Measuring the Variability in Effectiveness of a Retrieval System

Mehdi Hosseini[1], Ingemar J. Cox[1], Natasa Millic-Frayling[2], and Vishwa Vinay[2]

[1] Computer Science Department, University College London
{m.hosseini,i.cox}@cs.ucl.ac.uk
[2] Microsoft Research Cambridge
{natasamf,vvinay}@microsoft.com

Abstract. A typical evaluation of a retrieval system involves computing an effectiveness metric, e.g. average precision, for each topic of a test collection and then using the average of the metric, e.g. mean average precision, to express the overall effectiveness. However, averages do not capture all the important aspects of effectiveness and, used alone, may not be an informative measure of systems' effectiveness. Indeed, in addition to the average, we need to consider the variation of effectiveness across topics. We refer to this variation as the *variability in effectiveness*. In this paper we explore how the variance of a metric can be used as a measure of variability. We define a variability metric, and illustrate how the metric can be used in practice.

1 Introduction

A common practice in a comparative evaluation of information retrieval (IR) systems is to create a test collection comprising a document collection, a set of topics (queries) and associated relevance judgments, and to then measure effectiveness (performance[1]) of retrieval systems over such a collection. A typical evaluation of a system involves computing an effectiveness metric, e.g. average precision (AP), and then averaging across topics, e.g. computing the mean average precision (MAP), to characterize the overall system effectiveness. However, when used alone, averages do not capture all the important aspects of effectiveness. For example, averages may not reveal possibly large variations in effectiveness across topics. We maintain that, in addition to average effectiveness, one needs to consider the variation in effectiveness across topics. In particular, when two systems are not distinguishable based on their average, we can use the variations to contrast them. We refer to the cross-topic variation as the *variability in effectiveness*.

There are various ways in which variability could be measured. In this paper we explore how the variance of IR metrics, in particular, the variance in AP scores, can be used for this purpose. The IR community is, of course, familiar with variance, and uses it routinely to assess whether the difference in the averages

[1] In this paper we use effectiveness and performance interchangeably.

of two systems' effectiveness is significant or not. However, our use of variance is different, and we illustrate this next.

Consider a scenario illustrated in Figure 1a, in which we have two systems, A and B, each of which exhibits the same MAP score, but the variance of AP scores for System A is much larger than for System B. If the two systems are compared based on MAP alone, then a paired student t-test will conclude that the two systems are equivalent. However, in practice, users may observe a significant difference between the two systems. Qualitatively, Figure 1a shows that System A either gives very good or very poor responses to a query. In contrast, System B gives "satisfactory" responses to all queries, i.e. the responses of System B are neither very good nor very bad. Which system would a user prefer? The answer to this question is not entirely straightforward.

Consider a scenario in which users require AP scores to exceed a minimum threshold in order to be satisfied with the response of the system. This is depicted by the horizontal line in Figure 1a. In this case, System B always satisfies users, while half the time, System A fails to satisfy users, despite the fact that both systems have the same MAP. In this case, the system with lower variability is preferred. Now consider Figure 1b, in which we again have two systems, C and D, with the same MAP score. However, in this scenario, the MAP is lower than the threshold needed to satisfy users. In this case, the system with lower variance, D, never satisfies users. In contrast, System C, with high variance does satisfy users for some queries. In this case, the system with higher variability is preferred.

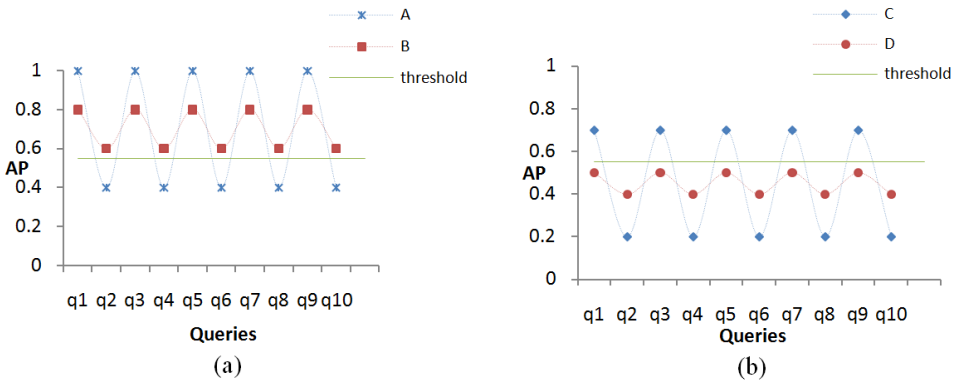

Fig. 1. Two IR systems with (a) equal MAP which is larger than a threshold needed to satisfy a user, and (b) two IR systems with equal MAP smaller than the threshold.

