Evaluating Evaluation Measures with Worst-Case Confidence Interval Widths

Tetsuya Sakai
Waseda University
tetsuyasakai@acm.org

ABSTRACT

IR evaluation measures are often compared in terms of rank correlation between two system rankings, agreement with the users' preferences, the swap method, and discriminative power. While we view the agreement with real users as the most important, this paper proposes to use the Worst-case Confidence interval Width (WCW) curves to supplement it in test-collection environments. WCW is the worst-case width of a confidence interval (CI) for the difference between any two systems, given a topic set size. We argue that WCW curves are more useful than the swap method and discriminative power, since they provide a statistically well-founded overview of the comparison of measures over various topic set sizes, and visualise what levels of differences across measures might be of practical importance. First, we prove that Sakai's ANOVA-based topic set size design tool can be used for discussing WCW instead of his CI-based tool that cannot handle large topic set sizes. We then provide some case studies of evaluating evaluation measures using WCW curves based on the ANOVA-based tool, using data from TREC and NTCIR.

CCS CONCEPTS

•**Information systems → Retrieval effectiveness;**

KEYWORDS

ANOVA; confidence intervals; effect sizes; evaluation measures; *p*-values; sample sizes; statistical significance

1 INTRODUCTION

IR systems are built to satisfy users' information needs, but it is not practical to make the users evaluate the systems all the time for the purpose of improving them—that would *annoy* the users, not satisfy them! Hence, we often turn to IR evaluation measures in laboratory experiments. But which IR measures are *good*?

In laboratory studies, evaluation measures are often compared in terms of *rank correlation* between two system rankings (e.g. [9]), agreement with the users' document preferences (e.g. [7]), the swap method (e.g. [8]), and discriminative power (e.g. [3, 4]). Since IR evaluation measures are often regarded as surrogates of *user satisfaction* or *user performance* measurements, we view the agreement with users as the most important, although it needs to be said that user preference studies often use hired assessors such as crowd workers intead of real users with an information need. Moreover, studies involving human assessors obviously incur costs.

To supplement user-based studies of IR evaluation measures, we propose to use *Worst-case Confidence interval Width* (WCW) curves in test-collection environments. WCW is the worst-case width of a confidence interval (CI) for the difference between any two systems, given a topic set size. We argue that WCW curves are more useful than the swap method and discriminative power, since they provide a statistically well-founded overview of the comparison of measures over various topic set sizes, and visualise what levels of differences across measures might be of practical importance. To this end, we leverage one of the publicly available *topic set size design* Excel tools of Sakai [6]. First, we prove that Sakai's ANOVA-based topic set size design tool[1] can be used for discussing WCW instead of his CI-based tool[2] that cannot handle large topic set sizes (See Section 2). We then provide some case studies of evaluating evaluation measures using WCW curves based on the ANOVA-based tool, using data from TREC and NTCIR.

2 PRIOR ART IN EVALUATING EVALUATION MEASURES

When a new IR evaluation measure is invented, a system ranking according to this measure (averaged over a set of topics) is often compared with another according to a well-established measure; *rank correlation* measures such as Kendall's τ or the top-heavy τ_{ap} [9] are often used to quantify the similarity between two rankings. However, this approach cannot tell us whether a measure is good or bad, due to the lack of a "correct" system ranking. It merely tells us whether a new measure is similar to an existing one or not; it only serves as a sanity check.

For a given query, a user sees two Search Engine Result Pages (SERPs) side by side, and says that $SERP_1$ is better than $SERP_2$ ("$SERP_1 > SERP_2$"). If an evaluation measure also says "$SERP_1 > SERP_2$," this is a *preference agreement*; if it says "$SERP_1 < SERP_2$," this is a preference disagreement. We can count the number of agreements over different queries and SERP pairs, and use it for comparing the "goodness" of evaluation measures. In practice, this approach also has a few limitations: (a) the judges employed in the preference assessments are often not real search engine users with an information need; (b) human assessments can be unreliable and/or inconsistent; and (c) hiring judges comes at a cost, no matter how small.

