Bias-Variance Decomposition of IR Evaluation

Peng Zhang[1], Dawei Song[1,2], Jun Wang[3], Yuexian Hou[1]
[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]The Computing Department, The Open University, United Kingdom
[3]Department of Computer Science, University College London, United Kingdom
{darcyzzj, dawei.song2010}@gmail.com, jun_wang@acm.org, yxhou@tju.edu.cn

ABSTRACT

It has been recognized that, when an information retrieval (IR) system achieves improvement in mean retrieval effectiveness (e.g. mean average precision (MAP)) over all the queries, the performance (e.g., average precision (AP)) of some individual queries could be hurt, resulting in retrieval instability. Some stability/robustness metrics have been proposed. However, they are often defined separately from the mean effectiveness metric. Consequently, there is a lack of a unified formulation of effectiveness, stability and overall retrieval quality (considering both). In this paper, we present a unified formulation based on the bias-variance decomposition. Correspondingly, a novel evaluation methodology is developed to evaluate the effectiveness and stability in an integrated manner. A case study applying the proposed methodology to evaluation of query language modeling illustrates the usefulness and analytical power of our approach.

Category and Subject Descriptors: H.3.3 [Information Search and Retrieval]

General Terms: Theory, Measurement, Performance

Keywords: Bias-Variance, Decomposition, Effectiveness, Stability, Robustness, Evaluation

1. INTRODUCTION

Recently, it has been noticed that when we are trying to improve the mean retrieval effectiveness (e.g., measured by MAP [3]) over all queries, the stability of performance across different individual queries could be hurt. For example, compared with a baseline (using the original query in the first round retrieval), query expansion based on pseudo relevance feedback can generally achieve better MAP, but it can hurt the performance for some individual queries [2].

In the literature, various stability (or robustness) measures, e.g., $R-Loss$ [2], Robustness Index [2], and $< Init$ [9] have been proposed. $R - Loss$ computes the *averaged* net loss of relevant documents (due to query expansion failure) in the retrieved documents. $<Init$ calculates the percentage of queries for which the retrieval performance of a query model is worse than that of the original query model. The robustness index is defined as $RI(Q) = (n_+ - n_-)/|Q|$ [2], where n_+ is the number of queries for which the performance is improved, n_- is the number of queries hurt, over the original model, and $|Q|$ is the number of all queries.

These existing measures have some limitations, as shown by the example in Table 1. Let us consider model A as the original query model, as well as regard B and C as two query expansion models. The example shows that A is less effective (with a lower MAP), but more stable (with a lower $< Init$) than B. In this case, we can not judge which model (A or B) has the better overall performance (considering both effectiveness and stability). For comparison of B and C, both $<Init$ and robustness index can not distinguish the stability between them. The MAPs of B and C are also the same. Intuitively, B would seem more stable due to the less variance of its AP values for different queries. However, the above stability metrics do not take into account such variance (denoted as Var in Table 1). For example, one way of computing the variance of AP for model A can be $[(0.3 - 0.2)^2 + (0.1 - 0.2)^2]/2 = 0.01$, which calculates the derivation of AP from its MAP (i.e., 0.2). In Table 1, Var distinguishes the models A, B and C: the smaller Var can indicate the better retrieval stability.

Now let us change the AP values of C to 0.32 and 0.11 (for q_1 and q_2 respectively), then its MAP, $< Init$ and Robustness Index become 0.215, 0 and 1 respectively, suggesting C is more stable but less effective than B. We are not sure which one (B or C) is overall better when considering both effectiveness and stability. This is mainly because the existing stability metrics are often defined separately from the effectiveness metric (MAP). There is a lack of a unified formulation of the effectiveness and stability to allow them to be looked at in an integrated manner.

