 2001; van Rijsbergen, 1997; Zhai, 2007). Our focus in this paper is on the language modeling (LM) approach.

The LM approaches (Ponte and Croft, 1998; Zhai and Lafferty, 2001) are derived by estimating how probable it is for a document to generate a query (Sparck Jones et al., 2003). There is no explicit *relevance* in the formulation in early LM approaches, where the query representation is the original query language model estimated by the maximum-likelihood method. Later on, the relevance model (RM) (Lavrenko and Croft, 2001) was developed by assuming that the query and its relevant documents are random samples from an underlying relevance model R. In practice, RM estimates an expanded query language model, which is generated from pseudo-relevant documents, rather than truly relevant documents. It is natural to assume that the truly relevant query language model (true query model for short in the rest of the paper) can be generated from the truly relevant documents given a query.

Despite its effectiveness in general, the expanded query model is often less stable in the sense that its performance is not stable across different individual queries (Collins-Thompson, 2009a). The expanded query model may perform less effectively than the original query model for some queries (Amati et al., 2004). Recently, many methods have been proposed to improve the robustness of query expansion. Tao and Zhai (2006) proposed a method to integrate the original query with feedback documents in a probabilistic mixture model and then regularize the parameter estimation. Li (2008) considered the original query as a short document, and investigated rank-related priors and term selection in RM. Lv and Zhai (2009) proposed to adaptively combine the original query and the feedback information. Collins-Thompson and Callan (2007) investigated the uncertainty of feedback-based query models and proposed to resample different feedback document models using Boot-

strap sampling. In (Collins-Thompson, 2009b; Dillon and Collins-Thompson, 2010), the risk and reward tradeoff and optimization for query expansion were discussed. Lv et al. (2011) proposed a FeedbackBoost method to improve the robustness of the expanded query model.

In our opinion, retrieval robustness can be considered as a combined criteria of retrieval effectiveness and retrieval stability. However, to our knowledge, existing work did not provide a formulation to decompose retrieval robustness into retrieval effectiveness and retrieval stability. In addition, the tradeoff between retrieval effectiveness and stability has not been studied via the bias-variance tradeoff. Moreover, existing work on query language modeling paid more attention to the retrieval performance than to the estimation quality with respect to the true query model.

The variance of retrieval performance across different queries has been investigated in the literature (Banks et al., 1999). The variation of the query difficulty/hardness across different topics was studied in the query expansion task (Amati et al., 2004). More recently, Robertson and Kanoulas (2012) have investigated the per-topic variance. Such variance comes from different per-topic AP values measured from different simulated document collections. They simulate a number of document collections from one existing collection. Robertson and Kanoulas (2012) did not adopt the bias-variance tradeoff to investigate the retrieval effectiveness and stability across topics/queries.

The proposed bias-variance analysis is different from the existing mean-variance analysis in document ranking (Wang, 2009; Wang and Zhu, 2009; Zhu et al., 2009). In mean-variance analysis, the variance is associated to the relevance score, while the bias and variance in our paper are associated to the retrieval performance and estimation quality. Moreover, in mean-variance analysis, the relationships between mean and variance have not been explored, while in our paper, the relationship between mean and variance is studied by looking at the tradeoff between the bias and variance.

Our work is also related to but different from the recent research on the exploration-exploitation tradeoff in interactive relevance feedback (Karimzadehgan and Zhai, 2010, 2012) and in online learning to rank (Hofmann et al., 2012). We formulate the bias and variance (see the next section) and analyze the tradeoff between them in query language modeling (see Section 4.2), while in (Karimzadehgan and Zhai, 2012), the bias and variance are not defined or formulated. Nevertheless, the exploration-exploitation tradeoff occurs in our experiments (see Section 5.4.4), in the sense that the estimated model may over-fit the relevant feedback documents, but may not fit the other relevant

documents which do not appear in the feedback document set.

3. Formulation of Bias and Variance

3.1. Introduction to Bias-Variance Analysis

The bias-variance analysis is a fundamental theory and has been extensively studied in parameter estimation (Lebanon, 2010; Duda et al., 2001), density estimation (Zucchini et al., 2005), linear regression (Geman et al., 1992), classification (Valentini et al., 2004; Lipka and Stein, 2011), and other areas (Bishop, 2006). We first briefly explain the classical bias-variance decomposition for the squared loss of the estimation.

Let us consider an estimator $\widehat{y}$ for the unknown true target y, where $\widehat{y}$ is determined by the sample X. For different sample X, the value of $\widehat{y}$ varies. Thus, $\widehat{y}$ can be considered as a random variable. The expected squared error loss of the estimation can be decomposed to bias and variance:

$$\begin{aligned} E(\widehat{y}-y)^2 &= E(\widehat{y}-E(\widehat{y})+E(\widehat{y})-y)^2 \\ &= E(\widehat{y}-E(\widehat{y}))^2+(E(\widehat{y})-y)^2 \\ &= Var(\widehat{y})+Bias^2(\widehat{y}) \end{aligned} \tag{1}$$

where the expectation E is computed over all possible $\widehat{y}$, $Bias^2(\widehat{y})$ computes the squared error (i.e., $(E(\widehat{y})-y)^2$) of the expected value $E(\widehat{y})$ with respect to the true value y, and $Var(\widehat{y})$ computes the variance of $\widehat{y}$ across all samples.

The above formulation is a general description of the bias and variance. It can be applied to specific areas with specific explanations. For instance, in parameter estimation, the task is to estimate a parameter (e.g., mean or variance) of the underlying distribution of a given data sample [2]. On different samples (or sampling distributions), the estimated values can be different. In regression or classification, the task is to estimate the response value (in regression) or the class labels (in classification) for any test data point, given a training sample (or called training set). On different training samples, the estimated values could be different (Geman et al., 1992; Bishop, 2006) and the estimated value can be considered as a random variable.

[2] In this paper, we consider the *sample* as a terminology of statistics and refer to each *sample* as a collection of data or information. In some other literature, a sample may be considered as a single data point.

Generally speaking, given the limited size for each sample, there is a *tradeoff* between bias and variance (Geman et al., 1992). For example, a simple estimation method involving less configurations (e.g., less parameters or assumptions) often has higher bias but lower variance, compared with a more complex method (Geman et al., 1992). This means that the expected estimation error of the simple method is often larger than that of the complex one, but the estimated values of the simple method over different samples are more stable than those of the complex one. To reduce the bias and variance simultaneously, one often needs more data (e.g., larger sample size or more training data) (Brain and Webb., 1999; Bishop, 2006; Perlich et al., 2003), or well designed methods (e.g., combination method, also called as ensemble method) (Valentini et al., 2004; Ghahramani et al., 2003). In the context of query language modeling, we will analyze the above factors that can affect the bias and variance in Section 4.2.

3.2. Bias and Variance Regarding Retrieval Performance

We now define the bias-variance analysis in IR. According to previous introduction, the bias considers the expected estimation value over all samples, while the variance represents the variability of the estimated values across different samples. In IR, for evaluating a retrieval model or a query model, we are concerned about its mean retrieval performance over all queries, and also the variability of retrieval performance across different queries. We can consider each query and its corresponding data (e.g., query terms, retrieved documents, or relevance judgements if available) as a sample to test the retrieval performance. Therefore, we can let the actual retrieval performance be a random variable, which can be different for different queries.

3.2.1. Bias-Variance based on Actual Performance $\widehat{P}$

Recall that we consider the actual performance (denoted $\widehat{P}$) as a random variable. For a query q_i, let its actual retrieval performance be $\widehat{P}_i$, and the corresponding performance target be P_i. In query model estimation, given the query q_i, $\widehat{P}_i$ and P_i correspond to the estimated query model and the true query model, respectively.

Now, let $P_i - \widehat{P}_i$ be the difference between $\widehat{P}_i$ and P_i, and the average difference over all queries is:

$$\frac{1}{m}\sum_i (P_i - \widehat{P}_i) = \frac{1}{m}\sum_i P_i - \frac{1}{m}\sum_i \widehat{P}_i \tag{2}$$

Table 2: Basic notations and descriptions related to query language modeling

Notation	Description
$\widehat{\theta}_{q_i}$	*estimated* query language model for q_i
θ_{q_i}	*true* query language model for query q_i
$\widehat{P}_i$	performance of an *estimated* query model $\widehat{\theta}_{q_i}$ for query q_i
P_i	performance of the *true* query model θ_{q_i} for query q_i
$\widehat{\eta}_i$	KL-divergence between true model θ_{q_i} and *estimated* model $\widehat{\theta}_{q_i}$
η_i	KL-divergence between true model θ_{q_i} and *true* model θ_{q_i}

where m is the number of all queries given a test collection.

We first look at the actual performance part, i.e., $\frac{1}{m}\sum_i \widehat{P}_i$, in Eq. 2. We can consider it as an expected value over all queries:

$$E(\widehat{P}) = \sum_i \widehat{P}_i \times p(q_i) = \frac{1}{m}\sum_i \widehat{P}_i \tag{3}$$

where $p(q_i)$ is uniform, meaning that all queries are treated equally. A lot of efforts in IR have been devoted to improve this expected performance. For instance, if the average precision (AP) is used as the performance metric, $\widehat{P}_i$ represents the AP for each individual query q_i and $E(\widehat{P})$ represents the mean average precision (MAP) over all queries. Note that other performance metrics can also be used in Eq. 3.

Now let us look at the performance target part $\frac{1}{m}\sum_i P_i$ in Eq. 2. Let $P \equiv \frac{1}{m}\sum_i P_i$, which actually denotes the upper bound of $E(\widehat{P})$. Let the difference between the actual mean performance and target mean performance can be defined as the performance bias:

$$Bias(\widehat{P}) = P - E(\widehat{P}) \tag{4}$$

The above $Bias(\widehat{P})$ is equivalent to $\frac{1}{m}\sum_i(P_i - \widehat{P}_i)$ in Eq. 2, which considers the average difference between the actual performance $\widehat{P}_i$ and the performance target P_i over all queries. From Eq. 4, it turns out that the higher $E(\widehat{P})$ (i.e., the actual MAP) is, the smaller performance bias would be, for the same set of queries and the same upper bound performance P.