The *swap method* [8] may be used to measure the consistency (i.e., "preference agreement with itself") of evaluation measures across different topic sets. Given a set of n topics, the set is split in half, and the number of inconsistent preferences (e.g., $SERP_1 > SERP_2$ with the first half but $SERP_1 < SERP_2$ with the second half) is counted, using different systems and different splits. As this method can

EVIA 2017, co-located with NTCIR-13, Tokyo, Japan.

[1] http://www.f.waseda.jp/tetsuya/CIKM2014/samplesizeANOVA.xlsx
[2] http://www.f.waseda.jp/tetsuya/FIT2014/samplesizeCI.xlsx

only consider half the original topic set size, Voorhees and Buckley used a simple extrapolation method to estimate what will happen for topic set sizes larger than n. However, estimating the swap rate for (say) $n = 100$ topics based on observations with (say) $n = 10, 25, 50$ topics may not be reliable. To directly consider the size n, *bootstrap samples* [3] can be used to replace the sampling-without-replacement approach of Voorhees and Buckley, but this method cannot consider topic set sizes larger than n either.

Given a set of runs and an evaluation measure, a p-value can be obtained for every system pair using an appropriate significance test, and the sorted p-values can be plotted against the system pairs [3, 4]: this is called the *discriminative power* curve. While highly discriminative measures are useful in the sense that they can obtain more statistically significant results in a given environment with exactly n topics, discriminative power does not provide a view over different choice of topics. Moreover, it is not clear, for example, a measure with 90% discriminative power should actually be preferred over one with 80% discriminative power.

Sakai [6] released three Excel tools based on *topic set size design*, which determines the number of topics n to create for a new test collection given a set of statistical requirements. His ANOVA-based tool takes the following as input: α (Type I error probability), β (Type II error probability), m (the number of systems to be compared in one-way ANOVA), $\hat{\sigma}^2$ (an estimate of the within-system variance for a particular evaluation measure), and *minD* (*minimum detectable range*); the tool returns the topic set size n that ensures $100(1-\beta)\%$ *statistical power* whenever the true difference between the best and the worst among the m systems is *minD* or larger. Whereas, his CI-based tool takes the following as input: α, $\hat{\sigma}_t^2$ (an estimate of the variance of the between-system differences in terms of a particular evaluation measure), and δ, which is exactly what we call WCW in this study; the tool returns the topic set size n that ensures that the width of the $100(1-\alpha)\%$ CI for any system pair is no larger than δ. Following Sakai, we simply let $\hat{\sigma}_t^2 = 2\hat{\sigma}^2$ for any evaluation measure.

While the relationship between *minD* for ANOVA and n can be plotted for different evaluation measures, this seems problematic as a way to compare evaluation measures, since, for example, a *minD* of 0.1 in term of one measure is not equivalent to a *minD* of 0.1 in terms of another. In contrast, if we plot δ against n, this is probably a more valid comparison since, at least for any normalised measures that lie in the $[0, 1]$ score range, we usually want the CI width to be as small as possible. This is why we propose to plot δ against topic set sizes to compare different measures. However, Sakai's CI-based tool cannot handle large topic set sizes: the limitation of his CI-based tool is due to that of Excel's `GAMMA` function: `GAMMA(172)` is greater than 10^{307} and cannot be computed [6]. Hence, we start by proving that his ANOVA-based tool can be used instead of the less robust CI-based one, for IR researchers to compare the statistical reliability of evaluation measures based on WCW.

3 PROOF THAT ANOVA-BASED TOPIC SET SIZE DESIGN CAN BE USED INSTEAD OF CI-BASED ONE

According to Sakai's CI-based topic set size design, the initial topic set size estimate for ensuring that the CI width for the difference in means for any two systems is no larger than $\delta(>0)$ is given by [6]:

$$n_{CI} = \frac{4\{z_{inv}(\alpha/2)\}^2\hat{\sigma}_t^2}{\delta^2} = \frac{4\{z_{inv}(\alpha/2)\}^2(2\hat{\sigma}^2)}{\delta^2}, \quad (1)$$

where $z_{inv}(P)$ is the upper z-value[3] for probability P. Subsequently, this estimate is incremented until it actually satisfies the requirement (α, δ). Thus, while the actual CI relies on a t-distribution, the method starts off with a standard normal distribution by assuming that the variance estimate $\hat{\sigma}_t^2$ is perfectly accurate[4]. This is why Eq. 1 involves a z-value rather than a t-value.