In this paper, we present a formulation based on a fundamental concept in Statistics, namely the bias-variance decomposition, to tackle the problem. In a nutshell, we view the unsatisfactory overall performance as one total error, which can be decomposed into bias and variance of the AP values of different queries. The bias captures the expected difference of the APs from their upper bounds (the best AP obtained by model T in Table 1). The variance has been illustrated earlier. The detailed formulation will be given in the next section. Briefly speaking, the smaller bias or variance reflects the better retrieval effectiveness or stability, respectively. The total error (denoted as $Bias^2 + Var$) can reflect the overall quality of a model. As an illustration,

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

Table 1: Examples of Bias-Variance (AP)

Model	A		B		C		T	
	q_1	q_2	q_1	q_2	q_1	q_2	q_1	q_2
AP	0.3	0.1	0.6	0.08	0.65	0.03	0.7	0.2
MAP	0.2		0.34		0.34		0.45	
Robust Index	0		0		0		1	
$< Init$	0		0.5		0.5		0	
$Bias$	0.25		0.11		0.11		0	
Var	0.01		0.0646		0.0961		0.0625	
$Bias^2 + Var$	0.0725		0.0797		0.182		0.0625	

in Table 1, $Bias^2 + Var$ indicates that (1) the target model T has the best overall retrieval quality; (2) the model B is more desirable than C; (3) in comparison with the baseline A, both query expansion models B and C reduce the bias but enlarge the variance as well as the total error.

Recently, Wang and Zhu [7] studied the mean-variance analysis regarding the document relevance scores. The per-topic variance of AP (of each query) was investigated in [6]. In this paper, we explore the variance of AP (across queries) and the bias-variance decomposition.

2. BIAS-VARIANCE DECOMPOSITION OF RETRIEVAL PERFORMANCE

In Statistics [1], bias and variance, which are measurements of estimation quality in different aspects, are decomposed from the expected squared error of the estimated values with respect to the target value. In this paper, we formulate the bias-variance decomposition in the IR evaluation scenario. Let us first assume there exists a target model T that can have an upper-bound performance for each query [1].

2.1 Bias-Variance of AP

We can let AP be a random variable over queries. $Bias(\text{AP})$ essentially calculates the average difference between the target AP (i.e. the AP of the target model T, denoted AP_T) and the actual AP of a retrieval model over all different queries. Specifically,

$$Bias(\text{AP}) = E(\text{AP}_T - \text{AP}) = E(\text{AP}_T) - E(\text{AP}) \qquad (1)$$

where the expectation $E(\cdot)$ is over a set of queries that are assumed to be uniformly distributed. $E(\text{AP}_T)$ corresponds to the target MAP (i.e., the MAP of the target model T, denoted MAP_T), and $E(\text{AP})$ corresponds to the MAP of the retrieval model under evaluation.

It turns out that the smaller bias indicates the better mean retrieval effectiveness over queries. In Table 1, the bias for model A is $\frac{1}{2}[(0.7 - 0.3) + (0.2 - 0.1)] = 0.25$. According to Eq. 1, it can be equivalently calculated as 0.45-0.2 = 0.25, where 0.45 is the target MAP and 0.2 is the MAP of A.

Similarly, the variance of the APs for different queries is defined as:

$$Var(\text{AP}) = E(\text{AP} - E(\text{AP}))^2 \qquad (2)$$

The smaller $Var(\text{AP})$ indicates the better retrieval stability of AP across queries.

Table 2: Examples of Additional Bias-Variance ($\widehat{\rho}$)

Model	A		B		C		T	
	q_1	q_2	q_1	q_2	q_1	q_2	q_1	q_2
$\widehat{\rho}$	0.4	0.1	0.1	0.12	0.05	0.17 0	0	0
$Bias$	0.25		0.11		0.11		0	
Var	0.0225		0.0001		0.0036		0	
$Bias^2 + Var$	0.0850		0.0122		0.0157		0	

We now add up the bias and variance, yielding:

$$\begin{aligned} & Bias^2(\text{AP}) + Var(\text{AP}) \\ &= [E(\text{AP}_T) - E(\text{AP})]^2 + E(\text{AP} - E(\text{AP}))^2 \qquad (3) \\ &= E(\text{AP} - \text{MAP}_T)^2 \end{aligned}$$

It turns out $E(\text{AP} - \text{MAP}_T)^2$ is not exactly the expected squared error $E(\text{AP} - \text{AP}_T)^2$. However, we will show later that $E(\text{AP} - \text{MAP}_T)^2$ is a simplified version of an expected squared error $E(\text{AP} - 1)^2$ in Section 2.3. This error formulation will help us understand the difference between different bias-variance decompositions (see Section 2.3).

