Topic Set Size Design with Variance Estimates from Two-Way ANOVA

Tetsuya Sakai
Waseda University, Japan.
tetsuyasakai@acm.org

ABSTRACT

Recently, Sakai proposed two methods for determining the topic set size n for a new test collection based on variance estimates from past data: the first method determines the minimum n to ensure high statistical power [22], while the second method determines the minimum n to ensure tight confidence invervals [23]. These methods are based on statistical techniques described by Nagata [15]. While Sakai [22] used variance estimates based on one-way ANOVA, Sakai [23] used the 95% percentile method proposed by Webber, Moffat and Zobel [38]. This paper reruns the experiments reported by Sakai [22, 23] using variance estimates based on two-way ANOVA [17], which turn out to be slightly larger than their one-way ANOVA counterparts and substantially larger than the percentile-based ones. If researchers should choose to "err on the side of over-sampling" as recommened by Ellis [10], the variance estimation method based on two-way ANOVA and the results reported in this paper are probably the ones researchers should adopt. We also establish empirical relationships between the two topic set size design methods, and discuss the balance between n and the pool depth pd using both methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

confidence intervals; effect sizes; evaluation; measures; sample sizes; statistical significance; test collections

1. INTRODUCTION

Recently, Sakai proposed two methods for determining the topic set size n for a new test collection based on variance estimates from past data: the first method determines the minimum n to ensure high statistical power [22], while the second method determines the minimum n to ensure tight confidence invervals (CIs) [23]. These methods are based on statistical techniques described by Nagata [15]. While Sakai [22] used variance estimates based on one-way ANOVA, Sakai [23] used the 95% percentile method proposed by Webber, Moffat and Zobel [38]. This paper reruns the experiments reported by Sakai [22, 23] using variance estimates based on two-way ANOVA [17], which turn out to be slightly larger than their one-way ANOVA counterparts and substantially larger than the percentile-based ones. If researchers should choose to "err on the side of over-sampling" as recommened by Ellis [10], the variance estimation method based on two-way ANOVA and the results reported in this paper are probably the ones researchers should adopt. We also establish empirical relationships between the two topic set size design methods, and discuss the balance between n and the pool depth pd using both methods.

The remainder of this paper is organised as follows. Section 2 discusses related work. Section 3 describes two topic set size design methods: the first method determines n based on power analysis [22], while the second one determines n based on CIs [23]. Section 4 describes three variance estimation methods, as variance estimates are required for topic set size design: the 95% percentile method of Webber *et al.* [38] later adopted by Sakai [23]; the one-way ANOVA-based method used by Sakai [22]; and an alternative method based on two-way ANOVA statistics, which we introduce in this paper. Section 5 reports on our new experiments using variance estimates based on the two-way ANOVA statistics and reruns the topic set size design experiments of Sakai [22, 23]. Finally, Section 6 concludes this paper.

2. RELATED WORK

2.1 Webber/Moffat/Zobel

Webber *et al.* [38] proposed procedures for building a test collection based on power analysis. They recommend adding topics and conducting relevance assessments incrementally while examining the achieved *statistical power* (i.e., the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true) and re-estimating the standard deviation σ_t of between-system performance differences. They considered the comparison of two systems only and therefore adopted the t-test; they did not address the problem of the *family-wise error rate* [3, 10]: if a pairwise t-test is conducted independently for m systems with a significance level of α (i.e., the probability of rejecting H_0 when it is true), the probability of detecting at least one nonexistent between-system difference amounts to $1-(1-\alpha)^{m(m-1)/2}$. Their experiments focused on Average Precision (AP), a binary-relevance evaluation measure. In order to estimate the standard deviation σ_t (or equivalently, the variance σ_t^2), they used a 95%-percentile method with existing data; this will be described in Section 4.1.

2.2 Sakai

While the aforementioned method by Webber *et al.* [38] assumed that information retrieval (IR) researchers can iteratively sample new topics and conduct relevance assessments while checking the achieved power and reestimating variances in order to compare a given pair of systems, Sakai [22, 23] addressed an arguably more practical question: "I want to build a new test collection. How many topics should I prepare?" His aim is to answer this question directly and simply at the beginning of test collection design.