Whereas, according to Sakai's ANOVA-based topic set size design, the initial topic set size estimate for ensuring $100(1-\beta)\%$ statistical power whenever the true difference between the best and the worst systems is *minD* or larger is given by [6]:

$$n_{ANOVA} = \frac{2\hat{\sigma}^2\lambda}{minD^2}, \quad (2)$$

where λ is a noncentrality parameter of a noncentral χ^2 distribution with $\phi = m - 1$ degrees of freedom; as discussed below, linear formulae are available for estimating λ from ϕ [2]. As Eq. 2 is based on a series of approximations, n_{ANOVA} is then incremented until it actually satisfies the requirement $(\alpha, \beta, minD, m)$.

Sakai [6] observed that, for the data he considered, "*the topic set size required based on the CI-based design with $\alpha = 0.05$ and $\delta = c$ is almost the same as the topic set size required based on the ANOVA-based design with $(\alpha, \beta, m) = (0.05, 0.20, 10)$ and $minD = c$, for any c.*" We analytically explain and generalise his observation as follows. From Eqs. 1 and 2, we have:

$$\frac{n_{ANOVA}}{n_{CI}} = \frac{\lambda\delta^2}{4\{z_{inv}(\alpha/2)\}^2 minD^2} = \frac{\lambda}{4\{z_{inv}(\alpha/2)\}^2}\left(\frac{\delta}{minD}\right)^2. \quad (3)$$

Here, note that $4\{z_{inv}(\alpha/2)\}^2$ is a constant for a given α; also, λ is a constant given α, β and m. Figure 1 visualises the relationship between the two constants for $\alpha = 0.01, 0.05$ and $\beta = 0.10, 0.20$, while varying the number of systems m. The linear formulae for approximating λ based on $\phi = m - 1$ [6] are provided in the bottom half of the figure. Figure 1 shows that

$$\lambda \approx 4\{z_{inv}(\alpha/2)\}^2 \quad (4)$$

holds when:

Condition (a) $\alpha = 0.05, \beta = 0.20, m = 10$; or
Condition (b) $\alpha = 0.05, \beta = 0.10, m = 5$; or
Condition (c) $\alpha = 0.01, \beta = 0.20, m = 18$; or
Condition (d) $\alpha = 0.01, \beta = 0.10, m = 10$.

Hence, whenever one of the above four conditions holds true, then from Eqs. 3 and 4 we obtain:

$$\frac{n_{ANOVA}}{n_{CI}} \approx \left(\frac{\delta}{minD}\right)^2. \quad (5)$$

Thus, when one of the above four conditions holds, by letting $\delta = minD$ in Eq. 5 we obtain $n_{ANOVA}/n_{CI} \approx 1$, that is, $n_{ANOVA} \approx n_{CI}$, regardless of the variance estimate $\hat{\sigma}^2$. *Q.E.D.*

Henceforth, we only consider the popular *Cohen's five-eighty convention* [1], i.e., $(\alpha, \beta) = (0.05, 0.20)$[5], and leverage **Condition (a)**

[3] `NORM.S.INV`$(1-P)$ with Microsoft Excel.
[4] Replacing the true population variance of a standard normal distribution with a sample variance constitutes the very definition of a t-distribution.
[5] Note that "eighty" refers to the statistical power: $100(1-\beta)\%$.