We now formulate the performance variance as

$$Var(\widehat{P}) = E(\widehat{P} - E(\widehat{P}))^2 \tag{5}$$

which represents the performance variability over different queries, and can indicate the stability of the retrieval performance. Again, in this paper, $E(\widehat{P})$ denotes MAP and $Var(\widehat{P})$ represents the variance of average precision of all concerned queries. We can denote the variance of average precision as VAP. The smaller VAP indicates the better stability of the retrieval performance [3].

Variance of AP (VAP) in fact computes the second central moment of AP, by considering the value of AP on different queries as a random variable. This is helpful to integrate VAP and MAP, the latter being the first moment of AP, into the bias-variance framework.

Now, we can add the bias and variance together, yielding

$$\begin{aligned} Bias^2(\widehat{P}) + Var(\widehat{P}) &= (E(\widehat{P}) - P)^2 + E(\widehat{P} - E(\widehat{P}))^2 \\ &= E(\widehat{P} - E(\widehat{P}) + E(\widehat{P}) - P)^2 \\ &= E(\widehat{P} - P)^2 \end{aligned} \tag{6}$$

This summed quantity $E(\widehat{P} - P)^2$ in Eq. 6 takes into account both performance bias and variance, which are related to retrieval effectiveness and stability, respectively, across all queries.

In our opinion, retrieval robustness is a combined criteria of retrieval effectiveness and stability. Both effectiveness and stability are important in evaluating the robustness of an IR system. Considering only one criteria (effectiveness or stability) is insufficient. Thus, the summed quantity in Eq. 6, which takes into account both retrieval effectiveness and stability, can be considered as a metric for the retrieval robustness. The bias-variance decomposition of $E(\widehat{P} - P)^2$ in Eq. 6 can naturally formulate the effectiveness-stability decomposition of retrieval robustness.

We do not argue that the overall quantity in Eq. 6 can cover every aspect of retrieval robustness in IR. However, it provides a decomposition perspective, which can help us understand and analyze the retrieval robustness. In addition, the bias-variance decomposition can help us analyze the tradeoff between the retrieval effectiveness and stability and then give us some clues on how to improve retrieval robustness.

[3] VAP is the variance of AP, and AP is the performance metric this paper is focused on. Indeed, the bias-variance analysis results are dependent on the choice of the effectiveness metric. If one changes the metric AP to its other forms, e.g., logAP, it would lead to different observations and analysis of the retrieval stability.

3.2.2. Additional Bias-Variance based on Difference between $\widehat{P}_i$ and P_i

In the above bias-variance formulation, the random variable is the actual performance $\widehat{P}$ which is different for different queries. Now, we are going to formulate an additional bias-variance based on the difference between the actual performance $\widehat{P}_i$ and the performance target P_i of each query q_i. First, let $\widehat{\rho}$ denote the random variable representing such a difference which can be different for differen queries. Specifically, let

$$\widehat{\rho}_i = P_i - \widehat{P}_i \tag{7}$$

and accordingly its target $\rho_i = P_i - P_i$. Obviously, $\rho_i = 0$ for each query. Then, we can let $\rho = 0$ be the target difference for each $\widehat{\rho}_i$.

Next, we can define the bias of the random variable $\widehat{\rho}$ as:

$$Bias(\widehat{\rho}) = E(\widehat{\rho}) - \rho = E(\widehat{\rho}) \tag{8}$$

where $E(\widehat{\rho})$ is an expectation value over all queries:

$$E(\widehat{\rho}) = \frac{1}{m}\sum_i \widehat{\rho}_i = \frac{1}{m}\sum_i (P_i - \widehat{P}_i) \tag{9}$$

It shows that $Bias(\widehat{\rho})$ is equivalent to $\frac{1}{m}\sum_i (P_i - \widehat{P}_i)$ and $Bias(\widehat{P})$ (in Eq. 4).

We now define the additional performance variance as

$$Var(\widehat{\rho}) = E(\widehat{\rho} - E(\widehat{\rho}))^2 \tag{10}$$

If P_i is a constant for every query q_i, then $Var(\widehat{\rho})$ would be equivalent to $Var(\widehat{P})$. To illustrate this, we can let $P_i = a$ for every query. Then,

$$\begin{aligned} Var(\widehat{\rho}) &= E(\widehat{\rho} - E(\widehat{\rho}))^2 \\ &= E(a - \widehat{P} - E(a - \widehat{P}))^2 \\ &= E(a - \widehat{P} - a + E(\widehat{P}))^2 \\ &= E(\widehat{P} - E(\widehat{P}))^2 \\ &= Var(\widehat{P}) \end{aligned} \tag{11}$$

In practice, it is not necessary to define every P_i as a constant (e.g., its maximum value), since there is a system variance of performance targets in terms of hardness of different queries. In other words, for different queries

q_i, P_i can be different. Given the existence of the system variance associated with P_i, $Var(\widehat{\rho})$ can be different from $Var(\widehat{P})$. Recall that in $Var(\widehat{\rho})$, the random variable is $\widehat{\rho}$, rather than the actual performance $\widehat{P}$.

We will also investigate the additional performance bias-variance by proposing a regularized $\widehat{\rho}$, in order to reduce the impact of the aforementioned system variance on the bias and variance. We can first regularize the actual performance $\widehat{P}_i$ of each query q_i. Specifically, we can let

$$\widehat{P}'_i = \frac{\widehat{P}_i}{P_i} \tag{12}$$

where $\widehat{P}'_i$ is the regularized actual performance by considering the hardness of a query. Accordingly, the target of $\widehat{P}'_i$ is P'_i=1 (a constant for all queries). In this manner, the system variance of the (regularized) performance target values can be eliminated. We can then define the regularized $\widehat{\rho}_i$ as:

$$\widehat{\rho}'_i = P'_i - \widehat{P}'_i = \frac{P_i - \widehat{P}_i}{P_i} \tag{13}$$

where $\widehat{\rho}'_i$ represents the regularized difference between the actual performance $\widehat{P}_i$ and the performance target P_i for each query q_i.

Based on $\widehat{\rho}'_i$, we can define another additional performance bias as $Bias(\widehat{\rho}')$ and variance as $Var(\widehat{\rho}')$, similarly to Eq. 8 and Eq. 10, respectively:

$$Bias(\widehat{\rho}') = E(\widehat{\rho}') - \rho' = E(\widehat{\rho}') \tag{14}$$

and

$$Var(\widehat{\rho}') = E(\widehat{\rho}' - E(\widehat{\rho}'))^2 \tag{15}$$

Next, we will present an example to discuss the relationships between performance bias-variance and additional performance bias and variance.

3.2.3. Examples on Different Performance Bias-Variance

Let us look at the example in Table 3, from which we can observe the results of different bias-variance defined on different variables. The methods A and B correspond to the original and expanded query models, respectively. In Table 3, for all three variables, the biases of the method A are larger than those of the method B, indicating that the expanded query model B is more effective than original model A, regardless of which variable is used.

Table 3: An example of different bias-variance based on different variables ($\widehat{P}$, $\widehat{\rho}$ and $\widehat{\rho}'$) for query estimation methods (A and B), with performance targets $P_1 = 0.7$ and P_2=0.2

Variable	$\widehat{P}$		$\widehat{\rho}$		$\widehat{\rho}'$	
Description	$\widehat{P}_i$: AP of q_i		$\widehat{\rho}_i = P_i - \widehat{P}_i$		$\widehat{\rho}'_i = (P_i - \widehat{P}_i)/P_i$	
Method	A	B	A	B	A	B
q_1	0.3	0.6	0.4	0.1	0.5714	0.1429
q_2	0.1	0.08	0.1	0.12	0.5	0.6
$Bias$	0.25	0.11	0.25	0.11	0.5357	0.3714
Var	0.01	0.0676	0.0225	0.0001	0.0013	0.0522

Regarding the variance, in Table 3, for the variables $\widehat{P}$ and $\widehat{\rho}'$, the variances of A are smaller than those of B, while for the variable $\widehat{\rho}$, the variance of A is larger than that of B. Recall that the smaller variance reflects the better stability. The variances based on $\widehat{P}$ and $\widehat{\rho}'$ can reveal that the original query model A is more stable than the expanded query model B. However, the variance based on $\widehat{\rho}$ indicates that A is less stable.

In the further analysis and experiments in the later sections, we will pay more attention to the performance variance based on $\widehat{P}$ and $\widehat{\rho}'$. The variance on $\widehat{P}$ directly computes the variance of the actual retrieval performance $\widehat{P}_i$ (e.g., AP) across queries. The variance on $\widehat{\rho}'$ takes into account the difference between the actual performance and the performance target for each query, as well as regularizes the variability of performance targets of different queries.. (see Eq. 13 and Eq. 15).

3.3. Estimation Bias and Variance

Now, we are going to formulate the estimation bias-variance, in order to *directly* investigate the estimation error (or quality) of an estimated query model with respect to the true query model. The estimation error or quality can be based on the divergence between the estimated query model $\widehat{\theta}_q$ and the true query model θ_q. The specific formulation of the estimated and true query models are given in the next section. Here, we focus on the formulation of bias and variance in the estimation process.

For each query q_i, we denote the true query model as θ_{q_i} and an estimated query model as $\widehat{\theta}_{q_i}$. The estimation error can be represented by the KL-

divergence [4] between the $\widehat{\theta}_{q_i}$ and θ_{q_i}:

$$\widehat{\eta}_i = D(\widehat{\theta}_{q_i}|\theta_{q_i}) \tag{16}$$

Then, the mean estimation error over all concerned queries can be defined as the expected value of $\widehat{\eta}$:

$$E(\widehat{\eta}) = \sum_i \widehat{\eta}_i \times p(q_i) = \frac{1}{m}\sum_i D(\widehat{\theta}_{q_i}|\theta_{q_i}) \tag{17}$$

where m denotes the number of queries and $p(q_i)$ is assumed to be uniform, meaning that all queries are treated equally. The expected estimation error in Eq. 17 represents the *bias* of the estimation.

More strictly, for each query q_i, we can consider $\widehat{\eta}_i$ to be an estimated value. The true value can be denoted as η_i, which corresponds to the case when the estimated query model $\widehat{\theta}_{q_i}$ (in Eq. 16) is the true query model θ_{q_i}. It is obvious that $\eta_i = 0$ for each query as $D(\theta_{q_i}|\theta_{q_i}) = 0$. Therefore, we can denote each η_i as η (=0), which is a constant for each query. Now, we have

$$Bias(\widehat{\eta}) = E(\widehat{\eta}) - \eta \tag{18}$$

which is the estimation bias. It equals to the expected value in Eq. 17. The smaller bias indicates the smaller expected estimation error, implying the higher expected estimation quality.