Sakai extended the work of Webber *et al.* in several ways: (a) He used not only the t-test but also one-way ANOVA with power analysis to consider the problem of ensuring high statistical power when comparing $m \geq 2$ systems [22]; he also proposed the alternative approach of setting the topic set size n to ensure tight CIs [23]; (b) He considered a variety of graded relevance evaluation measures, including those for search result diversification, and demonstrated that evaluation measures should be chosen at the test collection design phase as variances are heavily dependent on the choice of evaluation measure; (c) He adopted a time-honoured method for estimating σ^2, the performance variance of any system (from which σ_t^2, the variance of the between-system score *differences*, may be deduced) based on one-way ANOVA statistics, and performed variance pooling across multiple test collections [22]. His topic set size design tools based on the t-test, one-way ANOVA, and the CI are publicly available[1]; they can be used to determine n provided that an estimate of σ^2 is available.

While Sakai [23] estimated σ_t^2 using the 95%-percentile method of Webber *et al.* for his CI-based topic set size design method, Sakai [22] estimated σ^2 (and σ_t^2) using the aforementioned statistics from one-way ANOVA. In the present study, we repeat all of their experiments using variance estimates based on two-way ANOVA statistics, which turn out to be slightly larger than the one-way ANOVA counterparts and substantially larger than the 95%-percentile values.

2.3 Other Related Work

The IR community in the twentieth century was rather reluctant to conduct parametric significance testing; however, nowadays it is known that the t-test is relatively robust to assumption violations and applicable to IR evaluation [25]. Computer-based alternatives to classical significance testing, namely, the *bootstrap* [20, 28] and the *randomisation test* [29, 32] are also available. However, while several research disciplines have gone beyond significance testing and standardised the use of CIs and *effect sizes* (ESs) [9, 10], a similar *statistical reform* [11, 25] is yet happen in IR. That is, the use of CIs, power, and ESs are rather limited even though statistical significance alone is not as informative as is commonly assumed [3]. Exceptions include the work of Nelson [16] who stressed the importance of power and ESs in IR evaluation in 1989; that of Carterette and Smucker [5] who considered power analysis with the sign test for AP; that of Smucker and Clarke [30] who discussed ESs for their Time-Biased Gain measure.

The Generalisability Theory (GT) has also been shown to be useful for assessing the test collection reliability [1, 4, 33]: while both the GT approach and ours rely on variance estimates from past data, Urbano, Marrero and Martín [33] point out that the reliability indicators obtained from GT are difficult to interpret. We leave the comparison of our methods with the GT approach for the purpose of topic set size design as future work. Yet other alternatives to classical significance testing include Killeen's p_{rep} [14] and the Baysian approach to hypothesis testing [2, 13], but these are also beyond the scope of this study.

Besides the above statistically motivated studies, topic-splitting heuristics have been used in the literature to answer the following question: "I have n topics: are $n/2$ topics good enough for predicting what will happen with the other $n/2$ topics?" (e.g. [21, 36, 37, 39]). Sakai [20] showed that his *discriminative power* method using bootstrap tests can provide results similar to topic splitting while using the full n topics; this method has been used by several researchers for comparing evaluation measures (e.g. [12, 19, 26, 31]). The topic set size design methods we examine in this study may also be regarded as alternative ways to assess evaluation measures: they translate the statistical stability of measures into practical significance, namely, the assessment cost [22, 23].

3. TOPIC SET SIZE DESIGN METHODS

3.1 Determining n based on Power Analysis

Sakai [22] released two simple Excel tools which allow researchers to determine n based on power analysis. The first tool is based on the paired t-test and concerns comparisons of $m = 2$ systems; the second is based on one-way ANOVA and concerns comparisons of $m \geq 2$ systems. It was demonstrated that when $m = 2$, the two tools give very similar results, with the ANOVA tool giving slightly higher estimates. The present study uses the one-way ANOVA tool as this is more general than the t-test version. The input to the ANOVA-based tool are as follows:

α, β: The probability of Type I error α and that of Type II error β [24]. The Excel tool contains four sheets for $(\alpha, \beta) = (0.01, 0.10), (0.01, 0.20), (0.05, 0.10), (0.05, 0.20)$.

m: The number of systems that will be compared $(m \geq 2)$.