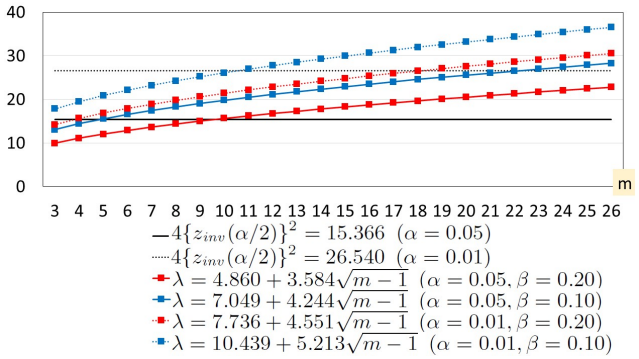

Figure 1: The noncentrality parameter λ vs. $4\{z_{inv}(\alpha/2)\}^2$

Table 1: $\hat{\sigma}^2$: estimates of within-system variances. *md* stands for measurement depth (i.e., document cutoff).

(a) TREC03-04Robust (*md* = 1000), from Sakai [6]			
AP	.0471	nDCG	.0456
Q	.0465	nERR	.1145
(b) TREC11-12WebAdhoc (*md* = 10), from Sakai [6]			
AP	.0824	nDCG	.0441
Q	.0368	nERR	.0863
(c) TREC11-12WebDiversity (*md* = 10), from Sakai [6]			
α-nDCG	.0779	D-nDCG	.0340
nERR-IA	.0842	D#-nDCG	.0504
(d) NTCIR-12 STC1C (*md* = 10), from Sakai [5]			
nG@1	.1144	std-AB nG@1	.0193
P+	.0943	std-AB P+	.0186
nERR	.0867	std-AB nERR	.0182

mentioned above. Figure 2 compares, for different and quite extreme values of the variance estimate $\hat{\sigma}^2$, the topic set size curve using the CI-based tool with $\alpha = 0.05$ and one using the ANOVA-based tool with $\alpha = 0.05, \beta = 0.20, m = 10$. Due to the aforementioned limitation of the CI-based tool, it was not possible to obtain the entire curves with this tool. On the other hand, it is clear that the ANOVA-based curves can serve as highly accurate surrogates for the CI-based curves *and* can handle large topic set sizes. In summary, to discuss WCW, we can always use the more robust ANOVA-based tool and treat the *minD* values as if they are δ values.

4 WCW-BASED EVALUATION OF EVALUATION MEASURES: CASE STUDIES

Having proven that the ANOVA-based tool can be used instead of the less robust CI-based tool, we now demonstrate how different evaluation measures can be compared using WCW curves obtained with the ANOVA-based tool.

Table 1 shows the variance estimates $\hat{\sigma}^2$ of various evaluation measures reported in the literature [5, 6]. For the purpose of the present study, the knowledge of each evaluation measure is not necessary; the measures with a prefix "std-AB" denote *standardised* measures of the original measures, where the raw score for each topic is transformed based on a set of known systems, to absorb the hardness of that topic as well as its variation across systems [5]. Given a topic-by-run score matrix for a particular evaluation measure, $\hat{\sigma}^2$ can easily be obtained as the residual variance of ANOVA. While some evaluation measures are substantially less stable across topics than others (e.g., Compare nERR and nDCG in Table 1(a)), it is not clear just from this table how such differences will actually impact our evaluation results.