2.2 Additional Bias-Variance

To clarify the above observation, let us first look at the decomposition of the expected squared error $E(\text{AP} - \text{AP}_T)^2$. We need to define another random variable

$$\widehat{\rho} = \text{AP}_T - \text{AP} \qquad (4)$$

As an illustration, for the model A in Table 2, $\widehat{\rho}$ is 0.4 (0.7-0.3) and 0.1 (0.2-0.1), for q_1 and q_2 respectively.

Let ρ_T be the target value of $\widehat{\rho}$, which is indeed 0, since for model T, $\widehat{\rho} = \text{AP}_T - \text{AP}_T$=0. We need ρ_T in the following analysis, since the target value is an important component in the standard definition of bias. We define

$$Bias(\widehat{\rho}) = E(\widehat{\rho}) - \rho_T \qquad (5)$$

where $E(\widehat{\rho})$ is the averaged $\widehat{\rho}$ over all queries. Since ρ_T is 0, $Bias(\widehat{\rho})$ equals to $E(\widehat{\rho}) = E(\text{AP}_T - \text{AP})$, which is $Bias(\text{AP})$.

The variance based on $\widehat{\rho}$, denoted $Var(\widehat{\rho})$, computes the variance of the difference between the AP (of the test model) and the AP target across all queries. Formally,

$$Var(\widehat{\rho}) = E(\widehat{\rho} - E(\widehat{\rho}))^2 \qquad (6)$$

As shown in Table 2, $Var(\widehat{\rho})$ of B and C are smaller than $Var(\widehat{\rho})$ of A, indicating that B and C are more stable than A. This observation is different from $< Init$ and $Var(\text{AP})$ which shows that B and C are less stable than A in Table 1.

Now, we derive the decomposition of $E(\text{AP} - \text{AP}_T)^2$:

$$\begin{aligned} E(\text{AP} - \text{AP}_T)^2 &= E(\widehat{\rho} - \rho_T)^2 \\ &= E(\widehat{\rho} - E(\widehat{\rho}))^2 + (E(\widehat{\rho}) - \rho_T)^2 \qquad (7) \\ &= Var(\widehat{\rho}) + Bias^2(\widehat{\rho}) \end{aligned}$$

The above equations show the expected squared error $E(\text{AP} - \text{AP}_T)^2$ can be exactly decomposed into bias and variance on $\widehat{\rho}$. The expected squared error $E(\text{AP} - \text{AP}_T)^2$ always computes the error of the target model as zero, no matter whether or not the target model still has room for improvement. It is likely that the current best performance for some queries can be further advanced in the near future.

[1] We will relax this constraint in Sections 2.4.

2.3 Further Investigation on Decomposition of Expected Squared Error

In order to further investigate two aforementioned bias-variance decompositions, we set 1 (the maximum AP) as the upper-bound AP for each query. We can have an expected squared error $E(\mathrm{AP}-1)^2$ and its decomposition as:

$$\begin{aligned} & E(\mathrm{AP}-1)^2 \\ = & E(\mathrm{AP}-\mathrm{AP}_T+\mathrm{AP}_T-1)^2 \\ = & E(\mathrm{AP}-\mathrm{AP}_T)^2+E(\mathrm{AP}_T-1)^2+2E(\mathrm{AP}-\mathrm{AP}_T)(\mathrm{AP}_T-1) \end{aligned} \tag{8}$$

which shows that $E(\mathrm{AP}-\mathrm{AP}_T)^2$ is only one part of $E(\mathrm{AP}-1)^2$, and the target model T still has an error $E(\mathrm{AP}_T-1)^2$, suggesting there is still a room for improvement.

