Generalized bias-variance evaluation of TREC participated systems

Conference or Workshop Item

How to cite:

Zhang, Peng; Hao, Linxue; Song, Dawei; Wang, Jun; Hou, Yuexian and Hu, Bin (2014). Generalized bias-variance evaluation of TREC participated systems. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, pp. 1911–1914.

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1145/2661829.2661934
http://cikm2014.fudan.edu.cn

Generalized Bias-Variance Evaluation of TREC Participated Systems

Peng Zhang[1], Linxue Hao[1], Dawei Song[1,2], Jun Wang[3], Yuexian Hou[1], Bin Hu[4]
[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
[4]Ubiquitous Awareness and Intelligent Solutions Lab, Lanzhou University, China
{darcyzzj, haolinxue, dawei.song2010}@gmail.com,
jun_wang@acm.org, yxhou@tju.edu.cn, bh@lzu.edu.cn

ABSTRACT

Recent research has shown that the improvement of mean retrieval effectiveness (e.g., MAP) may sacrifice the retrieval stability across queries, implying a tradeoff between effectiveness and stability. The evaluation of both effectiveness and stability are often based on a baseline model, which could be weak or biased. In addition, the effectiveness-stability tradeoff has not been systematically or quantitatively evaluated over TREC participated systems. The above two problems, to some extent, limit our awareness of such tradeoff and its impact on developing future IR models. In this paper, motivated by a recently proposed bias-variance based evaluation, we adopt a strong and unbiased "baseline", which is a virtual target model constructed by the best performance (for each query) among all the participated systems in a retrieval task. We also propose generalized bias-variance metrics, based on which a systematic and quantitative evaluation of the effectiveness-stability tradeoff is carried out over the participated systems in the TREC Ad-hoc Track (1993-1999) and Web Track (2010-2012). We observe a clear effectiveness-stability tradeoff, with a trend of becoming more obvious in more recent years. This implies that when we pursue more effective IR systems over years, the stability has become problematic and could have been largely overlooked.

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

General Terms: Theory, Measurement, Performance

Keywords: Evaluation, Effectiveness-stability tradeoff, Bias-variance tradeoff, Virtual target model

1. INTRODUCTION

While IR research is often focused on improving the retrieval effectiveness (e.g., mean average precision), the performance could become instable across queries. Such an effectiveness-stability tradeoff has been evidenced by some recent experiments [1, 3, 8, 9], but the tradeoff has not been systematically and quantitatively evaluated over the systems participated in TREC tasks.

The effectiveness evaluation has a long history in IR. One of the most commonly used effectiveness metrics is the mean average precision (MAP). The evaluation often involves the comparison with a baseline. Armstrong et al. [2] argued that the effectiveness evaluation based on a weak baseline is not sufficient to prove the effectiveness of a test method. On the other hand, the retrieval stability issue is an emerging topic. To address this issue, some risk metrics (e.g., RI(reliability of improvement) and U_{risk}) [7, 4] have been proposed. These risk metrics usually rely on the performance comparison against a single baseline method. However, Dincer et al. [5] described that the evaluation based on a single baseline system is biased since different baselines can yield different risk values. It is suggested that a less biased approach is to construct the baseline from a set of different systems/runs [5]. To sum up, we need a strong and unbiased baseline to evaluate the effectiveness-stability tradeoff.

In this paper, our "baseline" is a virtual target model made from the best performance (for each query) among all the IR systems in a TREC task. The concept of the virtual target model was mentioned in [8]. Indeed, to evaluate the retrieval effectiveness and stability in an integrated manner, the bias-variance metrics of average precision (AP) were proposed in [8]. However, the experiments in [8] are limited to query expansion using a single target model, and have nothing to do with the systems that participated in a real TREC task.

We will propose a generalized formulation of the bias-variance analysis with respect to IR performance metrics (such as MAP and ERR) and construct a systematic evaluation of TREC participated systems. Specifically, we generalize the original bias-variance framework by 1) making the bias-variance definition more consistent with the definition in estimation theory, 2) extending from the bias-variance of AP to MAP and any other mean effectiveness metrics, and 3) explicitly formulate the target model in the evaluation metrics. We extend AP to MAP for avoiding the adverse influence of some too big/small AP values in the computation of variance. In addition, we propose to quantify the tradeoff degree by two standard correlation coefficient measures. Furthermore, we carry out systematic evaluation of the systems that participated in the TREC Ad-hoc Track (1993-1999) and Web Track (2010-2012).