For the estimation variance, we can have

$$Var(\widehat{\eta}) = E(\widehat{\eta} - E(\widehat{\eta}))^2 \tag{19}$$

which represents the variance of the estimation error for different individual queries (i.e., q_i's). The estimation variance represents estimation stability.

By adding the squared bias and variance, we get

$$\begin{aligned} Bias^2(\widehat{\eta}) + Var(\widehat{\eta}) &= (E(\widehat{\eta}) - \eta)^2 + E(\widehat{\eta} - E(\widehat{\eta}))^2 \\ &= E(\widehat{\eta} - E(\widehat{\eta}) + E(\widehat{\eta}) - \eta)^2 \\ &= E(\widehat{\eta} - \eta)^2 \end{aligned} \tag{20}$$

which can represent the total estimation error.

[4] Other Divergence measures (e.g., JS-divergence) could be used in Eq. 16.

4. Bias-Variance Analysis of Query Language Models

In Section 4.1, we give a brief introduction to some background knowledge of the general language modeling (LM) and query language modeling approaches. In Section 4.2, we first formulate the true query model, then present a systematic bias-variance analysis of various estimated query models which reflect different key factors that can affect the model estimation.

4.1. Background of Language Modeling

The query-likelihood (QL) approach (Ponte and Croft, 1998; Zhai and Lafferty, 2001) is a standard LM approach and uses the original query representation. It can be formulated as:

$$p(q_i|\theta_d) = \prod_{j=1}^{m_{q_i}} p(q_{i,j}|\theta_d) \tag{21}$$

where $p(q_i|\theta_d)$ is the query-likelihood, q_i $(q_{i,1}q_{i,2}\cdots q_{i,m_{q_i}})$ is the given original query, m_{q_i} is q_i's length, and θ_d is the smoothed language model for a document d. The QL aims to estimate the probability that this document d generates the query q_i.

The Relevance Model (RM) (Lavrenko and Croft, 2001), as a relevance-based language model and a typical query expansion method, is used to estimate an expanded query language model based on relevance feedback:

$$p(w|\widehat{\theta}_{q_i}^{(f)}) = \sum_{d\in D} p(w|\theta_d)\frac{p(q_i|\theta_d)p(\theta_d)}{\sum_{d'\in D} p(q_i|\theta_{d'})p(\theta_{d'})} \tag{22}$$

where $\widehat{\theta}_{q_i}^{(f)}$ represents the feedback-based expanded query model, $p(\theta_d)$ represents the prior probability of document d, D denotes a set of feedback documents that generate the expanded query model, $p(q_i|\theta_d)$ computes the query-likelihood (QL) score, and the normalized QL score serves as the document weight:

$$S_{q_i}(d) = \frac{p(q_i|\theta_d)p(\theta_d)}{\sum_{d'\in D} p(q_i|\theta_{d'})p(\theta_{d'})} \tag{23}$$

In practice, the documents in D are pseudo-relevant feedback documents, i.e., top-ranked documents retrieved by the QL model (as the first-round retrieval method). After the query expansion, the document ranking is based on the second-round retrieval using the expanded query model.

For any estimated query model, the document retrieval can be based on the negative KL-Divergence (Lafferty and Zhai, 2001) between the estimated query language model $\widehat{\theta}_{q_i}$ and document language model θ_d:

$$-D(\widehat{\theta}_{q_i}|\theta_d) = -H(\widehat{\theta}_{q_i}, \theta_d) + H(\widehat{\theta}_{q_i}) \tag{24}$$

where $H(\widehat{\theta}_{q_i}, \theta_d)$ is the cross entropy between $\widehat{\theta}_{q_i}$ and θ_d, and $H(\widehat{\theta}_{q_i})$ is the entropy of the $\widehat{\theta}_{q_i}$. Each kind of estimated query model can be regarded as one estimation method for the query language model [5].

.

4.2. Analyzing Query Language Models

4.2.1. True Query Model

We first define a form of the true query model. Assuming that the true information need can be reflected or represented by the truly relevant documents, the true query language model should be generated from the truly relevant documents (see also the motivation behind the true query model in the literature review) as follows:

$$p(w|\theta_{q_i}) = \sum_{d \in D_R} p(w|\theta_d) \frac{1}{|D_R|} \tag{25}$$

where θ_{q_i} represents the true query model, D_R denotes the set of all truly relevant documents, given the query q_i. The weights of all documents in the set D_R in Eq. 25 are uniform since they have the same relevance judgements (i.e., 1) given binary judgement values. We think this is a reasonable way of deriving the true query model for the purpose this paper.

4.2.2. Factors Affecting Bias and Variance

We first describe various factors that have an influence on the query model estimation. First, the choice to use original query model or expanded query model would result in different kinds of estimated query models. Second, we consider different combinations (with different combination coefficients) of the original and expanded query models. Third, the change of document weight (in Eq. 23) in RM can lead to different estimation for the query

[5] When we mention query model without specifying any query, it generally refers to an estimation method of query model.

language model. At last, it is important whether or not we have part of true relevance information, e.g., relevance judgements, in building the expanded query models in Eq. 22.

The aforementioned factors actually correspond to the factors that can affect the bias and variance. In Section 3.1, we have mentioned three factors. They are model complexity, model design, and training data size. Regarding query model estimation, the difference between original model and expanded model is related to the model complexity. The expanded query model is often more complex in the sense that: 1) it adopts additional assumptions (Lavrenko and Croft, 2001), e.g., it assumes that the top-ranked documents are relevant; 2) it often involves more parameters, e.g., the number of expanded query terms or the number of feedback documents. The combination strategy and document weight issue are related to the model design. The use of true relevance information can be somewhat considered as use of training data. We emphasize that we do not incorporate any machine learning algorithms (e.g., regression or classification) in our current study.

We now briefly mention different estimated query models for which we will analyze. These models [6] include: 1) original query model and expanded query model; 2) combined query model by original and expanded query models; 3) expanded query model with smoothed document weights for the feedback documents; 4) expanded query model with true relevance information, e.g., some known non-relevant documents. 5) expanded query model with both true relevance information and smoothed document weights.

4.2.3. Original and Expanded Query Models

First, we denote $\widehat{\theta}_{q_i}^{(o)}$ as the original query language model, which is a maximum likelihood estimate of the original query term representation. $\widehat{\theta}_{q_i}^{(f)}$ in RM (see Eq. 22) represents a feedback-based expanded query model.

The expanded query model can usually outperform the original one in terms of the retrieval effectiveness over all queries. As a result, the performance bias of the expanded query model will be smaller than that of the original one. However, for some individual queries, the inclusion of non-relevant documents in pseudo-relevance feedback set can hurt the perfor-

[6] Since our focus is the bias-variance analysis, we may only adopt some basic methods or simple versions of concerned models. This can help us reduce the number of parameters in the retrieval models and it is more feasible to adjust no more than one parameter at a time (if possible) to observe the trends of the changing bias and variance.

mance. Intuitively, a poor initial ranking (by original query) would include many non-relevant feedback documents that are mis-ranked highly. Therefore, for those queries with poor initial performance, query expansion is more likely to hurt the performance, than those queries with better initial performance. A possible consequence after query expansion is that a poor initial performance would become even worse, while a better initial performance would become even better. This can result in the performance variance of the expanded query model being bigger than that of the original one.

For the estimation bias-variance, recall that it is directly related to the divergence/similarity between the estimated query model and the true query model. The original query model $\widehat{\theta}_{q_i}^{(o)}$ is very sparse, in the sense that it only contains the original query terms. On the other hand, the true query model θ_{q_i} (in Eq. 25) and the expanded query model $\widehat{\theta}_{q_i}^{(f)}$ by RM (in Eq. 22) do not have such a sparsity problem since they are generated from a set of documents. Due to the range of KL-divergence in [0,+∞] and the sparsity of the original query model, the scale of $D(\widehat{\theta}_{q_i}^{(o)}|\theta_{q_i})$ and the scale of $D(\widehat{\theta}_{q_i}^{(f)}|\theta_{q_i})$ are quite different – the former values are often much larger than latter values. As a result, the estimation bias (based on KL-divergence) of the original query model $\widehat{\theta}_{q_i}^{(o)}$ will often be much bigger than the expanded model $\widehat{\theta}_{q_i}^{(f)}$. In addition, due to the aforementioned scale difference, the KL-divergence-based estimation variance of the original query model can also be bigger than that of the expanded model. To sum up, in KL-divergence-based estimation bias-variance, the expanded query model often has smaller bias, and can also have smaller variance, compared with the original query model.

The trend of estimation bias-variance can be different when we use other divergence metrices, e.g., JS-divergence (JSD). JSD's range is [0,1], which can be thought of as a normalized range of [0, +∞]. The range [0,1] is also the same as the range of retrieval performance (e.g., Average Precision (AP) or Precision). Therefore, it is more likely that the bias-variance tradeoff can occur in the estimation bias-variance using JS-divergence.

4.2.4. Combination between Original and Expanded Query Models

The combination between original and expanded query models was widely studied in the literature (Abdul-Jaleel et al., 2004; Tao and Zhai, 2006; Li, 2008; Lv and Zhai, 2009). Basically, the combination can be formulated as

$$\widehat{\theta}_{q_i}^{(c)} = \lambda\widehat{\theta}_{q_i}^{(o)} + (1-\lambda)\widehat{\theta}_{q_i}^{(f)} \tag{26}$$

where $\widehat{\theta}_{q_i}^{(c)}$ is the combined query model, λ is the combination coefficient of the original query $\widehat{\theta}_{q_i}^{(o)}$, and $1-\lambda$ is the coefficient of the feedback-based expanded query model $\widehat{\theta}_{q_i}^{(f)}$. The combined query model in Eq. 26 is often referred to as RM3 (Abdul-Jaleel et al., 2004).

In Section 3.1, we mentioned that the *combination* method may reduce the bias and variance simultaneously. Therefore, it is expected that the *combined* query model $\widehat{\theta}_{q_i}^{(c)}$ could reduce bias and variance simultaneously, if a proper combination coefficient λ is used. Here, we will investigate how the combined query model can reduce the bias and/or variance, for different kinds of bias-variance formulation.