$minD$: The *minimum detectable range* [22]. That is, whenever the performance difference between the best and the worst systems is $minD$ or higher, we want to ensure a power of β given the significance level of α.

$\hat{\sigma}^2$: The estimated variance of a system's performance, under the *homoscedasticity* (i.e., equal variance) assumption [3, 22]. That is, it is assumed that the scores of the i-th system obey $N(\mu_i, \sigma^2)$ where σ^2 is common to all systems. This variance is heavily dependent on the evaluation measure.

3.2 Determing n based on Confidence Intervals

Sakai [23] released a simple Excel tool which allow researchers to determine n based on the tightness of a CI between any two systems. This method is closely related to the paired t-test and considers $m = 2$ only. The input to the CI-based tool are:

α One minus the *confidence level*. This α is the same α used in significance testing, and is usually set to $\alpha = 0.05$ in order to achieve 95% confidence [9].

δ The upperbound we impose on the width of any CI. That is, we want the CI for any system pair to be no wider than δ.

$\hat{\sigma}_t^2$ The estimated variance of the performance difference between Systems X and Y. The paired t-test and the analogous CI are based on the assumption that the system scores obey $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively, where $\mu_\bullet$ and $\sigma_\bullet^2$ are the population means and variances. It then follows that the performance difference obeys $N(\mu_X - \mu_Y, \sigma_t^2)$ where $\sigma_t^2 = \sigma_X^2 + \sigma_Y^2$. Hence, when the aforementioned per-system variance estimate $\hat{\sigma}^2$ is available under the homoscedasticity assumption, a reasonable estimate of σ_t^2 may be $\hat{\sigma}_t^2 = 2\hat{\sigma}^2$ [22].

4. VARIANCE ESTIMATION METHODS

4.1 95% Percentile Method

Here we describe the method used by Webber *et al.* [38] and later adopted by Sakai [23] for estimating σ_t^2 from a given matrix of evaluation scores with n_C topics and m_C systems, where C represents an existing test collection. For each of the $k =$

[1] http://www.f.waseda.jp/tetsuya/tools.html

Table 1: TREC test collections and runs used for estimating σ^2. The web track relevance grades [6, 7] were mapped to our relevance levels as follows: -2 and 0 →L0 (i.e., nonrelevant); 1 →L1; 2 →L2; 3 →L3; 4 →L4.

short name	track	topics	runs	pool depth	relevance levels	documents
(a) task: adhoc/news						
TREC03new	2003 robust	50 (601-650)	78	125	L0-L2	the Congressional Record)
TREC04new	2004 robust	49 (651-700 minus 672)	78∗	100	L0-L2	528,155 (disks 4+5 minus
(b) task: adhoc/web						
TREC11w	2011 web - ad hoc	50	37	25	L0-L3	approx. one billion
TREC12w	2011 web - ad hoc	50	28	20/30	L0-L4	(clueweb09)
(c) task: diversity/web						
TREC11wD	2011 web - diversity	50 (same as TREC11w)	25	25	L0-L3 per intent	approx. one billion
TREC12wD	2011 web - diversity	50 (same as TREC12w)	20	20/30	L0-L4 per intent	(clueweb09)

∗ TREC 2004 description-only runs excluded (the set of runs used by Webber, Moffat and Zobel [38])

$m_C(m_C-1)/2$ system pairs ($b=1,\ldots,k$), the method first computes an unbiased estimate of the population variance of the between-system difference in terms of a particular measure:

$$V^b_C = \frac{\sum_{j=1}^{n_C}(d^b_{C,j}-\bar{d}^b_C)^2}{n_C-1}\ , \qquad (1)$$

where $d^b_{C,j}$ is the difference between X and Y for the j-th topic and the b-th system pair, and $\bar{d}^b_C$ is the mean difference for the same pair. The k values are then sorted, and the 95th percentile is taken to be $\hat{\sigma}^2_t$.