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661934.

2. GENERALIZED BIAS-VARIANCE EVALUATION

2.1 Bias-Variance of AP

In [8], the performance difference between a current model under evaluation and the target model is viewed as a kind of error. By assuming the target AP is 1, which is AP's maximal value, this error can be formulated as $E(AP-1)^2$. It can be decomposed into bias and variance:

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

where the expectation $E(\cdot)$ is computed over all queries and $E(AP)$ computes the mean of AP, i.e., the MAP of the current model. A smaller $E(AP-1)^2$ means that the current model is closer to the target model. Since $E(AP)$ computes MAP, the smaller bias (i.e., $[E(AP)-1]^2$) can reflect the better retrieval effectiveness. The smaller variance of AP suggests that the current model is more stable.

Since the maximal AP (i.e., 1) is an ideal case, a practical version of $Bias^2(AP)$ is also defined in [8] as $[E(AP)-MAP_T]^2$, where MAP_T is the upper-bound MAP that is achieved by a single model in the reported case study [8]. In [8], it states that this practical bias has the similar trend as the original bias in Eq. 1. Using the practical bias, the sum of bias and variance can be:

$$\begin{aligned} & Bias^2(AP) + Var(AP) \\ &= [E(AP)-MAP_T)]^2 + E(AP-E(AP))^2 \\ &= E(AP-MAP_T)^2 \end{aligned} \tag{2}$$

In [8], the retrieval effectiveness-stability tradeoff has been observed based on the above bias-variance in a case study.

2.2 Generalized Bias-variance Formulation

Now, we introduce a generalized bias-variance formulation. Our motivation is that in the estimation theory [6], bias and variance are defined to evaluate the estimation quality of the distribution parameter (i.e., mean) of a variable. If AP is treated as a variable like in [8], it is more general to define bias-variance on the mean of AP's distribution across queries, i.e., Mean of Average Precision (MAP).

2.2.1 Bias-Variance Decomposition on MAP's Squared Error

We first introduce an expected squared error based on MAP as:

$$\int_{Q_S} [MAP(f,Q_S) - MAP(f_T,Q)]^2 p(Q_S) dQ_S \tag{3}$$

where f is a current model under evaluation, and f_T is the target model which has the best performance for each query; Q_S is a query sample of the query population Q which contains all the possible queries. This expected squared error considers the MAP's difference between the current model f on the query samples Q_S and the target model f_T on the query population Q.

To facilitate the bias-variance decomposition, we can rewrite Eq. 3 and decompose it as:

$$\begin{aligned} & E[MAP(f,Q_S) - MAP(f_T,Q)]^2 \\ &= E[MAP(f,Q_S) - E(MAP(f,Q_S))]^2 \\ &\quad + [E(MAP(f,Q_S)) - MAP(f_T,Q)]^2 \\ &= Var(MAP(f,Q_S)) + Bias^2(MAP(f,Q_S)) \end{aligned} \tag{4}$$

where the expectation $E(\cdot)$ is computed over all query samples. If we consider each query as a query sample, the bias-variance of MAP in Eq. 4 is equivalent to the bias-variance of AP in Eq. 2 . Therefore, the bias-variance of AP can be considered as a *special case* of the bias-variance of MAP.

In addition to the single query sampling, we can partition all the queries (denoted as Q) in a test collection into several subsets Q_S, and treat each query subset as a query sample. In the experiments, we adopt two partition methods: one is random partitioning and the other is based on query difficulty (detailed in Section 3.2.2).

Under the above query sample configurations, the term of bias in Eq. 4 can be derived as:

$$\begin{aligned} & Bias^2(MAP(f,Q_S)) \\ &= [E(MAP(f,Q_S)) - MAP(f_T,Q)]^2 \\ &= [MAP(f,Q) - MAP(f_T,Q)]^2 \end{aligned} \tag{5}$$

It computes the derivation of $MAP(f,Q)$ $w.r.t.$ $MAP(f_T,Q)$ which explicitly formulates a virtual target model f_T. f_T is constructed by assigning the best performance (for each query among all the considered systems/runs in a retrieval task) to it. Dincer et al. [5] described that the evaluation based on a single baseline system is biased since different baselines can yield different risk values. The virtual target model f_T designed in this paper is unbiased, since it is constructed by all various considered systems.