For performance bias-variance, as discussed previously, one reason why the expanded query model has larger variance is that, for some queries, the performance can be hurt after query expansion when non-relevant terms are brought into query models. One solution can be combining it with the original query model, which can boost the weights of original query terms while reducing the influence of non-relevant terms in the expanded query model. This can actually prevent the query drifting from the underlying information need (Zighelnic and Kurland, 2008). If the downside performance can be prevented, this could reduce the variance of the expanded query model. On the other hand, the bias can also be reduced if the retrieval performance on average can be improved, given appropriate combination parameters. To sum up, the combined query model with a proper combination coefficient is expected to reduce both bias and variance, which balances the advantages and disadvantages of the original and expanded query models.

With regard to the estimation bias-variance, when λ is approaching 1, the combined query model is getting close to the original query model and will suffer from the sparsity problem as in the original query model. This can lead to not only the increasing bias but also the increasing variance (along with the increasing λ), as we discussed in Section 4.2.3.

4.2.5. Expanded Query Model with Smoothed Document Weights

Recall that in the true query model (see Eq. 25), the document weights are kept uniform, leading to the most smooth document weight distribution. We think that it is worthwhile to investigate the bias-variance of the expanded query model with smoothed document weights.

To facilitate the investigation, we adopt a simple document weight s-

moothing method (Zhang et al., 2011), which can be formulated as:

$$\widetilde{S_{q_i}}(d) = \frac{[S_{q_i}(d)]^{\frac{1}{s}}}{\sum_{d' \in D}[S_{q_i}(d')]^{\frac{1}{s}}} \tag{27}$$

where $\widetilde{S_{q_i}}(d)$ is the smoothed document weight, $S_{q_i}(d)$ is the original document weight, and $s(s > 0)$ is a parameter that controls the smooth degree of document weights. When $s = 1$, the document weights are unchanged. The larger the s is, the greater degree of the smoothing would be. For example, assuming the original weights are 0.6250 and 0.3750 for d_1 and d_2, and the parameter s is 3, then the smoothed document weights are 0.5425 and 0.4575, meaning the document weight distribution becomes more smooth.

Using the smoothed document weights, the estimated query model can be formulated as:

$$p(w|\widehat{\theta}_{q_i}^{(s)}) = \sum_{d \in D} p(w|\theta_d)\widetilde{S}_{q_i}(d) \tag{28}$$

where $\theta_{q_i}^{(s)}$ can be referred to as smoothed query model which is the expanded query model with smoothed document weights $\widetilde{S}_{q_i}(d)$ (see Eq. 27).

The above smoothing method can improve the document weight smoothness among relevant documents in the pseudo-relevant feedback (PRF) document set. The improved smoothness can also broaden the topic coverage of the expanded query, in order to prevent too many weights on the topics represented in topmost documents which might be non-relevant. It has been shown that properly smoothing the document weights (with moderate smoothing parameters) can improve the effectiveness (measured by MAP) of feedback-based query expansion (Zhang et al., 2010, 2011). On the other hand, for some individual queries, smoothing may affect the discriminativity between the relevant documents and non-relevant document in the PRF document set. For instance, if too much smoothing is imposed and the weights of every PRF documents are the same, no documents will have discriminative weights, even for the relevant ones. To sum up, smoothing the weights of feedback documents (with moderate smoothing parameters) can improve the retrieval effectiveness, but may hurt the performance for some individual queries, leading to the drop of retrieval stability. This then results in a performance bias-variance tradeoff.

The document weight smoothing can play a bigger role in reducing the estimation bias/variance, than in reducing the performance bias/variance.

This is because the estimation bias/variance directly computes the estimation error of the estimated query model with respect to the true query model. In the true query model used in our analysis (see Eq. 25), the weights of relevant documents are the same (i.e., very smooth). The smoothing method can improve the smoothness among relevant documents (in the feedback documents), which makes the estimated query model closer to the true one.

4.2.6. Expanded Query Model with Available Non-Relevant Data

One of the reasons for the stability problem of query expansion is that the expanded query model is often generated from a mixture of relevant and non-relevant documents. As a result, the expanded query term distribution is actually a mixture distribution of relevant terms and non-relevant ones. It is argued, that the retrieval performance can be improved if one can remove the non-relevant distribution from the mixture distribution (Zhang et al., 2009)). In accordance with the assumption in (Zhang et al., 2009), we assume that part of non-relevance information is known. Specifically, we assume that a certain ratio (denoted as parameter r_n) of non-relevant documents is known. We then derive an expanded query model based on RM with part of known non-relevant documents (denoted as D_N) removed:

$$p(w|\widehat{\theta}_{q_i}^{(-n)}) = \sum_{d \in D - D_N} p(w|\theta_d) S_{q_i}(d) \tag{29}$$

where $\widehat{\theta}_{q_i}^{(-n)}$ is the estimated query model, $D - D_N$ is the set of remaining documents, and $S_{q_i}(d)$ is the original document weight computed by the normalized QL score (see Eq. 23). Note that the non-relevant documents are selected in a top-down manner from the initial ranking of feedback documents, since the top non-relevant documents with bigger document weights have more influence on the query expansion.

As the parameter r_n increases, more non-relevant documents can be removed from the PRF document set, meaning the PRF documents are *purer* to be truly relevant. It also means that we have more relevance judgements as r_n increases. It is expected that as the parameter r_n increases, bias and variance can be reduced simultaneously, in terms of both performance bias-variance and estimation bias-variance.

4.2.7. Expanded Query Model with Document Weight Smoothing and Non-Relevant Data

Now, let us consider the idea of combining the use of both relevance information (Section 4.2.6) and document weight smoothing (Section 4.2.5). We then come up with the estimated query model as follows.

$$p(w|\widehat{\theta}_{q_i}^{(-ns)}) = \sum_{d \in D - D_N} p(w|\theta_d)\widehat{S}_{q_i}(d) \tag{30}$$

If one removes all non-relevant documents (i.e., $r_n = 1$) in Eq. 29, the process to smooth the document weights (with increasing smoothing parameter s) can be considered as an attempt to gradually approach the true query model in Eq. 25. By using true relevance information (when $r_n = 1$) and document weight smoothing together, it is expected that the bias and variance can drop simultaneously along with the increasing smoothing parameter s.

5. Experiments

In this section, we are going to evaluate each estimated query model described in the previous section. We first summarize a number of hypotheses, drawing on the analysis in Section 4.2.

5.1. Hypotheses

$h1$: For the original query model and the expanded model by RM, the performance bias-variance tradeoff will occur. The estimation bias-variance tradeoff may not occur when using KL-divergence.

$h2$: For the combined query model, the performance bias-variance tradeoff will occur. A proper combination coefficient can reduce the performance bias and variance simultaneously. The KL-divergence-based estimation bias-variance tradeoff may not occur.

$h3$: For the smoothed query model, the performance bias-variance tradeoff will occur. Compared with the performance bias and variance, the estimation bias-variance tradeoff is less likely to occur.

$h4$: For the expanded query model with available true relevance information (e.g., explicit relevance feedback [7]), the performance bias and variance

[7] In this study, we are using the relevance judgements in the test collection to simulate the explicit relevance feedback of users.

can be reduced simultaneously. The estimation bias and variance can be also reduced simultaneously.

$h5$: For the expanded query model with available true relevance information and document weight smoothing, there is a trend of performance bias and variance can be reduced simultaneously. The estimation bias and variance can be also reduced simultaneously.

The factors which can affect bias and variance include not only various *query model factors* described in Section 4.2.2, but also the *evaluation factors*, e.g., different test query sets and test document collections. We will explain different observations on different evaluation factors in the experiments.

5.2. Evaluation Set-up

The evaluation involves four standard TREC collections, including WSJ (87-92, 173,252 documents), AP (88-89, 164,597 documents) in TREC Disk 1 & 2, ROBUST 2004 (528,155 documents) in TREC Disk 4 & 5, and WT10G (1,692,096 documents). These data sets involve a variety of texts, e.g., newswire articles and Web/blog data. Both WSJ and AP data sets are tested on queries 151-200, while the ROBUST 2004 and WT10G collections are tested on queries 601-700 and 501-550, respectively. The *title* field of the queries is used. Lemur 4.7 (Ogilvie and Callan, 2002) is used for indexing and retrieval. All collections are stemmed using the Porter stemmer and stop words are removed in the indexing process.

The first-round retrieval is carried out by a baseline language modeling (LM) approach, i.e., the query-likelihood (QL) model (Ponte and Croft, 1998; Zhai and Lafferty, 2001) as described in Eq. 21, which uses the original query model. The smoothing method for the document language model is the Dirichlet prior (Zhai and Lafferty, 2001) with fixed value $\mu = 700$.

After the first-round retrieval, the top n ranked documents are selected as the pseudo-relevance feedback (PRF) documents for the query expansion task. We report the results with respect to $n = 30$. Nevertheless, we have similar observations on other n (e.g., 50, 70). The Relevance Model (RM) in Eq. 22, is used as the basic method for query expansion. The number of expanded terms is fixed as 100. For any query model (including the original one), 1000 documents are retrieved by the negative KL-divergence measure.

5.3. Evaluation Metrics

Average precision (AP) is used as the performance metric for each query q_i, and the mean average precision (MAP) is used to measure the retrieval

Table 4: Retrieval effectiveness-stability of original query model (QL, $\lambda = 1$) and expanded query model (RM, $\lambda = 0$)

Collections	**WSJ8792**			**AP8889**		
Topics	Topics 151-200			Topics 151-200		
Metrics	MAP(%)	VAP (%)	$<Init$(%)	MAP(%)	VAP (%)	$<Init$(%)
QL	31.25	5.567	—	30.43	6.255	—
RM	37.01*	6.367	30	38.10*	8.368	30

Collections	**ROBUST2004**			**WT10G**		
Topics	Topics 601-700			Topics 501-550		
Metrics	MAP(%)	VAP (%)	$<Init$(%)	MAP(%)	VAP (%)	$<Init$(%)
QL	29.15	4.121	—	19.78	2.213	—
RM	33.26*	5.550	45	21.59*	2.929	46

*Statistically significant improvements over QL at level 0.05 by Wilcoxon signed rank test,

effectiveness over a set of queries. Then, in Eq. 3, $E(\widehat{P})$ represents MAP, and the larger MAP corresponds to the smaller performance bias $Bias(\widehat{P})$ (see Eq. 4). The variance of average precision (VAP), which can be represented by $Var(\widehat{P})$ in Eq. 5, captures the performance variance and can indicate the retrieval stability. The smaller VAP, the better stability is. We will also report the results of the additional performance bias-variance (i.e., $Bias(\widehat{\rho}')$ and $Var(\widehat{\rho}')$ in Section 3.2.2) based on $\widehat{\rho}'$, i.e., the regularized $\widehat{\rho}$.