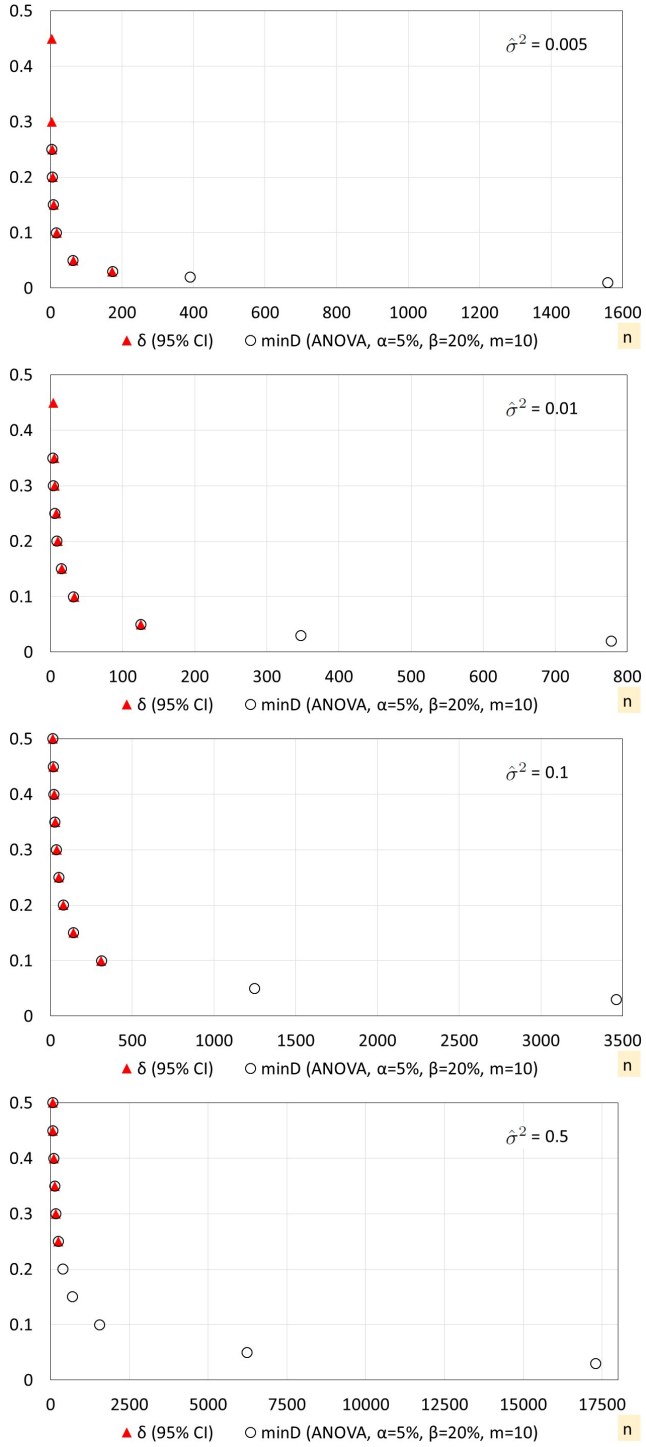

Figure 2: The actual relationship between δ for CI and *minD* for ANOVA in topic set size design.

Figure 3 shows the WCW curves that correspond to the variances shown in Table 1, for $\alpha = 0.05$, i.e., 95% CIs. For each evaluation