We can also decompose $E(\mathrm{AP}-1)^2$ as:

$$\begin{aligned} & E(\mathrm{AP}-1)^2 \\ = & E[\mathrm{AP}-E(\mathrm{AP})+E(\mathrm{AP})-1]^2 \\ = & E(\mathrm{AP}-E(\mathrm{AP}))^2+[E(\mathrm{AP})-1]^2 \\ = & Var(\mathrm{AP})+(\mathrm{MAP}-1)^2 \\ = & (1-\mathrm{MAP})^2+Var(\mathrm{AP}) \end{aligned} \tag{9}$$

It turns out the variance parts in Eq. 3 and Eq. 9 are the same (i.e., $Var(\mathrm{AP})$). The term $(1-\mathrm{MAP})^2$ in Eq. 9 has the same trend as $Bias^2(\mathrm{AP})$ (i.e., $(\mathrm{MAP}_T-\mathrm{MAP})^2$), where MAP_T is the upper bound of the MAP. The above observations show that the decomposition in Eq. 3 can be considered as a simplified version of the decomposition in Eq. 9.

2.4 Comparison between Two Bias-Variance

First, the bias-variance of $\widehat{\rho}$ requires a target AP for every query. On the other hand, the bias-variance of AP (in Eq. 3) can be used when only a target MAP (i.e., MAP_T) is given. Specifically, two biases (i.e., $Bias(\mathrm{AP})$ and $Bias(\widehat{\rho})$) are equivalent and both can be computed by $\mathrm{MAP}_T-\mathrm{MAP}$. Regarding variance, the variance of $\widehat{\rho}$ requires AP_T for each query (see Eq. 6 and Eq. 4), while the variance of AP can be calculated without knowing AP_T (see Eq. 2). Thus, the bias-variance based on AP is more flexible for practical use.

Second, the bias-variance of $\widehat{\rho}$ is an exact decomposition of $E(\mathrm{AP}-\mathrm{AP}_T)^2$ and it always regards the target model as a zero-error model. However, the decomposition of $E(\mathrm{AP}-1)^2$ in Eq. 8 shows that the target model can still has error. On the other hand, under the bias-variance of AP, the variance for the target model still exists, indicating that although we can assume a target model as an upper-bound at a certain stage, it can be further improved.

2.5 Bias-Variance Evaluation

To carry out IR evaluation based on the proposed the bias-variance decomposition methods, one needs to choose the upper-bound settings (i.e., the target model T). There can be a number of ways. First, the upper-bound AP for each query can be simply set as 1, which corresponds to a perfect target model. Second, the upper-bound AP for each query can be obtained based on the best AP among evaluated retrieval models in the historic TREC results. Third, one can simply set an upper-bound MAP (i.e., MAP_T).

The first and second upper-bound settings correspond to a *virtual* target model. For example, in the second setting, the target model collects the best AP among many retrieval models. The third setting can correspond to a *real* target model (see the model settings in the experiments.) With the first and second upper-bound settings, either of bias-variance metrics (based on AP or $\widehat{\rho}$) can be adopted. The third setting is suitable for bias-variance of AP (see Section 2.4). Once an upper-bound setting and a variable (AP or $\widehat{\rho}$) are chosen, we can calculate bias-variance figures for a series of models and then see which model can have the minimum bias, or minimum variance, or the sum of them (i.e., the total error). We can also get an overview on the trend of bias and variance over parameters.

3. EMPIRICAL EVALUATION

We use the query modeling as an example of bias-variance evaluation. Due to the page limit, we only report the evaluation results on two standard TREC collections, i.e., WSJ (87-92) over queries 151-200 and ROBUST 2004 over queries 601-700. The *title* field of the queries is used. Lemur 4.7 [5] is used for indexing and retrieval. The top $n = 30$ ranked documents in the initial ranking by the original query model are selected as the pseudo-relevance feedback (PRF) documents. The number of expanded terms is fixed as 100. 1000 documents are retrieved by the KL-divergence model [5].