The summed quantity of performance bias and variance (e.g., the $E(\widehat{P} - P)^2$ in Eq. 6), which takes into account both bias and variance, can be considered as a retrieval robustness metric integrating retrieval effectiveness and stability. We will refer $E(\widehat{P} - P)^2$ as $bias^2{+}var$.

Another robustness metric, i.e., $<Init$ in (Zighelnic and Kurland, 2008), which tests the percentage of queries for which the retrieval performance is worse than that of the initial ranking (i.e. QL), is also adopted for comparison. $<Init$ is dependent on the performance of original query model and thus is not applicable to the initial ranking (Zighelnic and Kurland, 2008). On the other hand, the summed metric $E(\widehat{P} - P)^2$ is independent of the baseline method (i.e., the initial ranking by original query model).

The estimation bias-variance directly tests the estimation quality of each query model with respect to the true query model. The evaluation metrics are the estimation bias and variance proposed in Section 3.3. Specifically, they are KL-divergence based $Bias(\widehat{\eta})$ and $Var(\widehat{\eta})$.

5.4. Bias-Variance Results for Different Query Models

5.4.1. Original and Expanded Query Models

As we can see from Table 4, on four collections, the expanded query models computed by RM are more effective than the original ones used in the query likelihood (QL) model. This can be observed from the experimental results, that RM significantly outperforms QL in MAP on every collection.

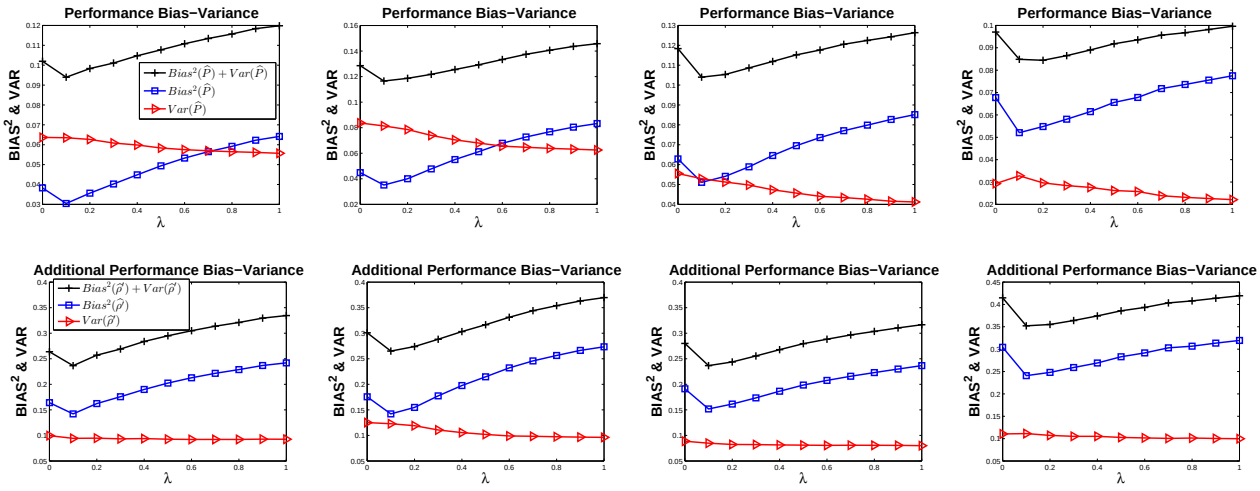
On the other hand, $<Init$ shows that for at least 30% queries (or even 46% on WT10G), the MAP decreases after the query expansion. In addition, the variance of average precision (denoted as VAP) over different queries increases on each collection, meaning that query expansion hurts the retrieval stability. Therefore, we can verify that there is a tradeoff between the retrieval effectiveness and stability.

This tradeoff corresponds to the performance bias-variance tradeoff, which can be observed in Figure 1, where $\lambda = 0$ corresponds to the expanded query model by RM and $\lambda = 1$ to the original model in the combined query model (see Eq. 26). This tradeoff supports our hypothesis $h1$ in Section 5.1.

Now, we look at the results of $bias^2 + var$ plotted in the first row of Figure 1. As mentioned in Section 5.3, $bias^2 + var$ is robustness metric which considers both retrieval effectiveness and stability. It shows that the original query model is more robust than the expanded query model on WSJ8792, AP8889 and ROBUST2004. On WT10G, the original query model is slightly less robust than the expanded query model.

The robustness reflected by $bias^2 + var$ is different from the robustness reflected by another robustness metric $<Init$, which shows the percentage of queries for which the performance is worse than the initial ranking by original query model. Therefore, in any case, the $<Init$ for the original query model will always be 0. This means that the original query model will always be the most robust query model, no matter how bad its MAP is. Therefore, the metric $<Init$, which is dependent on the initial ranking, is not applicable to the original query model.

Let us look at the results (shown in the second row of Figure 1) concerning the additional bias-variance based on $\widehat{\rho}'$ (i.e., the regularized $\widehat{\rho}$), in which the system variance of the (regularized) performance target has vanished. It shows that on all the four collections, the expanded query model ($\lambda = 0$) has smaller bias but larger variance, than the original query model ($\lambda = 1$). The above results show a clear tradeoff of the (additional) performance bias and variance based on $\widehat{\rho}'$, which supports the hypothesis $h1$.



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 1: Performance bias-variance of the combined query model. The x-axis shows λ values from [0,1] with increment 0.1, and the y-axis represents the bias-variance results. $Bias^2$ (which is proportional to $Bias$) is marked with "blue square", Var is marked with "red triangle" and the sum of $Bias^2$ and Var is marked with "black plus sign". The 1^{st} row is the bias-variance based on AP, while the 2^{nd} row is the bias-variance based on $\widehat{\rho}'$.

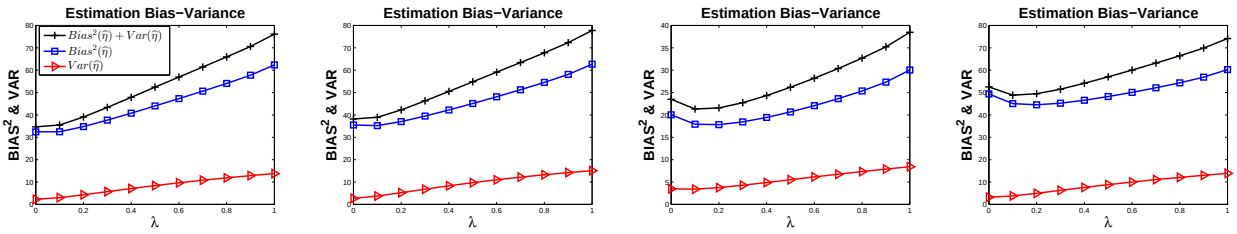
Figure 2 shows the results about the estimation bias and variance of the original and expanded query models, where $\lambda = 1$ corresponds to the original query model based on query likelihood (QL) and $\lambda = 0$ corresponds to the expanded query model by RM. It shows that the KL-divergence based $Bias(\widehat{\eta})$ and $Var(\widehat{\eta})$ do not have a tradeoff on all collections. In Figure 3, we also plotted the JS-divergence based $Bias(\widehat{\eta}')$ and $Var(\widehat{\eta}')$, which has a clear tradeoff on all collections.

The reason why there is no tradeoff between $Bias(\widehat{\eta})$ and $Var(\widehat{\eta})$ is mainly because of the sparsity of the original query model and the range of KL-divergence in [0,+∞] as we discussed in Section 4.2.3.

5.4.2. Combined Query Models with Different Combination Coefficient

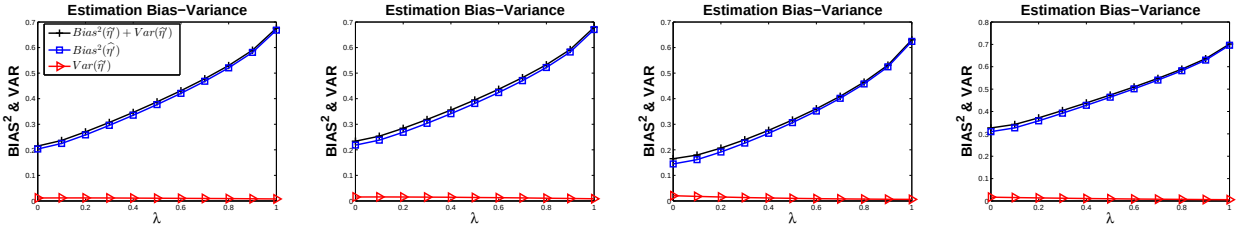
Here, we evaluate the combined query model [8], which is the combination (see Eq. 26) of the original query model (when $\lambda = 1$) and the expanded query model by RM (when $\lambda = 0$). The experimental results are shown in

[8]This is often referred to as RM3 (Abdul-Jaleel et al., 2004).



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 2: Estimation bias-variance based on KL-divergence of the combined query model. The x-axis shows λ values from [0,1] with increment 0.1.



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 3: Estimation bias-variance based on JS-divergence of the combined query model.

Figure 1, where the parameter λ is the combination coefficient with respect to the original query model and λ is chosen from the interval [0,1].

In Figure 1, as λ increases from 0 to 1 with increment 0.1, in most cases, the bias-variance tradeoff happens, evidenced by the fact that the bias and variance change in opposite trends in most cases. Only for a small λ (e.g., 0.1), both the bias and variance can be reduced. The above observation is consistent with our hypothesis $h2$ in Section 5.1.

The original query model only contains original query terms. Then, original query terms in the original query model have much bigger weights than those in the expanded query model. Therefore, a small λ (e.g., 0.1) can adjust the weights of original query terms in the expanded model, while preventing the expanded model from being dominated by original query terms. Evaluation results regarding the robustness metric $bias^2$+var in Figure 1 also suggest that the combined query model with a small λ (e.g., 0.1) can be the most robust one among the models with different λ values.