This example highlights three important points. First, the average of a metric does not always provide sufficient information with which to judge a system. Second, variability can be used not just for significance testing, but also to characterize systems with similar average performance. Third, a preference

for systems with high or low performance variability depends on the relative performance of systems in comparison with a user's satisfaction threshold.

Of course, this is an artificial example. However it is common for real systems to exhibit statistically identical mean performance, yet exhibit different levels of variance. For example, Table1 shows two experimental runs [1] from the Robust track of TREC 2004 [11]. For each of them we compute MAP across 199 topics (351-450 and 600-700), removing one topic that had no relevant documents in the collection. We also calculate the standard deviation of AP scores as our measure of variability. The two runs have the same MAP value but different standard deviations. Using the paired t-test reveals that there is no significant difference in the MAP values while applying a statistical test to assess the equality of standard deviation (levene's test, see Section 2) confirms that the difference in the standard deviations is statistically significant.

Runs	**MAP**	**SD(AP)**	**Paired t-test**	**Levene's test**
uogRobSWR5	0.304	0.24	$p = 0.96 \gg 0.05$	$p = 0.007 \ll 0.05$
NLPR04clus10	0.304	0.20		

Table 1. Two experimental runs from the robust track of TREC 2004. The corresponding MAP values and standard deviations of AP scores, SD (AP), are measured over 199 topics.

In Section 2.1, we discuss related work and in Section 2.2 two statistical tests are introduced which are used to determine whether the difference between the variances of two systems' effectiveness is statistically significant. In Section 3, we describe several experiments. We first highlight the problem of using standard deviation (SD) of a metric which has bounded values as a measure of variability. In order to overcome the problem we consider two transformations that have been used in the context of information retrieval. Then we evaluate the systems that participated in two tracks of TREC 2004. When pairs of systems are compared, we find that 26% of these pairs have no statistical difference in their MAP scores. We refer to these systems as "ties". However, 44% of the ties have statistically different levels of variability. Finally, we consider the minimum number of topics required to run a robust comparison in terms of variability. Using two test collections, Robust track and Web track of TREC 2004, we observe that 90 topics are required to run a comparison with less than a 5% error rate. We finally discuss our findings in Section 4.

2 BACKGROUND

In this section, we first briefly discuss related work and then introduce two statistical significant tests used to compare variabilities in systems' effectiveness.

[1] A single retrieval system can have several settings. In TREC each setting is referred to as a run. For our purposes we will treat each run as a search system.

2.1 Related Work

The topic of variability in effectiveness has received little attention in IR research. Perhaps, the most of prior work related to variability is to do query expansion. Query expansion methods typically yield good improvements in mean average precision but are unstable and have high variance across queries [2]. Collins-Thompson [1] proposed a model of evaluating effectiveness of query expansion methods by using a risk-reward tradeoff where reward was defined as the percentage gain for MAP relative to the original, un-expanded query, and the risk reflected the number of relevant documents that were lost due to the expansion. Such a risk measurement is solely based on the number of relevant documents. In contrast, the percentage MAP gain depends on not only the number of relevant documents retrieved but also the ranks of them. Perhaps the variability in effectiveness, as defined in this paper, can be an alternative measure of risk where both number of relevant documents and corresponding ranks are taken into account.

C.T. Lee et al. [5] proposed a novel weighted average (generalized adaptive-weight mean) to rank systems' effectiveness where the weights reflected the ability of the test topics to differentiate among the retrieval systems. The variance of the AP scores was indirectly incorporated into measuring systems' effectiveness. In particular, they used the Euclidean distance to characterize the dispersion of AP scores. However, effectiveness of their system ranking and comparison was not evaluated in detail. We observe that the performance scores (AP values) are bounded in [0, 1] and expect that the approach will be affected by the boundary conditions, 0 and 1, as we discuss in Section 3. In this paper, we propose a way to overcome the issues of a bounded score distribution and its effect on the variance.