From the formulation of the bias in Eq. 5, we can know that the better the retrieval effectiveness (measured by MAP of the current model f) is, the smaller the bias will be. However, the better retrieval effectiveness does not guarantee a better retrieval stability, which will be measured by the variance term (i.e., $Var(MAP(f,Q_S))$ in Eq. 4).

2.2.2 Bias-Variance Decomposition on Normalized MAP

In Eq. 4, the variance term is a direct way to measure the variance of MAP of the current model. However, in this manner, even the target model may have a big variance value because of the variability of the query difficulty across different query subsets. Therefore, we normalize $MAP(f,Q_S)$ as:

$$MAP^c(f,Q_S) = MAP(f,Q_S))/MAP(f_T,Q_S).$$

The target MAP, i.e., $MAP(f_T,Q_S)$ will be 1 for all the query samples after the normalization. Then, the bias-variance formulation on the normalized MAP can be

$$\begin{aligned} & E[MAP^c(f,Q_S) - MAP^c(f_T,Q)]^2 \\ &= E[MAP^c(f,Q_S) - 1]^2 \\ &= E[MAP^c(f,Q_S) - E(MAP^c(f,Q_S))]^2 \\ &\quad + [E(MAP^c(f,Q_S)) - 1]^2 \\ &= Var(MAP^c(f,Q_S)) + Bias^2(MAP^c(f,Q_S)) \end{aligned} \tag{6}$$

Table 1: Datasets and topics used for Ad-hoc Track.

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

where the variance considers the variance of the normalized MAP. We refer this variance as the normalized variance. In this manner, the normalized variance of the target model becomes zero. Now, we can focus on the performance variability caused by the model/system rather than the query difficulty across different query samples. Note that if we use the normalized MAP, the smaller normalized bias can still imply a better retrieval effectiveness.

2.2.3 A General Expected Squared Error and its Bias-Variance Decomposition

In addition to MAP, we can use other metrics (e.g., ERR or NDCG) in Eq. 3. Denoting a mean performance metric as M, we have a more general expected squared error:

$$\int_{Q_S} [M(f, Q_S) - M(f_T, Q)]^2 p(Q_S) dQ_S \qquad (7)$$

Correspondingly, we can have the bias-variance decomposition as we did for MAP and the normalized MAP. The retrieval effectiveness-stability tradeoff reflected by the bias-variance formulation of the metric ERR will also evaluated in our experiments.

3. EXPERIMENTS

3.1 Evaluation Set-up

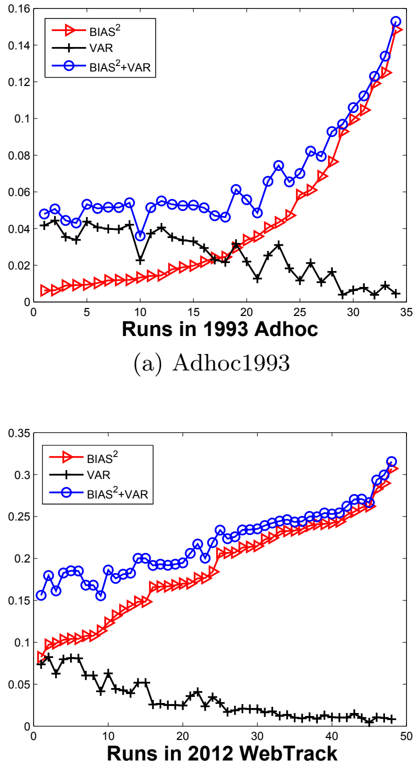
We carry out bias-variance evaluation on Ad-hoc Track and Web Track. For each task, we evaluated all the submitted systems/runs for several years. Table 1 shows the document collections and query topics used for Ad-hoc Track. For Web Track from 2010 to 2012, ClueWeb09 dataset and the query topics (provided by organizers) are used, based on one task (i.e., adhoc task) on Web Track. The Web Track 2013 data are not available to us.