Now, let us look at the results about the additional bias-variance based on $\widehat{\rho}'$ in the second row of Figure 1. The results shows similar trends with

the bias-variance results based on AP, and support the hypothesis $h2$.

In Figure 2, as λ increases, the KL-divergence based $Bias(\widehat{\eta})$ and $Var(\widehat{\eta})$ both increase. This observation supports the hypothesis $h2$. We also plotted the JS-divergence based $Bias(\widehat{\eta}')$ and $Var(\widehat{\eta}')$ in Figure 3, which shows a clear tradeoff on AP8889, ROBUST2004 and WT10G. This indicates that the bias-variance tradeoff is dependent on the scales of the corresponding metrics (see the discussions in Section 4.2.3).

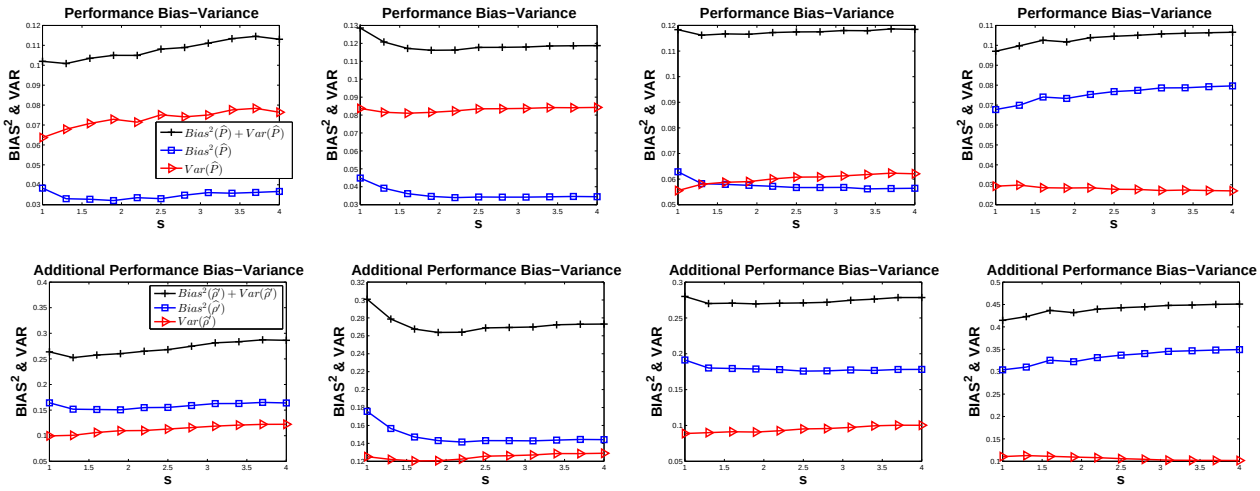
5.4.3. Expanded Query Model with Smoothed Document Weights

Now, we evaluate the expanded query model by RM with smoothed document weights (described in Section 4.2.5). Recall that the bigger the smoothing parameter s is, the more smooth the document weights would be. For RM, we can consider its smoothing parameter s as 1, meaning the document weights remain unchanged. Therefore, RM corresponds the smoothed model when $s = 1$ in Figure 4.

Let us look at the performance bias-variance shown in Figure 4, where parameter s is chosen from the range of [1,4] with the increment 0.3. Along with the increasing smoothing parameter s, the performance bias drops on WSJ8792 ($s < 1.9$), AP8889, ROBUST2004, and increases on WT10G. On the other hand, the performance variance increases on WSJ8792, AP8889 ($s > 1.6$) and ROBUST2004, and drops on WT10G. To sum up, we can observe a clear bias-variance tradeoff on each collection. The above evidence supports our hypothesis $h3$.

We now explain why the observations on WSJ8792, AP8889 and ROBUST2004 are different from those on WT10G. Smoothing also help improve the smoothness of relevant feedback documents in generating the estimated query model. Intuitively, a better initial ranking can have more relevant feedback documents, in which case the smoothing can be more helpful. The initial ranking performance averaged over all queries on WSJ8792, AP8889 and ROBUST2004 is better than that on WT10G (see MAP of QL in Table 4). Therefore, on the first three collections, it is more likely that smoothing can improve MAP and reduce performance bias.

Even if the mean performance on WSJ8792, AP8889 and ROBUST2004 is improved, the performance of some individual queries (with relatively poor initial ranking) can be hurt or can not be improved. This can cause the instability of retrieval performance across queries and increase the performance variance on the three collections. We observe that higher variance of initial ranking performance over queries is more likely to cause the increasing per-



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

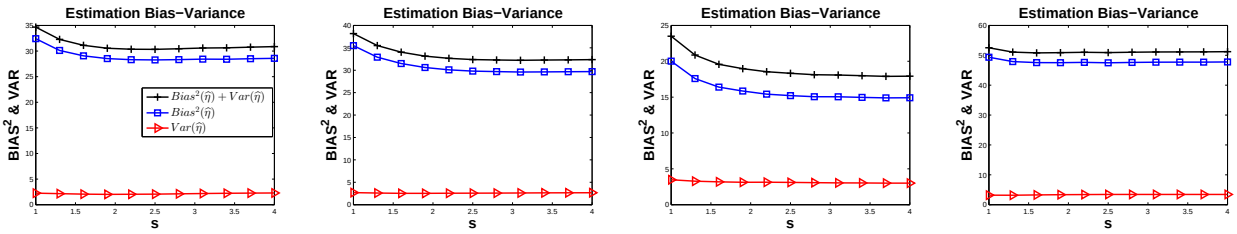
Figure 4: Performance bias-variance of the smoothed query model. The x-axis shows smoothing parameter s from [1,4] with increment 0.3. The 1^{st} row is the bias-variance based on AP, while the 2^{nd} row is the bias-variance based on $\widehat{\rho}'$.

formance variance (based on AP) of the expanded model in Figure 4. The variances of the original query model (see VAP of QL in Table 4) on WSJ8792, AP8889 and ROBUST2004 are all higher than that on the WT10G.

Now, let us look at the results (shown in Figure 4) related to the additional bias-variance based on $\widehat{\rho}'$. It shows that on ROBUST2004 and WT10G, the bias-variance tradeoff obviously occurs. On WSJ8792, compared with the RM-based expanded model (when $s = 1$), the smoothed query model (when $s < 1.9$) has a smaller bias but bigger variance, which also shows a tradeoff.

We now report the estimation bias-variance results plotted in Figure 5, where $s = 1$ corresponds to expanded query model by RM with its original document weights. Recall that the bigger the smoothing parameter s is, the more smoothing would be imposed on the document weights.

We observe that document weight smoothing can help reduce the estimation bias. Regarding the variance, Figure 5 shows that as s increases, the variance almost remains unchanged. The above results support the hypothesis $h3$. As we explained in Section 4.2.5, document weight smoothing can play an important role in the reduction of estimation bias and variance.



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 5: Estimation bias-variance (based on KL-divergence) of the smoothed query model. The x-axis shows smoothing parameter s from [1, 4] with increment 0.3.

5.4.4. Expanded Query Model with Available Non-Relevant Data

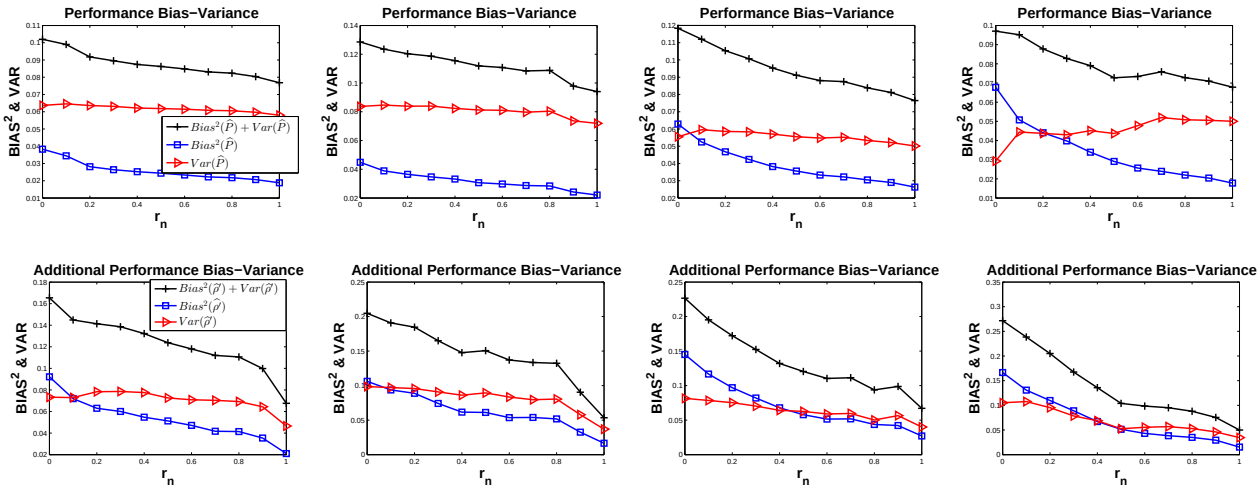
In this subsection, we carry out experiments for the expanded query model by RM with part of non-relevant data available. According to Section 4.2.6. a certain percentage (denoted as r_n) of non-relevant documents are assumed to be available and we simply remove those non-relevant documents (see Eq. 29) in generating the query model. Thus, the expanded query model (by RM only) corresponds to the model when $r_n = 0$ in Figure 6, meaning that no non-relevant data is available.

Let us see the performance bias-variance plotted in the first row of Figure 6, where parameter r_n is in the interval [0,1] with increment 0.1. It clearly shows that on WSJ8792, AP8889, ROBUST2004, performance bias and variance (based on AP) can be reduced simultaneously. The above evidences support our analysis in Section 4.2.6 and the hypothesis $h4$.

The trend of performance variance on WT10G is different from those on the first three collections. The initial ranking on the WT10G is poor (see MAP of QL in Table 4) and then for many queries there are a large number of non-relevant feedback documents in the feedback document set. For those queries, after removing some non-relevant ones, most remaining documents could be still non-relevant and the room for performance improvement is very small. Meanwhile, there may exist some other queries for which the performance improvement can be bigger. As a result, the performance variance will be increased.

For the robustness metric $bias^2 + var$ in Figure 6, it also has a dropping trend on each collection, meaning that removing more non-relevant documents have a good combined effect of effectiveness and stability on each collection.