2.2 Statistical Significance Tests

Tests of statistical significance have been thoroughly discussed in the IR literature. The common statistical significance tests used in IR experiments are student's paired t-test, wilcoxon signed rank, and sign test. The assumptions which these tests are based on were discussed in [4]. In addition, the use of two sampling-based tests, bootstrap shift method and fisher's randomization, in IR was discussed in [10]. Sakai [8] also discussed the use of paired bootstrap test in IR which was a combination of the bootstrap shift method and student's t-test.

These tests, for example, make use of the variance of AP scores to determine whether the difference in two MAP scores is statistically significant. Here, we are interested in determining whether the difference in variabilities, as measured by variance or standard deviation, of two systems' effectiveness is statistically significant. The statistical community has, of course, addressed this and we briefly describe two tests, the F-test and Levene's test.

F-test This test first defines a ratio of the standard deviations of two systems' effectiveness measured across a set of topics. Therefore, if σ_A and σ_B are the

standard deviations of AP scores of systems A and B, the ratio is calculated as:

$$F = \frac{\sigma_A}{\sigma_B} \tag{1}$$

In the F-test the null hypothesis and the alternative hypothesis is defined as below:

$$H_0 : \sigma_A = \sigma_B \text{ (the null hypothesis)}$$
$$H_1 : \sigma_A \neq \sigma_B \text{ (the alternative hypothesis)}$$

The more the ratio deviates from 1, the stronger the evidence for unequal variances. The null hypothesis is rejected if the ratio was larger than a critical value. The critical value is adjusted based on a significance level, e.g. $\alpha = 0.05$ or $\alpha = 0.01$.

There is a limiting condition in F-test assuming that the distribution of AP scores is normal. However, our experiment in Section 3.3 shows that this assumption is not necessarily true. In order to deal with this restriction, we also consider the Levene's test which does not have such an assumption.

Levene's Test Levene's test is used to assess whether k sample groups have the same standard deviation [6]. Levene's test does not have the normality assumption. The statistic is obtained from one-way analysis of variance (ANOVA), where each observation, in our case each AP score, is replaced with its absolute deviation from the associated group's mean. In our case the group mean is the MAP value. Let $z_{ij} = |AP_{ij} - MAP_i|$, where AP_{ij} is the measured AP value of the i^{th} system on the j^{th} query. Levene's test defines a ratio as:

$$W_0 = \frac{\sum_i n_i(\bar{z}_i - \bar{z})^2 \times \sum_i (n_i - 1)}{(g-1) \times (\sum_i \sum_j (z_{ij} - \bar{z}_i)^2)} \tag{2}$$

where g is the number of sample groups which in our case is 2 indicating the number of systems, and n_i is the number of observations in the i^{th} group (in our case it is equal to the number of topics):

$$\bar{z}_i = \frac{\sum z_{ij}}{n_i} \text{ and } \bar{z} = \frac{\sum\sum z_{ij}}{\sum n_i}$$

The null hypothesis is rejected if W_0 was larger than a critical value that is adjusted with regard to a significance level. Replacing the group mean, MAP_{ij}, with the median of observations, median(AP), in forming z_{ij} defines W_{50}. We use W_{50} instead of W_0 when the AP distributions suffer from skewness.

3 Experiments

In Section 3.1, we examine the performance of various systems involved in the Web and Terabyte tracks of TREC 2004. This study reveals a curious phenomenon - systems with average performances, measured by a bounded IR metric, e.g. MAP, near to 0.5 have larger variances than systems with average performances near to each of the two boundaries, 0 and 1. This phenomenon is an

artifact of the fact that the metric's scores are bounded in [0,1]. Section 3.2 proposes two transformations of the metric's scores in order to eliminate this artifact. Section 3.3 then considers all pairs of systems participating in two test collections of TREC 2004: the Web and Robust tracks. Student t-tests show that 26% and 28% of pairs, respectively, are ties, i.e. there is no statistically significant difference in the averages of transformed AP scores. If the variability in effectiveness of these ties is examined, then the F-test shows that 33% and 34% of ties have statistically significant differences in variance. When Levene's test is used, 47% and 38% of ties have statistically significant differences in variance. Finally, in Section 3.4, we explore the effect that the size of a topic set has on the system comparison using variability in effectiveness. We observe that one needs to consider a sample of 90 topics to obtain an error rate smaller than 0.05.

