

Revisiting the Effect of Topic Set Size on Retrieval Error

Wei-Hao Lin
Language Technologies Institute
School of Computer Science
5000 Forbes Ave
Pittsburgh PA 15213
U.S.A.
whlin@cs.cmu.edu

Alexander Hauptmann*
Language Technologies Institute
School of Computer Science
5000 Forbes Ave
Pittsburgh PA 15213
U.S.A.
alex@cs.cmu.edu

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software—*performance evaluation*

General Terms

Experimentation, Measurement

Keywords

Test Collections, Measurement Error

1. INTRODUCTION

Evaluating retrieval systems in a controlled environment with a large set of topics has been the core paradigm in the information retrieval community. Voorhees and Buckley proposed to estimate the reliability of retrieval experiments by calculating the probability of making wrong effectiveness judgments between two retrieval systems over two retrieval experiments[2], which is called Retrieval Experiment Error Rate (REER) in this paper. They have successfully shown how the topic set sizes affect the retrieval experiment reliability. However, the REER model in the previous work was empirically justified without providing a derivation based on statistical principles. We fill this gap and show that REER can indeed be derived from statistical principles. Based on the derived model we can explain why a successful experiment design depends on factors including a sufficient number of topics, large enough measurement score difference between systems, and a homogeneous distribution of retrieval scores for topics and systems, which reduces the variance of the score differences.

2. RETRIEVAL EXPERIMENT ERROR RATES

In a TREC-like retrieval experiment, a test document collection and a topic set \mathcal{T} of size $|\mathcal{T}|$ are given to participants, and participants run their retrieval system over the test collection and return rank lists for each topic. Human assessors

*This work was supported in part by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037.

manually identify the relevant documents for each topic, and an evaluation metric, for example, Mean Average Precision (MAP)¹ is calculated to objectively compare the effectiveness of the retrieval systems.

Now consider two retrieval systems A and B . Denote $A_1, A_2, \dots, A_{|\mathcal{T}|}$ and $B_1, B_2, \dots, B_{|\mathcal{T}|}$ as average precisions of System A and System B for each topic in the topic set \mathcal{T} , respectively. Each A_i is assumed to be sampled independently from an unknown but identical distribution F_A with mean μ_A and variance σ_A^2 , and B_i is sampled independently from another unknown but identical distribution F_B with mean μ_B and variance σ_B^2 . By definition MAP of System A is the average of $A_1, A_2, \dots, A_{|\mathcal{T}|}$, denoted as \bar{A} , and similarly the MAP of System B is \bar{B} . By the central limit theorem, \bar{A} and \bar{B} are approximately normally distributed,

$$\bar{A} \sim N(\mu_A, \frac{\sigma_A^2}{|\mathcal{T}|}) \quad (1)$$

$$\bar{B} \sim N(\mu_B, \frac{\sigma_B^2}{|\mathcal{T}|}) \quad (2)$$

The MAP difference between two system is a random variable, denoted as $D = \bar{X} - \bar{Y}$. Since \bar{X} and \bar{Y} are independent, it follows that D is normally distributed,

$$D \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2 + \sigma_Y^2}{|\mathcal{T}|}) \quad (3)$$

Now we can formalize REER proposed in [2] as the probability of the event that the results of two retrieval experiments are contradictory, i.e. the sign of the MAP difference in the first experiment D_1 is different from the sign of the MAP difference in the second experiment D_2 ,

$$\begin{aligned} \text{REER} &= \Pr(D_1 \times D_2 < 0) \\ &= \Pr(D_1 > 0, D_2 < 0) + \Pr(D_1 < 0, D_2 > 0) \end{aligned} \quad (4)$$

Since the two retrieval experiments are conducted independently, the joint probability of D_1 and D_2 is the product

¹Note that the derivation in the section is not restricted to MAP, and it applies to other metrics like Precision at 100.

²If a further assumption that F_A and F_B are normal is made, one can carry out the usual two-sample t -test procedure to compare if MAPs of two retrieval systems indeed differ. However, neither [2] nor we make this assumption.

of the individual event probabilities,

$$\begin{aligned} \text{REER} &= \Pr(D_1 > 0) \times \Pr(D_2 < 0) + \\ &\quad \Pr(D_1 < 0) \times \Pr(D_2 > 0) \\ &= (1 - \Pr(D_1 \leq 0)) \times \Pr(D_2 < 0) + \\ &\quad \Pr(D_1 < 0) \times (1 - \Pr(D_2 \leq 0)) \end{aligned} \quad (6)$$