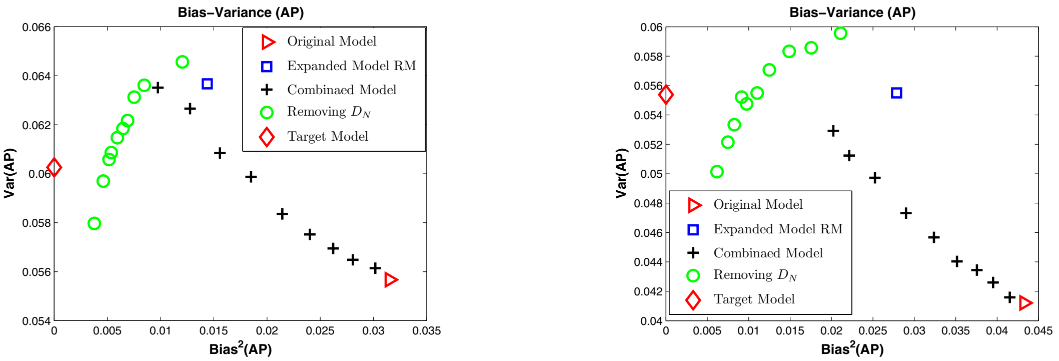
Model Settings We customize a number of query models according to three factors (i.e., model complexity, model combination, ground-truth data) that can generally influence the bias-variance tradeoff [1]. The first model is the *original query model* which is a maximum likelihood estimation of original query. We also evaluate an *expanded query model* RM (i.e., RM1 in [4]) which is generated from feedback documents D. Compared with the original query model, the expanded query model is more complex due to the fact that it has more terms and additional assumptions (e.g., assuming that all feedback documents are relevant). The above two models can be combined, leading to a *combined query model* (also called as RM3). Let λ be the combination coefficient which is in the range (0,1). When λ gets close to 0, the combined model is towards the RM, and otherwise towards the original model. In RM, we can gradually remove a percentage (denoted by r_n) of non-relevant documents D_N along the initial ranking of feedback documents, based on the available relevance judgements (as ground-truth) [8]. We finally derive a *target model* which are generated by keeping only relevant documents D_R in D:

$$p(w|\theta_q^{(t)}) \propto \sum_{d\in D_R} p(w|\theta_d) \tag{10}$$

which gives the best MAP over the aforementioned query models. In Eq. 10, the document weights are set as uniform (different from the weights in RM), since the relevant judgements of relevant documents are the same (i.e., 1).

Bias-Variance Metrics Setting We report the bias-variance on AP in our evaluation, since the target query model used in our evaluation only guarantees the upper-bound MAP. According to the discussion in Sections 2.4 and 2.5, the bias-variance of AP is more suitable.

Evaluation Results We first look at the bias-variance results in Figure 1. Recall that the smaller bias or variance (based on AP) corresponds to the better retrieval effectiveness or stability, respectively. Figure 1(a) shows that on WSJ8792, the original query model has the smallest variance (the highest stability), but the largest bias (the lowest MAP). This is a tradeoff between bias and variance. One



(a) WSJ8792

(b) ROBUST2004

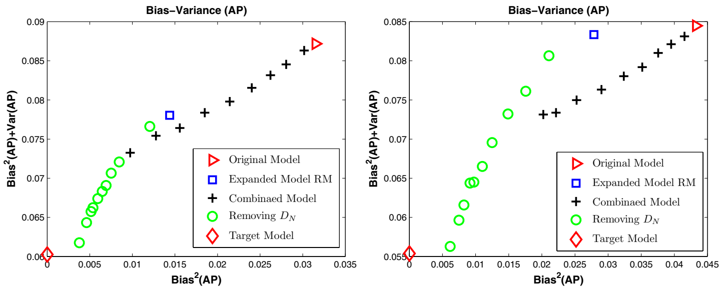
Figure 1: Results of Bias and Variance of AP of all the concerned query models. The x-axis shows the squared bias and the y-axis shows the variance.