3.1 The Variance of a Bounded Metric

Figure 2 plots the standard deviation in AP scores as a function of MAP, for systems participating in the Web and Terabyte tracks of TREC 2004. The Web track involves 74 systems and 225 topics. The Terabyte track involves 70 systems and 49 topics.

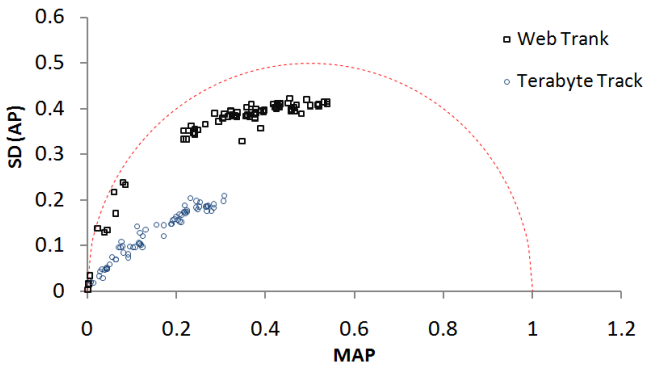

Fig. 2. The standard deviation of AP values (SD (AP)) versus MAP. The standard deviation is bounded in a semicircle with centre (0.5, 0.0) and radius 0.5.

We note that some systems have similar MAP values. Thus, it would be beneficial to use additional criteria, e.g., variability in effectiveness, to differentiate their performance. This is discussed shortly. However, the most striking feature in Figure 2 is an unexpected trend: the monotonic relationship between standard deviation and MAP, i.e. the larger the MAP value, the larger the variance in AP scores.

We believe this relationship is due to the bounded nature of AP metric, i.e. the fact that the metric's values fall within [0,1]. This bounds the standard devi-

ation of AP scores to a semicircle as shown in Figure 2 and proven in Appendix A. Therefore, the retrieval system with MAP close to one of the boundaries, 0 or 1, are more likely to have a smaller variance than those with MAP near to the center, 0.5. For this reason using the standard deviation of the raw AP scores is not a reliable measure of variability. In fact, this is true for any other bounded IR metrics, e.g. the reciprocal rank, as shown in Figure 3. The figure shows how the variance decreases as the MRR increases above 0.5. Again, this is expected since now the variation above the mean is limited by the upper bound of one on reciprocal rank.

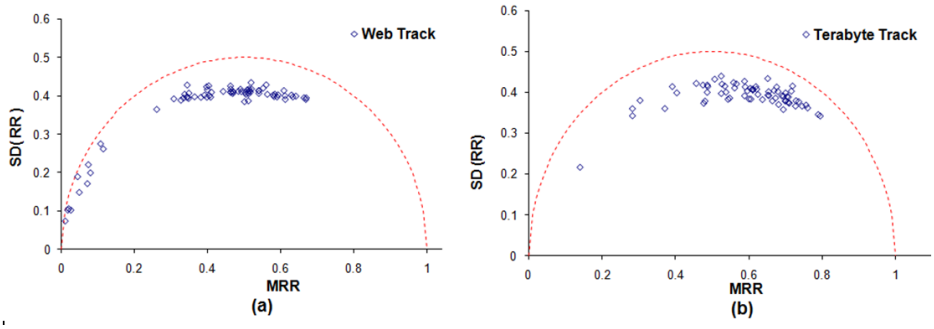

Fig. 3. MRR versus the standard deviation of RR values from: (a) runs participating in the Web track 2004, (b) runs participating in the Terabyte track 2004.

In order to overcome this issue, we consider functions that map values from [0,1] to $(-\infty, +\infty)$. We favor mappings that produce a symmetric distribution in the transformed space, akin to the normal distribution, if possible. We can then define the variability in effectiveness as the variance of the transformed values of a metric.