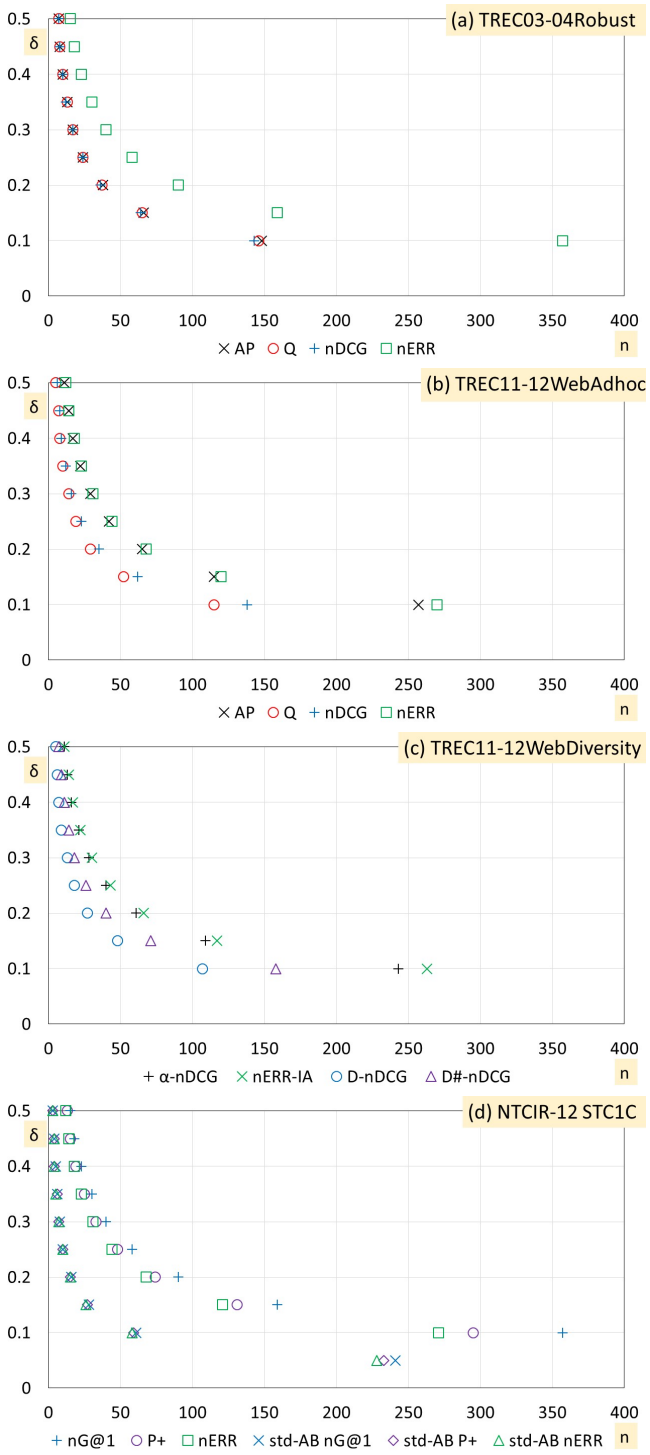

Figure 3: WCW curves for 95% CIs.

measure, the δ is plotted against the required topic set size n; the curve was obtained by entering different values of *minD* (i.e., δ) into the ANOVA-based tool (with $\alpha = 0.05, \beta = 0.20, m = 10$) and recording the resultant n. The advantages of the proposed WCW-based comparison of evaluation measures are as follows:

- Unlike discriminative power and the swap method, we can easily consider a wide range of topic set sizes;
- For a particular topic set size, we can easily compare across different evaluation measures, since an evaluation measure with a small WCW is usually more desirable than one with a large WCW under the same condition;
- The WCW curves can visualise the differences among measures *that practically matter*.

For example, from Figure 3(b), when the topic set size is $n = 50$, it is clear that the WCW of nDCG and that of Q are about the same (around 0.16), while those of AP and nERR are substantially larger (around 0.23). Similarly, from Figure 3(d), while it is clear that the *standardised* ("std-AB") measures have substantially lower WCW values than the unstandardised ones, the differences within the set of standardised measures are probably not of practical importance, as indicated by the near-perfect overlaps of the curves.

5 CONCLUSIONS AND FUTURE WORK

We proposed to evaluate evaluation measures by comparing the WCW for various topic set sizes, using an existing ANOVA-based tool instead of the less robust CI-based tool. We proved the relationship between these two topic set size design methods, and demonstrated the advantages of WCW curves over well-known methods such as the swap test and discriminative power. It is hoped that this method will supplement user-based studies of evaluation measures.

REFERENCES

[1] Paul D. Ellis. 2010. *The Essential Guide to Effect Sizes.* Cambridge University Press.
[2] Yasushi Nagata. 2003. *How to Design the Sample Size (in Japanese).* Asakura Shoten.
[3] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics based on the Bootstrap. In *Proceedings of ACM SIGIR 2006.* 525–532.
[4] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of ACM SIGIR 2007.* 71–78.
[5] Tetsuya Sakai. 2016. The Effect of Score Standardisation on Topic Set Size Design. In *Proceedings of AIRS 2016 (LNCS 9994).* 16–28.
[6] Tetsuya Sakai. 2016. Topic Set Size Design. *Information Retrieval Journal* 19, 3 (2016), 256–283. http://link.springer.com/content/pdf/10.1007%2Fs10791-015-9273-z.pdf
[7] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do User Preferences and Evaluation Measures Line Up?. In *Proceedings of ACM SIGIR 2010.* 555–562.
[8] Ellen M. Voorhees and Chris Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of ACM SIGIR 2002.* 316–323.
[9] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *Proceedings of ACM SIGIR 2008.* 587–594.