reason in Statistics language is that the expanded models are all complex than the original one. The bias of the expanded model RM is much smaller, while its variance is much larger, than the original model. The different variants of the combined model (with λ in [0.1, 0.9] with increment 0.1) are lying between the expanded and original models. We can also see that 2 (out of 9) combined models (when λ=0.1 and 0.2) has smaller bias and smaller variance than the expanded query model RM. This suggests that the combined query model with good parameters can achieve better effectiveness and stability over RM. Among the combined query models, bias and variance are negatively correlated, indicating a bias-variance tradeoff occurs. On the other hand, if we remove the non-relevant documents D_N (with r_n in [0.1, 1] with increment 0.1) in RM, the bias and variance are positively correlated, and 9 (out of 10) models (when $r_n = 0.2$ to 1) can reduce both the bias and variance over RM. The target model has zero bias, although there is a variance of its AP across different queries. On ROBUST2004, the results are similar. The differences are that: 1) by removing non-relevant documents, only 6 (out of 10) models (when $r_n = 0.5$ to 1) reduce both bias and variance over RM; 2) there are 3 (out of 9) (when $\lambda = 0.1$ to 0.3) variants of the combined model for which both bias and variance can be smaller than those of RM.

Figure 2 shows $Bias^2 + Var$ results for all the concerned models. On both collections, the target method achieves the smallest $Bias^2 + Var$ and the original query model achieves the largest $Bias^2 + Var$. Recall that $Bias^2 + Var$ represents the total error and the smaller error can reflect the better overall quality (or performance), by considering both the effectiveness and stability in one single criterion ($Bias^2 + Var$). The results also suggest that the model combination and the relevance judgements are helpful to reduce $Bias^2 + Var$ over the original model and the expanded model RM.

4. CONCLUSIONS AND FUTURE WORK

In this paper, a novel evaluation strategy based on the bias-variance decomposition has been presented. We formulate two forms of the bias-variance decomposition and theoretically compare them. We also use query expansion as an example to demonstrate the use of the bias-variance for IR evaluation and analysis, in terms of bias, variance or sum of them. In the future, we will further explore the other upper-bound settings and variables discussed in Section 2.5, test with more retrieval models and explore the deep insights behind the bias-variance trends.



(a) WSJ8792

(b) ROBUST2004

Figure 2: Results of $Bias^2$ and $Bias^2$+Var of AP of all the concerned query models. The x-axis shows the squared bias and the y-axis shows the $Bias^2 + Var$.

5. ACKNOWLEDGMENTS

This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No. 2013CB329304), the Natural Science Foundation of China (grant No. 61272265, 61070044), and EU's FP7 QONTEXT project (grant No. 247590).

6. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[2] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In CIKM '09, pages 837–846, 2009. ACM.

[3] G. V. Cormack and T. R. Lynam. Statistical Precision of Information Retrieval Evaluation. In SIGIR '06, pages 533–540, 2006. ACM.

[4] V. Lavrenko and W. B. Croft. Relevance-based language models. In *SIGIR '01*, pages 120–127, 2001.

[5] P. Ogilvie and J. Callan. Experiments using the lemur toolkit. In *TREC '02*, pages 103–108, 2002.

[6] S. E. Robertson and E. Kanoulas. On per-topic variance in ir evaluation. In *SIGIR '12*, pages 891–900, 2012.

[7] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09*, pages 115–122, 2009.

[8] P. Zhang, Y. Hou and D. Song. Approximating true relevance distribution from a mixture model based on irrelevance data. In *SIGIR '09*, pages 107–114, 2009.

[9] L. Zighelnic and O. Kurland. Query-drift prevention for robust query expansion. In *SIGIR '08*, pages 825–826, 2008.