3.2 The Variability of Transformed Scores

We illustrate our approach by considering two transformations that have been used in IR and observe the properties of the transformed scores. The first is the *logit* function used by Cormak and Lynam [3] as a parametric estimate to deal with the asymmetric AP distribution. The logit is defined as: $logit(x) = log(\frac{x}{1-x})$ for x in (0,1). The boundary points, 0 and 1, are replaced by ϵ and $1-\epsilon$ respectively, for a small value of $\epsilon > 0$, and then transformed, letting $\epsilon \to 0$.

The second transformation is the *standardized score* or *z-score* whose use in IR was motivated by Webber et al. [13]. It is defined as $z = \frac{(x-\bar{x})}{\sigma}$, where x is a metric's score, e.g. an AP score. In addition, $\bar{x}$ and σ are the average and standard deviation of a set of scores measured across a set of retrieval systems

on a fixed topic. Hence, for a particular topic it is defined as

$$z = \frac{(AP - Mean(AP)_{systems})}{SD(AP)_{systems}} \tag{3}$$

In the following, we observe several properties of the logit(AP) and the standardized z-scores.

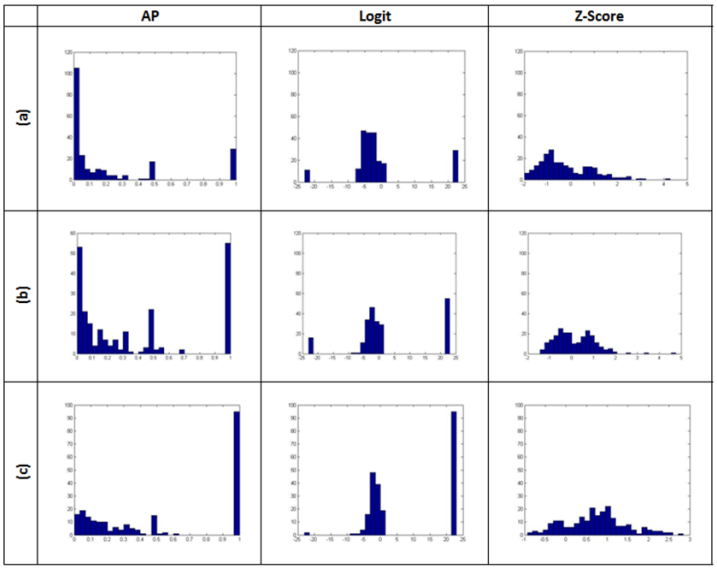

Fig. 4. The frequency distributions before and after transformation of three runs in Web track. (a) a run with a low MAP, (b) a run with a medium MAP, (c) a run with a high MAP.

Boundary Values and Score Distributions Figure 4 shows three runs of the Web track collection: (a) with a low MAP value, (b) with a medium MAP value, and (c) with a high MAP value. The distributions of the AP scores before and after both the logit and z-score transformations are presented as frequency histograms. The logit and z-score transformations differ significantly in the way they handle boundary values. The logit transformation transforms the boundaries to extreme values in the transformed space. This is observed by the extreme values at each end of the distributions in the middle column. In contrast, the z-score transformation disperses the boundaries smoothly as illustrated in the right column. In addition, the z-score transformation helps eliminate the source

of variance coming from topic difficulty[1] before measuring the variability of system effectiveness itself.

We now consider the variability in a system's effectiveness as the standard deviation of the transformed AP values. Let MLAP refer to the mean of the logit-transformed AP values and let MSAP refer to the mean of the standardized z-transformed AP values. Figure 5 shows the scatterplots of the standard deviations in transformed AP values as a function of their mean values, MLAP and MSAP. As seen in the figure, the logit and z-score transform the scores in different ranges. In addition, there is no longer a monotonic relationship between the values of mean and variability.