MAP is the effectiveness measure for Ad-hoc Track, and the main effectiveness measure for Web Track is expected reciprocal rank at 20 documents (denoted as ERR@20). For a better distinguishability, we adopt (M)ERR@20 to represent the mean value of ERR@20 on all test queries.

3.2 Evaluation Results

3.2.1 Results of Bias-variance of AP and ERR@20

We analyze the tradeoff between the squared bias and variance of AP on Ad-hoc Track from 1993 to 1999 and ERR@20 on Web Track from 2010 to 2012. In Table 2(a), we quantify the tradeoff by the Pearson and Spearman correlation coefficient of $Bias^2$ and Var. It shows that the correlation coefficients are often strongly negative (i.e., $r < -0.7$) on Ad-hoc Track and Web Track, indicating quite significant tradeoff between effectiveness and stability of runs/systems.



(a) Adhoc1993

(b) WebTrack2012

Figure 1: Results of $Bias^2$,Var and $Bias^2 + Var$ of MAP and (M)ERR@20 based on query difficulty partition, where x-axis represents the runs participated in Ad hoc 1993 and Web Track 2012.

3.2.2 Results of Bias-variance of MAP and (M)ERR@20

The partition strategy of Q can affect the bias-variance results. We first randomly partition the whole query set and each partitioned subset includes 10 queries. The process is repeated for ten times, based on which the average value of these bias and variance is computed.

We can obverse $Bias^2$ and Var of MAP and (M)ERR@20 in Table 2(b). On both tracks, it shows a clear bias-variance tradeoff, evidenced by the strongly negative correlation coefficients. The absolute values of correlation coefficients of MAP and (M)ERR@20 are often smaller than those of AP. This indicates that the variance is more smooth which is consistent with our motivation in designing the variance of MAP (see the last paragraph in Introduction).

There is a problem of random partition: the bias-variance results are different for different random partitioning processes. Therefore, we partition all the queries into several subsets on the basis of query difficulty. The query difficulty is measured by the best performance (i.e., the best AP) of a given query. The lower the best AP is, the more difficult the corresponding query is. We rank all the queries based on the query difficulty degree and group them into several subsets based on the rank, with each subset including 10 queries. Note that we have similar results when we set each subset as different sizes, or measure the query difficulty based on the average performance (among systems) of a query.

Table 2(c) shows bias-variance results based on the query difficulty partition. We still see a clear bias-variance trade-

Table 2: Correlation between $Bias^2$ and Var on four configurations: (a)AP/ERR@20, (b)MAP/(M)ERR@20 based on random partition, (c)MAP/(M)ERR@20 based on query difficulty partition, (d)Normalized MAP/(M)ERR@20 based on query difficulty partition.

	(a)		(b)		(c)		(d)	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Adhoc1993	**-0.8732**	**-0.8686**	**-0.8611**	**-0.9016**	**-0.8761**	**-0.9224**	-0.3433	-0.1878
Adhoc1994	**-0.8640**	**-0.7533**	**-0.7482**	-0.5824	**-0.8332**	**-0.8837**	-0.3830	-0.2616
Adhoc1995	**-0.9376**	**-0.8937**	**-0.9048**	**-0.8783**	**-0.8822**	**-0.9288**	**-0.8021**	-0.6775
Adhoc1996	**-0.8949**	**-0.8132**	**-0.8410**	**-0.7416**	**-0.8315**	**-0.8674**	-0.4553	-0.4079
Adhoc1997	**-0.9139**	**-0.8835**	**-0.8008**	**-0.7047**	**-0.8174**	**-0.8803**	-0.6005	-0.5983
Adhoc1998	**-0.8981**	**-0.8356**	**-0.8513**	**-0.8187**	**-0.8158**	**-0.9091**	-0.6330	-0.6592
Adhoc1999	**-0.9109**	**-0.7501**	**-0.8812**	**-0.7423**	**-0.8605**	**-0.7872**	-0.6859	-0.5622
WebTrack2010	**-0.7981**	**-0.7377**	**-0.8264**	**-0.8743**	**-0.7138**	**-0.8690**	-0.4498	-0.4372
WebTrack2011	**-0.7687**	**-0.7319**	-0.6556	-0.5912	**-0.8259**	**-0.8098**	-0.3890	-0.3586
WebTrack2012	**-0.9509**	**-0.9738**	**-0.9123**	**-0.9446**	**-0.9299**	**-0.9647**	**-0.7719**	**-0.7566**