Let us also look at the results (shown in the second row of Figure 6)



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

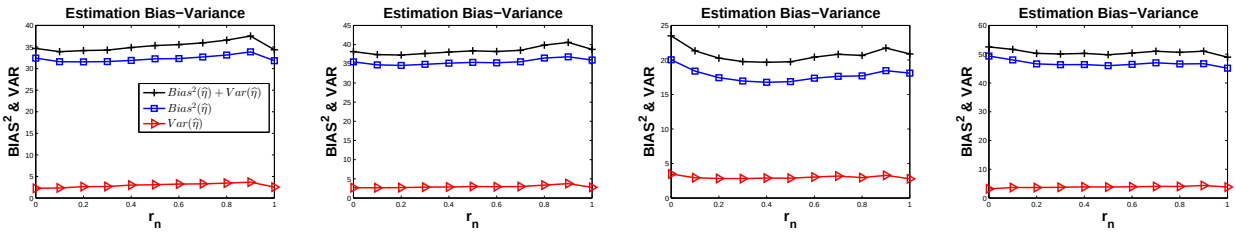
Figure 6: Performance bias-variance of the expanded query model with non-relevant data. The x-axis shows non-relevance percentage r_n from [0,1] with increment 0.1. The 1^{st} row is the bias-variance based on AP, while the 2^{nd} row is the bias-variance based on $\widehat{\rho}'$.

regarding the additional bias-variance based on $\widehat{\rho}'$. It shows that on all the four collections, the bias and variance can be reduced simultaneously, which supports the hypothesis $h4$.

Now, we evaluate the estimation bias and variance of the expanded query model by RM with non-relevant documents available. The results are plotted in Figure 7. It is expected, that by increasing r_n (i.e., removing more non-relevant documents in RM), the estimation quality of the estimated query model can be improved. Note that the expanded query model by RM corresponds to the $r_n = 0$, meaning that no non-relevant data is used. In Figure 7, $Bias(\widehat{\eta})$ first drops and then increases, and $Var(\widehat{\eta})$ has a increasing trend, along with the increasing r_n. This result does not support the hypothesis $h4$, which states that the bias and variance can be reduced simultaneously.

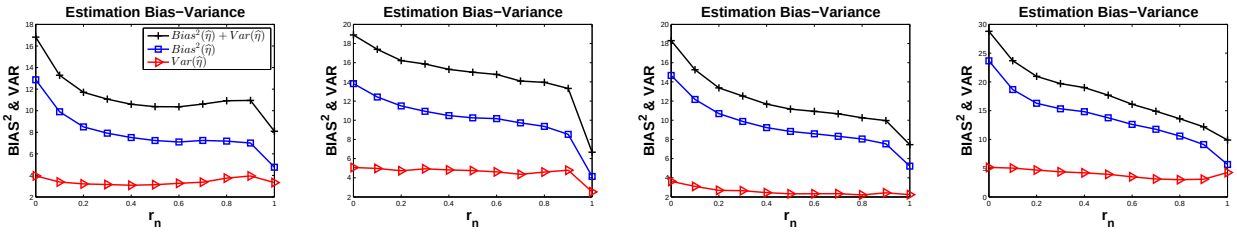
Note that the true query model used for estimation bias-variance evaluation is derived from all the truly relevant documents (see Section 4.2.1), except for the results in Figure 8. In Figure 8, we use truly relevant *feedback* documents as the D_R in Eq. 25 to generate the true query model, which shows that the estimation bias and variance can be often reduced simultaneously.

The above results show that as r_n increases, the estimated query model will gradually get closer to the true query model based on relevant feedback



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 7: Estimation bias-variance (based on KL-divergence) of the expanded query model with non-relevant data available. The x-axis shows the non-relevance percentage r_n from [0,1] with increment 0.1.



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

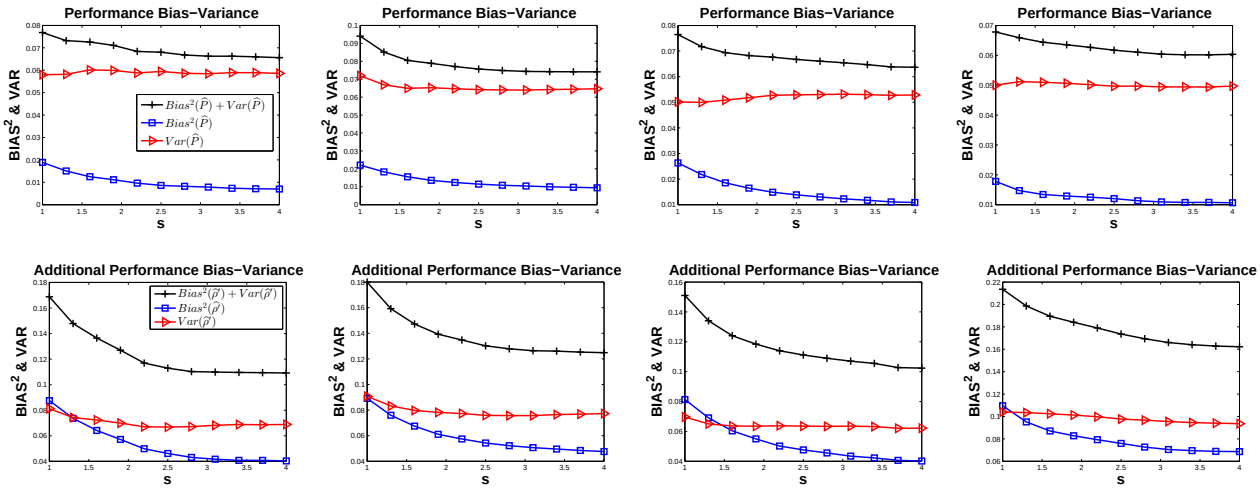
Figure 8: Estimation bias-variance (based on KL-divergence) of the expanded query model with non-relevant data available. The true query model is based on Eq. 25 but using all the relevant *feedback* documents.

documents, but may get far away from the true query model based on all relevant documents. In other words, the estimated query model can *overfit* the relevant feedback documents, but may not fit the other relevant documents which do not occur in the feedback document set.

5.4.5. Expanded Query Model with Relevant Feedback Documents and Document Weight Smoothing

Now, we evaluate the query model described in Eq. 30 in Section 4.2.7. This query model integrates the use of removing all the non-relevant documents and smoothing document weights.

Figure 9 shows performance bias-variance results. From the first row of Figure 9, we observe that when s starts to increase, the performance variance (based on AP) can increase a little bit on WSJ8792, ROBUST2004 and WT10G. This is because when s starts to increase, for some queries, the performance improvements are slow, while for other queries, the improvements



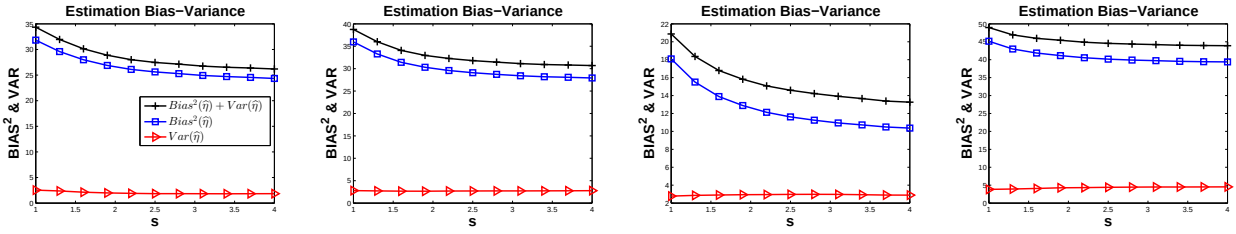
(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 9: Performance bias-variance of the expanded query models on relevant feedback documents with smoothed document wight. The x-axis shows smoothing parameter s from [1,4] with increment 0.3. The 1^{st} row is the bias-variance based on AP, while the 2^{nd} row is the bias-variance based on $\widehat{\rho}'$.

can be relatively rapid, leading to the slightly increased VAP. However, as we can see from Figure 9, there is a clear drop of performance variance when $s > 1.6$ on WSJ8792, $s > 2.2$ on ROBUST2004 and $s > 1.3$ on WT10G. With respect to the performance bias, it has a clear decreasing trend on all collections. Therefore, the performance bias and variance (based on AP) can be simultaneously reduced, which supports the hypothesis $h5$. Figure 9 also shows that as s increases, $bias^2 + var$ keeps dropping.

Let us observe the results of additional performance bias-variance based on $\widehat{\rho}'$ in the second row of Figure 9. As s increases, there is an obvious trend that bias and variance can be reduced simultaneously on each test collection. This observation can support the hypothesis $h5$.

The experimental results in Figure 10 show that as the smoothing parameter s increases, the estimation bias $Bias(\widehat{\eta})$ drops, and the estimation variance $Var(\widehat{\eta})$ almost remains unchanged. This result does not support the hypothesis $h5$, which states that the bias and variance can be reduced simultaneously. It should be noted that the total estimation error (measured by $Bias^2(\widehat{\eta}) + Var(\widehat{\eta})$) keeps decreasing, which indicates that the overall estimation quality is improved as the smoothing parameter s increases.



(a) WSJ8792 (b) AP8889 (c) ROBUST2004 (d) WT10G

Figure 10: Estimation bias-variance based on (KL-divergence) of the expanded query model based on relevant feedback documents with smoothed document wight. The x-axis shows smoothing parameter s from [1, 4] with increment 0.3.

6. Conclusions and Future Work

6.1. Conclusions

In this paper, we propose a novel bias-variance analysis framework to study the tradeoff between the retrieval effectiveness and stability of query language modeling approaches in the pseudo relevance feedback context. Specifically, we propose a performance bias-variance formulation. This enables us to better analyze and understand the retrieval performance using the bias-variance analysis, which is a fundamental theory in machine learning and statistical estimation. We also go beyond the retrieval performance by directly measuring how closely an estimated query model can approach the true query model derived from the truly relevant documents. This leads to the estimation bias-variance formulation, which is based on the divergence between the estimated query model and the true query model.