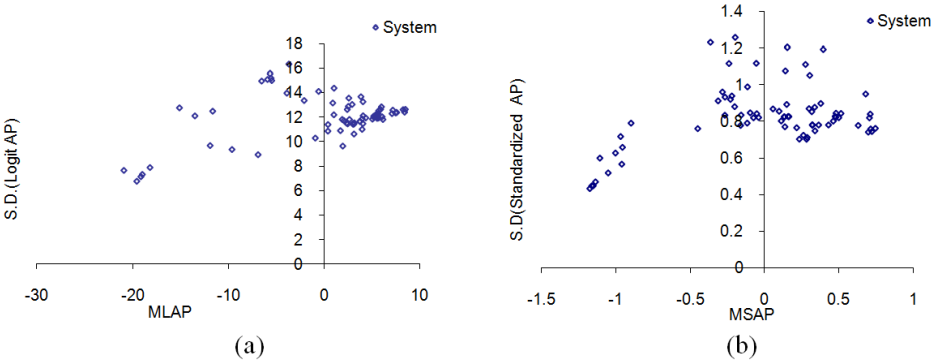

Fig. 5. Variability in effectiveness versus mean of transformed AP values: logit (a) and the z-score transformation (b).

3.3 Variability as a Tie Breaker

We consider all pairs of the top 75% (ordered by MAP) of systems participating in either the Robust or Web track of TREC 2004. We compare systems based on the mean of the standardized z-scores (MSAP). We use the paired t-test to measure the significance of MSAP differences. We set the significance level to 0.05. For all the ties we use the F-test and Levene's test to investigate the proportion of ties for which the variabilities in effectiveness's scores are significantly different.

As seen in Table 2 for the Robust track, 30% of pairs are considered ties, when using AP score, and 26% are considered ties in the transformed space. Interestingly, before transformation, the F-test cannot distinguish any statistical difference in variability, and the Levene's test can only break 11% of the ties. In contrast, after transformation into the z-space, the F-test can distinguish

[1] A topic is regarded as difficult if the range of effectiveness scores measured across a set of systems is small and near to zero.

Collections	Pairs	Status	Ties	Broken ties	
				F-test	Levene
Robust	3321	before transformation	997 (30%)	0 (0%)	106 (11%)
		after transformation	857 (26%)	280 (33%)	404 (47%)
Web	1485	before transformation	469 (31%)	1 (0.002%)	21 (0.04%)
		after transformation	415 (28%)	140 (34%)	158 (38%)

Table 2. The variability in effectiveness as a tie breaker: number of pairs, ties and broken ties in two tracks of TREC 2004.

between 33% and Levene's test can distinguish between 47% of the tied pairs. A similar effect before and after transformation is observed for the Web track.

3.4 The Effect of Topic Set Size on Measuring Variability in Effectiveness

If we are to use variability to characterize systems, it will be useful to know how many topics are needed to reliably compare two systems in terms of variability in effectiveness. Indeed, we will need to know how likely a decision would change if we compare systems using a different topic set. This performance variation across topic sets has previously been studied in the context of average performance [12]. We perform the same experiment to compare variabilities in systems' effectiveness.

```
foreach topic set size c from 10, 20, 30, ... , 100 {
  set the counters to 0;
   foreach TREC test collection t {
     foreach pair of systems A and B from track t {
       foreach trial from 1 to 50
         select two disjoint sets of topics X and Y of size c from t;
          if ( the difference between the variabilities is significant){
              d_X=SD(A,X)-SD(B,X);
              d_Y=SD(A,Y)-SD(B,Y);
              increment counter;
              if(d_X * d_Y < 0) {
                   increment swap counter; } } }}
     error-rate (c) =swap counter /counter;}
```

Fig. 6. Calculating error rates. SD(A, X) is the standard deviation of AP scores of system A measured on the topic Set X.

In our experiment, we first fix the topic set size, and then compute the variabilities in effectiveness of a pair of systems, A and B. Let us assume that System A is less volatile than System B based on this measurement. We then estimate the probability of a changed decision, i.e. finding System B to be less

volatile than System A. We estimate this probability by comparing the two systems across several trials that use different topic sets and then counting how many times the preference decision changes. Finally, to estimate the average probability of changing a decision, we repeat the process on different pairs of systems. This average probability (across systems) is called the *error rate*. The whole process is repeated for different sizes of topic sets.