off. Figure 1 visualizes the tradeoff on the Ad-hoc 1993 and Web track 2012. In Figure 1, all the systems are sorted in an ascending order of $Bias^2$. In addition to bias and variance, it also plots the sum of them ($Bias^2+Var$). We can observe that the Var of most systems/runs increases along with the decrease of $Bias^2$. This stresses a research question: how to get a better retrieval stability (a lower variance) when we pursue a high effectiveness (a lower bias).

In [8], it is stated that the smaller $Bias^2 + Var$ reflects the better overall retrieval performance (considering both effectiveness and stability). In Figure 1, we observe that the lowest $Bias^2 + Var$ is not corresponding to the lowest squared bias. This verifies that the best overall performance is not only determined by the best effectiveness.

3.2.3 Results of Normalized Bias-Variance of MAP and (M)ERR@20

We now report the evaluation results when the bias and variance are all obtained by a normalization process imposed on $MAP(f, Q_S)$. As discussed in Section 2.2.2, the normalization is expected to help us focus more on the performance variability caused by the model/system rather than the variability of the query difficulty.

In Table 2(d), there still exists a tradeoff between $Bias^2$ and Var of regularization MAP and (M)ERR@20 (because of the negative correlation coefficients), but it is less serious compared with the former results. This is because that we have normalized the performance variability caused by the variability of query difficulty.

One interesting point is that if we look at the correlation coefficients along the TREC years, we can observe that the coefficients often become closer to -1 (a more obvious bias-variance tradeoff) along with the increasing "recentness" of years. This indicates a more obvious effectiveness-stability tradeoff, in more recent years.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a generalized bias-variance evaluation strategy and evaluated the IR systems which participated in Ad-hoc track(1993-1999) and Web track (2010-2012). The bias-variance evaluation results can show the effectiveness-stability trends with respect to different systems, tasks, and years. We observe clear bias-variance tradeoff, which indicates retrieval effectiveness-stability tradeoff of the participated systems. In addition, the experimental results (in Figure 1) show that the improvement of effectiveness does not always mean the improvement of the overall retrieval performance (considering both effectiveness and stability). Moreover, the effectiveness-stability tradeoff could become more obvious in more recent TREC years (see Table 2(d)). Since participated systems (especially before 2013) are mainly designed to achieve better performance based on traditional effectiveness metrics, we can speculate that the stability of IR may become more problematic yet could have been overlooked in TREC contests over years.

Currently, we do not know the detailed algorithms/methods behind those tested systems, so our evaluation has a limitation that we do not fully understand why some systems have good effectiveness but bad stability, or are both effective and stable. In the future we will carry out a systematical evaluation of a large number of typical IR models, and analyze such questions in depth, for more insights in how to improve both retrieval effectiveness and stability.

5. ACKNOWLEDGMENTS

The authors would like to thank Mr. Qian Yu for the help on the experiments. This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No.2013CB329304, 2014CB744604), and the Natural Science Foundation of China(grant No. 61402324, 61272265, 61105072).

6. REFERENCES

[1] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *ECIR*, pages 127–137. Springer, 2004.

[2] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM*, pages 601–610. ACM, 2009.

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

[4] K. Collins-Thompson, P. N.Bennett, F. Diaz, and C. Clarke. Trec 2013 web track guidelines.

[5] B. T. Dinçer, I. Ounis, and C. Macdonald. Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In *ECIR*, pages 26–38. Springer, 2014.

[6] G. Lebanon. Bias, variance, and mse of estimators, 2010.

[7] T. Sakai, T. Manabe, and M. Koyama. Flexible pseudo-relevance feedback via selective sampling. *TALIP*, 4(2):111–135, 2005.

[8] P. Zhang, D. Song, J. Wang, and Y. Hou. Bias-variance decomposition of ir evaluation. In *SIGIR*, pages 1021–1024. ACM, 2013.

[9] P. Zhang, D. Song, J. Wang, and Y. Hou. Bias–variance analysis in estimating true query model for information retrieval. *IP&M*, 50(1):199–217, 2014.