$\Pr(D \leq 0)$ is the cumulative density function of D . From (3) we know D is normally distributed, and hence we can represent $\Pr(D \leq 0)$ in the standard normal cumulative density function Φ ,

$$\Pr(D \leq 0) = \Phi \left(\frac{-(\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{|\mathcal{T}|}}} \right) \quad (7)$$

Plug (7) back into (6), we finally obtain REER as follows,

$$\text{REER} = 2\Phi \left(\frac{-(\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{|\mathcal{T}|}}} \right) \left(1 - \Phi \left(\frac{-(\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{|\mathcal{T}|}}} \right) \right) \quad (8)$$

From (8) it can be easily shown that the range of REER fall between 0 and 0.5 for the Φ function ranges between 0 and 1.

2.1 Approximation

Voorhees and Buckley empirically fitted REER in the following model [2],

$$\text{REER} = b_1 \exp(-b_2 |\mathcal{T}|) \quad (9)$$

where b_1 and b_2 are two parameters. At the first sight the empirical model in (9) and our theoretically derived model in (8) bear no resemblance, but we will show that the empirical model in fact is an approximation of the theoretical model³

The theoretical REER (8) is not in closed form because of the integral in the standard normal cumulative density function Φ ,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp(-x^2/2) dx \quad (10)$$

$$= \frac{1}{2} (1 + \text{Erf}(\frac{z}{\sqrt{2}})) \quad (11)$$

where z is a standard normal random variable, and Erf is the so-called error function.

There have been efforts to approximate the Erf function in closed form, one of which is proposed by Williams[3] as follows,

$$\text{Erf}(x) \approx \sqrt{1 - \exp\left(\frac{-2x^2}{\pi}\right)} \quad (12)$$

By replacing the Erf function in (11) with the approximation in (12), the theoretic REER model in (8) can be approximated as follows,

$$\text{REER} \approx \frac{1}{2} \exp \left(-\frac{2}{\pi} \frac{(\mu_A - \mu_B)^2}{\sigma_A^2 + \sigma_B^2} |\mathcal{T}| \right) \quad (13)$$

If we compare the approximation in (13) with the empirical model in (9), they are clearly in exactly the same form. Therefore, we show that the empirical REER model proposed in [2] is indeed an approximation of the theoretical REER.

³Note that our goal here is not to approximate REER but to show the connection between the exponential form of the empirical REER model in (9) and the theoretical REER model in (8).

3. DISCUSSIONS

The theoretical REER derivation in (8) does not only explain why the empirical REER formula in (9) fitted the TREC evaluation results so well in [2], but also point out three important factors in designing successful retrieval experiments:

Sufficient number of topics By increasing the topic set size, i.e. $|\mathcal{T}|$, the theoretic REER model in (8) predicts that REER will decrease accordingly, that is, we can be more confident about the effectiveness judgments. This is consistent with the empirical findings in [2].

Large score differences If MAPs of two systems differ much, i.e. $\mu_X - \mu_Y$ is large, the theoretic REER model in (8) predicts that REER will be smaller. This explains why “a large enough difference between two effectiveness scores” is a general rule of thumb for acceptable experiment design [1].

Small score variances The last factor is the score variances, i.e. $\sigma_A^2 + \sigma_B^2$. The smaller the score variances, the lower the REER. Consequently when the topic difficulties vary much, the performance of a retrieval system will fluctuate greatly, resulting in bigger variance of the MAP difference. We estimate the variances of the MAP differences for TREC-3 and TREC-6 at the selected MAP difference levels⁴, as shown in Table 1. The variance in TREC-6 is larger than that in TREC-3, according to (8), REER in TREC-6 will be higher than that in TREC-3 at the same MAP difference level. This is consistent with the TREC participants’ impression that TREC-3 is easier than TREC-6 [2].

MAP Difference Level	Variances in TREC-3	Variances in TREC-6
$0.01 \leq \mu_X - \mu_Y < 0.02$	0.000464	0.000783
$0.02 \leq \mu_X - \mu_Y < 0.03$	0.000294	0.000814
$0.03 \leq \mu_X - \mu_Y < 0.04$	0.000434	0.000684

Table 1: The variances at different MAP difference level in TREC-3 and TREC-6.

4. REFERENCES

- [1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40. ACM Press, 2000.
- [2] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323. ACM Press, 2002.
- [3] J. D. Williams. An approximation to the probability integral. *The Annals of Mathematical Statistics*, 17(3):363–365, September 1946.

⁴Past TREC evaluation results can be found at <http://trec.nist.gov/results.html>.