The algorithm for computing the error rate is shown in Figure 6. It is based on the algorithm described in [12]. In our experiment, we compute the error rate for all the pairs regardless of their absolute differences. We run 50 different trials using different combinations of topics in the two disjoint topic sets. Furthermore, as Sanderson and Zobel [9] suggested, we only consider pairs with statistically significant differences in variability, as measured by Levene's test with a significance level of 0.05.

Once again we use the runs participating in the Robust track of TREC 2004 using topics 351-450 and 601-700 (199 topics), and the runs participating in the Web track (225 topics). In this experiment, only the top 75% of systems (ranked by MAP) are considered to prevent the poorly performing runs from having an effect on our conclusion [12]. Thus, our data collection consists of 135 runs and 4806 pairs of runs. Note that we transform AP scores using the z-score before measuring variability. The resulting error rate is shown in Figure 7. As expected, the curve shows that the error rate decreases as the topic set size increases. The experiment indicates that 90 topics are required to obtain an error rate less than 0.05. With 80 topics the measured error rate is 0.052 and with 90 topics it is 0.038.

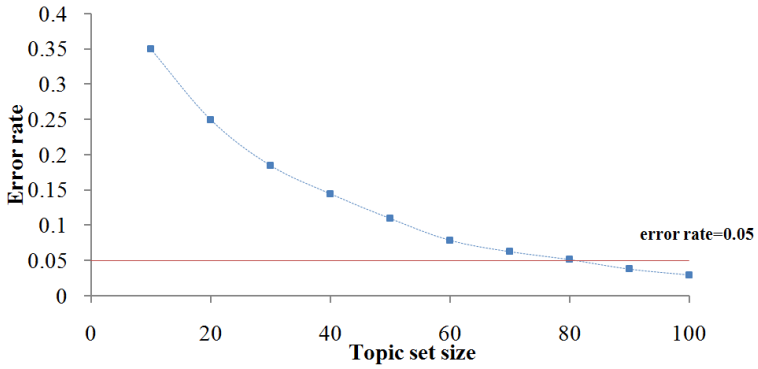

Fig. 7. Error rate versus topic set size using two TREC test collections: web track and robust track of TREC 2004.

4 Summary and Discussion

The average of effectiveness, measured across a topic set, does not capture all the important aspects of effectiveness and, used alone, may not be an informative measure of a system's effectiveness. We defined variability in effectiveness as the standard deviation of effectiveness scores measured across a set of topics. We proposed that a mean-variance graph helps demonstrate effectiveness in a two-dimensional space rather than ranking systems based on their average effectiveness. Our investigation revealed that the bounded values of a metric yield a curious phenomenon where values of average around 0.5 are accompanied with higher variances. We attributed this to the fact that the metric values fall within [0, 1]. This bounds the standard deviation of the scores to a semicircle as proven in Appendix A. Hence, retrieval systems with average effectiveness close to each of the two boundaries have smaller variances than those with average away from the boundaries. However, there might be also other reasons. For example, when the distribution is not symmetric, standard deviation cannot explain the dispersion properly. In Figure 4 it was shown that the distributions of AP scores were skewed toward the upper boundary, 1, and was completely asymmetric. We used two transformation methods to deal with this problem and showed how they differentiate systems effectiveness with the same average score. We finally discussed the minimum sample size required to estimate the variability in effectiveness. In our experiments we observed that 90 topics were required to obtain an error rate less than 0.05.

This paper only considered standard deviation as the measure of variability while it would be interesting to consider other measures, e.g. interquartile range and median absolute deviation. In addition, there are several ways to transform scores in a more symmetric space. For example, one might consider both logit and z-score transformation together. That is, the AP scores are first transformed by logit to $(-\infty , +\infty)$ and then z-score is used to deal with extreme values. We also note that the minimum sample size reported in Section 3.4 was averaged across different pairs of systems. As truly shown by Lin and Hauptmann [7], the minimum sample size varies across pairs of systems, and it depends on the difference between two systems' average effectiveness scores and corresponding variances.