Based on four query modeling factors, i.e., query model complexity, query model combination, document weight smoothness and non-relevant documents removal, we analyze a number of representative query model estimation methods and present five hypotheses based on our analysis. In order to test the hypotheses, we then construct a systematic evaluation on four TREC datasets. Experimental results of the performance bias-variance (based on AP and $\widehat{\rho}'$) generally support the hypotheses. This shows that the tradeoff between retrieval effectiveness and stability can be studied through the perspective of bias-variance tradeoff. In the experiments, we have explained when the bias-variance tradeoff can occur, and when the bias and variance can be reduced simultaneously. For example, in Section 4.2.4 and Section 5.4.2, we have explained why the performance bias-variance tradeoff will occur, and why a proper combination coefficient (e.g., $\lambda = 0.1$) can

reduce the performance bias and variance simultaneously.

The experimental results of the estimation bias-variance support the hypotheses $h1-h3$, but do not support the hypotheses $h4-h5$. The hypotheses $h1-h3$ are about whether or not the bias-variance tradeoff will occur, and $h4-h5$ are about when the bias and variance can be reduced simultaneously. The results in Section 5.4.4 show that the non-relevant documents removal can not reduce the estimation bias and variance simultaneously. As we explained, for some queries, this strategy may over-fit the relevant feedback documents, but may not fit the other documents that do not appear in the feedback document set, thus can move away from the true query model defined in Eq. 25. After we have removed all non-relevant documents in the feedback document set, we then use the document weight smoothing to improve the estimation quality. The corresponding results in Section 5.4.5 show that the estimation bias can be reduced, while the estimation variance almost remains unchanged. The total estimation error (summed over the estimation bias and variance) can be reduced by the document weight smoothing method.

The above observations show that improving retrieval performance do not guarantee the improvement of the estimation quality. For example, in Section 5.4.4, in addition to the different trends between the performance bias/variance and estimation bias/variance, the trends of the overall retrieval performance (reflected by the sum of the performance bias and variance) are different from the trends of the overall estimation quality (reflected by the sum of the estimation and variance). This could lead to a future research direction to analyze the estimation quality of the query language modeling, rather than the retrieval performance only.

6.2. Potential Impact and Future Work

This research may potentially lead to a novel evaluation strategy. Specifically, the estimation bias-variance formulation can provide novel metrics (e.g., estimation bias and estimation variance) to evaluate the estimation quality with respect to the true query model. In addition, the summed quantity of performance bias and performance variance (see Eq. 6) can be a kind of robustness metric, which indicates the overall retrieval performance (taking into account both retrieval effectiveness and stability). We carried out a preliminary exploration of the bias-variance decomposition of the overall retrieval performance (taking into account both retrieval effectiveness and stability) based on Average Precision (AP) in Zhang et al. (2013). One

may also introduce different weights for bias and variance respectively, in the summation of them, to reflect how the retrieval robustness should be decomposed differently in different scenarios. Moreover, we will investigate other performance metrics in the performance bias-variance analysis.

Based on the proposed bias-variance analysis and evaluation methodology, we can study other query language model estimation methods (e.g., models in (Collins-Thompson, 2009b; Dillon and Collins-Thompson, 2010; Lv et al., 2011)). The proposed bias-variance analysis could also be applied to study the bias-variance of other IR models in terms of their retrieval effectiveness and stability. For instance, we may be able to study the model complexity of other IR models (e.g., ranking functions). The combination of two query models can be extended to the combination/ensemble of multiple (tens or hundreds) rankers in the web search scenario. In machine learning, ensemble learners can reduce both bias and variance simultaneously. In principle, the ensemble learners correspond to the combined rankers. As another example, the document weight smoothness can be related to the diversity of topic coverage of feedback documents. Further, we may explore non-relevant documents removal in the implicit feedback or interactive feedback scenario.

Moreover, we can explicitly formulate the bias-variance for the social and personalized search where the tradeoff between the mean effectiveness over a collective user population and the variance across individual users needs to be balanced and properly modeled. We expect that the bias-variance analysis proposed in this paper can potentially serve as a start point for the above interesting research directions.

Acknowledgement

The authors would like to thank anonymous reviewers for their constructive comments. This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No. 2013CB329304,2014CB744604), the Natural Science Foundation of China (grant No. 61272265, 61070044, 61105702), and EU's FP7 QONTEXT project (grant No. 247590).

References

Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Metzler, D., Smucker, M. D., Strohman, T., Turtle, H., Wade, C., 2004. Umass at trec 2004: Novelty and hard. In: TREC '04.

Amati, G., Carpineto, C., Romano, G., Bordoni, F. U., 2004. Query difficulty, robustness and selective application of query expansion. In: European Conf. on IR Research. Springer, pp. 127–137.

Banks, D., Over, P., Zhang, N.-F., May 1999. Blind men and elephants: Six approaches to trec data. Inf. Retr. 1 (1-2), 7–34.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Brain, D., Webb., G., 1999. On the effect of data set size on bias and variance in classification learning. In: Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales. pp. 117–128.

Collins-Thompson, K., 2009a. Accounting for stability of retrieval algorithms using risk-reward curves. In: In Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation. pp. 27–28.

Collins-Thompson, K., 2009b. Reducing the risk of query expansion via robust constrained optimization. In: CIKM '09. ACM, New York, NY, USA, pp. 837–846.

Collins-Thompson, K., Callan, J., 2007. Estimation and use of uncertainty in pseudo-relevance feedback. In: SIGIR '07. pp. 303–310.

Dillon, J. V., Collins-Thompson, K., 2010. A unified optimization framework for robust pseudo-relevance feedback algorithms. In: CIKM '10. pp. 1069–1078.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification (2nd Edition), 2nd Edition. Wiley-Interscience.

Fuhr, N., 2001. Language models and uncertain inference in information retrieval. In: Proceedings of the Language Modeling and IR workshop. pp. 6–11.

Geman, S., Bienenstock, E., Doursat, R., January 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4, 1–58.

Ghahramani, Z., , Ghahramani, Z., chul Kim, H., 2003. Bayesian classifier combination.

Hofmann, K., Whiteson, S., de Rijke, M., 2012. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. Information Retrieval Journal.

Karimzadehgan, M., Zhai, C., 2010. Exploration-exploitation tradeoff in interactive relevance feedback. In: CIKM '10. pp. 1397–1400.

Karimzadehgan, M., Zhai, C., 2012. A learning approach to optimizing exploration-exploitation tradeoff in relevance feedback. Journal of Machine Learning Research.

Lafferty, J. D., Zhai, C., 2001. Document language models, query models, and risk minimization for information retrieval. In: SIGIR '01. pp. 111–119.

Lafferty, J. D., Zhai, C., 2003. Probabilistic relevance models based on document and query generation. In: Language Modeling and Information Retrieval. Kluwer Academic Publishers, pp. 1–10.

Lavrenko, V., Croft, W. B., 2001. Relevance-based language models. In: SIGIR '01. pp. 120–127.

Lebanon, G., 2010. Bias, variance, and mse of estimators.

Li, X., 2008. A new robust relevance model in the language model framework. Inf. Process. Manage. 44 (3), 991–1007.

Lipka, N., Stein, B., 2011. Robust models in information retrieval. In: 8th International Workshop on Text-based Information Retrieval (TIR).

Lv, Y., Zhai, C., 2009. Adaptive relevance feedback in information retrieval. In: CIKM '09. pp. 255–264.

Lv, Y., Zhai, C., Chen, W., 2011. A boosting approach to improving pseudo-relevance feedback. In: SIGIR '11. pp. 165–174.

Maron, M. E., Kuhns, J. L., July 1960. On relevance, probabilistic indexing and information retrieval. J. ACM 7, 216–244.

Ogilvie, P., Callan, J., 2002. Experiments using the lemur toolkit. In: TREC '02. pp. 103–108.

Perlich, C., Provost, F. J., Simonoff, J. S., 2003. Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research 4, 211–255.

Ponte, J. M., Croft, W. B., 1998. A language modeling approach to information retrieval. In: SIGIR '98. pp. 275–281.

Robertson, S. E., Kanoulas, E., 2012. On per-topic variance in ir evaluation. In: SIGIR '12. pp. 891–900.

Robertson, S. E., Zaragoza, H., 2009. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3 (4), 333–389.

Sparck Jones, K., Robertson, S., Hiemstra, D., Zaragoza, H., 2003. Language modelling and relevance. In: Language Modeling and Information Retrieval. Kluwer Academic Publishers, pp. 57–31.

Tao, T., Zhai, C., 2006. Regularized estimation of mixture models for robust pseudo-relevance feedback. In: SIGIR '06. pp. 162–169.

Valentini, G., Dietterich, T. G., Cristianini, N., 2004. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. Journal of Machine Learning Research 5, 725–775.

van Rijsbergen, C. J., 1997. Readings in information retrieval. Ch. A non-classical logic for information retrieval, pp. 268–272.

Wang, J., 2009. Mean-variance analysis: A new document ranking theory in information retrieval. In: ECIR. pp. 4–16.

Wang, J., Zhu, J., 2009. Portfolio theory of information retrieval. In: SIGIR '09. pp. 115–122.

Zhai, C., 2007. A brief review of information retrieval models,. In: Technical report, Dept. of Computer Science, UIUC.

Zhai, C., Lafferty, J. D., 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In: SIGIR '01. pp. 334–342.

Zhang, P., Hou, Y., Song, D., 2009. Approximating true relevance distribution from a mixture model based on irrelevance data. In: SIGIR '09. pp. 107–114.

Zhang, P., Song, D., Wang, J., Hou, Y., 2013. Bias-variance decomposition of ir evaluation. In: SIGIR '13. ACM, pp. 1021–1024.

Zhang, P., Song, D., Wang, J., Zhao, X., Hou, Y., 2011. On modeling rank-independent risk in estimating probability of relevance. In: AIRS '11. pp. 13–24.

Zhang, P., Song, D., Zhao, X., Hou, Y., 2010. A study of document weight smoothness in pseudo relevance feedback. In: AIRS '10. pp. 527–538.

Zhu, J., Wang, J., Cox, I. J., Taylor, M. J., 2009. Risky business: modeling and exploiting uncertainty in information retrieval. In: SIGIR '09. pp. 99–106.

Zighelnic, L., Kurland, O., 2008. Query-drift prevention for robust query expansion. In: SIGIR '08. pp. 825–826.

Zucchini, W., Berzel, A., Nenadic, O., 2005. Applied smoothing techniques. In: Lecture notes, Institute for Statistics and Econometrics, University of Gottingen.