Mean and variability can be used to evaluate retrieval systems. One may define a new metric as a function of both mean and variability. Such a metric helps rank systems' effectiveness in a one-dimensional space by considering both mean and variability in effectiveness. In addition, by a hypothetical scenario we showed that how a threshold of user satisfaction helps make preference between volatile and stable systems. However, we need to at least deal with two issues here. Firstly, in order to measure users' satisfaction we need to evaluate systems from users' perspective, i.e. directly asking users to express the amount of satisfaction. Such a user-oriented evaluation method provides accurate results but it is extremely expensive and difficult to do correctly. We can also model users' satisfaction by using implicit feedbacks of users, e.g. click-through data in a search engine query log. This method is less expensive but inaccurate. Sec-

ondly, users' satisfaction threshold may vary across queries. Indeed, the scenario described in Section 1 was simplified by considering the threshold as a constant value. However, in practice, the threshold varies across queries since it is highly depended on users' information needs and their expectation of the result set. We will consider these issues for future work.

Acknowledgements

The authors thank Jun Wang and Jianhan Zhu of UCL and Stephan Robertson of Microsoft Research Cambridge for useful discussion on earlier drafts of this paper.

Bibliography

[1] K. Collins-Thompson. Robust word similarity estimation using perturbation kernels. In *ICTIR '09: Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 265–272, Berlin, Heidelberg, 2009. Springer-Verlag.

[2] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 303–310, New York, NY, USA, 2007. ACM.

[3] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 533–540, New York, NY, USA, 2006. ACM.

[4] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, New York, NY, USA, 1993. ACM.

[5] C. T. Lee, V. Vinay, E. Mendes Rodrigues, G. Kazai, N. Milic-Frayling, and A. Ignjatovic. Measuring system performance and topic discernment using generalized adaptive-weight mean. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 2033–2036, New York, NY, USA, 2009. ACM.

[6] H. Levene. Robust test for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotteling*, pages 278–292, 1960.

[7] W.-H. Lin and A. Hauptmann. Revisiting the effect of topic set size on retrieval error. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 637–638, New York, NY, USA, 2005. ACM.

[8] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 525–532, New York, NY, USA, 2006. ACM.

[9] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, New York, NY, USA, 2005. ACM.

[10] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632, New York, NY, USA, 2007. ACM.

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

[12] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 316–323, New York, NY, USA, 2002. ACM.

[13] W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, New York, NY, USA, 2008. ACM.

Appendix A

Lemma: For all data sets like $X = \{x_1, x_2, ..., x_N\}$ where $0 \leq x_i \leq 1$, the corresponding *mean-standard deviation* values, $(\bar{X}, S_x)$, are confined within a semicircle with center (0.5, 0) and radius r=0.5:

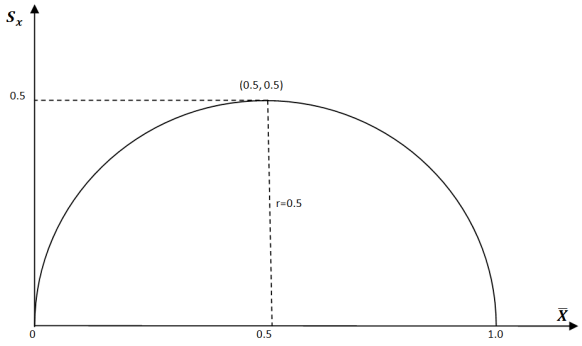

Fig. 8. The upper bound of standard deviation for scores bounded in 0 to 1.

$$(\bar{X} - \frac{1}{2})^2 + S_x^2 \leq (\frac{1}{2})^2;$$

$$\bar{X}^2 + S_x^2 \leq \bar{X} \tag{4}$$

Proof: with reference to the mean and variance:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

$$S_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2 \times \bar{X}(\frac{1}{N}\sum_{i=1}^{N} x_i) + \bar{X}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2$$

therefore:

$$\bar{X}^2 + S_x^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{6}$$

$x_i^2 \leq x_i$ because $0 \leq x_i \leq 1$; therefore:

$$\frac{1}{N}\sum_{i=1}^{N} x_i^2 \leq \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{X} \tag{7}$$

considering (6) and (7) together:

$$\bar{X}^2 + S_x^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \leq \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{X}$$

then we reach to (4):

$$\bar{X}^2 + S_x^2 \leq \bar{X}$$