4.2 One-way ANOVA-based Method

The 95% percentile method obtains $\hat{\sigma}^2_t$ directly from observed data, but there are time-honoured methods for estimating population variances in statistics. Sakai [22] used a method that uses one-way ANOVA statistics, since his topic set size design method was also based on one-way ANOVA. Let x_{ij} denote the performance score for the i-th system with topic j ($i=1,\ldots,m$ and $j=1,\ldots,n$); let $\bar{x}_{i\bullet}=\frac{1}{n}\sum_{j=1}^{n}x_{ij}$ (sample system mean) and $\bar{x}=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}$ (sample grand mean). In one-way ANOVA, the total variation $S_T=\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2$ is decomposed into between-system and within-system variations S_A and S_{E1} (i.e., $S_T=S_A+S_{E1}$), where

$$S_A = n\sum_{i=1}^{m}(\bar{x}_{i\bullet}-\bar{x})^2\ ,\ \ S_{E1}=\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i\bullet})^2\ . \qquad (2)$$

Furthermore , let:

$$V_A = S_A/\phi_A\ ,\ \ V_{E1}=S_{E1}/\phi_{E1}\ , \qquad (3)$$

where $\phi_A=m-1, \phi_{E1}=m(n-1)$. Then $F_0=V_A/V_{E1}$ is the test statistic for one-way ANOVA. Using the above basic statistics, the variance σ^2 can be estimated as follows [17]:

$$\hat{\sigma}^2 = \frac{m-1}{mn}(V_A-V_{E1})+V_{E1}\ . \qquad (4)$$

Here, $\frac{m-1}{mn}(V_A-V_{E1})$ is an estimate of the population between-system variance σ^2_A; and V_{E1} is an estimate of the population within-system variance σ^2_{E1}. These estimates are often used for estimating the population ESs for one-way ANOVA [17].

4.3 Two-way ANOVA-based Method

In this study, we explore an alternative time-honoured method for estimating σ^2. While one-way ANOVA decomposes the total variation S_T into between-system and within-system variations S_A and S_{E1}, two-way ANOVA (without replication) decomposes S_T into between-system, between-topic, and residual variations, S_A, S_B and S_{E2} (i.e., $S_T=S_A+S_B+S_{E2}$) [25], by utilising the fact that the scores $x_{\bullet j}$ correspond to one another. Let $\bar{x}_{\bullet j}=\frac{1}{m}\sum_{i=1}^{m}x_{ij}$ (sample topic mean) and

$$S_B = m\sum_{j=1}^{n}(\bar{x}_{\bullet j}-\bar{x})^2\ ,\ \ S_{E2}=\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i\bullet}-\bar{x}_{\bullet j}+\bar{x})^2\ . \qquad (5)$$

Furthermore, let:

$$V_B = S_B/\phi_B\ ,\ \ V_{E2}=S_{E2}/\phi_{E2}\ , \qquad (6)$$

where $\phi_B=n-1, \phi_{E2}=(m-1)(n-1)$. ($V_A$ is as defined in Eq. 3.) Then V_A/V_{E2} and V_B/V_{E2} are the test statistics for two-way ANOVA. Using the above basic statistics, the variance σ^2 can be estimated as follows [17]:

$$\hat{\sigma}^2 = \frac{m-1}{mn}(V_A-V_{E2})+\frac{1}{m}(V_B-V_{E2})+V_{E2}\ . \qquad (7)$$

Here, $\frac{m-1}{mn}(V_A-V_{E2})$ is an estimate of the population between-system variance σ^2_A (*cf.* Eq. 4); $\frac{1}{m}(V_B-V_{E2})$ is an estimate of the population between-topic variance σ^2_B; and V_{E2} is an estimate of the residual variance σ^2_{E2}. These estimates are often used for estimating the population ESs for two-way ANOVA. In this study, we use Eq. 7 to obtain $\hat{\sigma}^2$ (and $\hat{\sigma}^2_t=2\hat{\sigma}^2$) for topic set size design, while comparing the outcome with the results of Sakai [22, 23].

4.4 Pooled Variances

In each experiment, we pool variances obtained from multiple existing test collections. Let $\sigma^2_{t,C}$ be the 95%-percentile obtained using Eq. 1. Then σ^2_t may be estimated as follows:

$$\hat{\sigma}^2_t = \sum_C(n_C-1)\hat{\sigma}^2_{t,C}/\sum_C(n_C-1)\ . \qquad (8)$$

The ANOVA-based estimates $\hat{\sigma}^2$ are also pooled using the same formula.

5. EXPERIMENTS

Table 1 provides some statistics of the past data we used for obtaining variance estimates [22, 23]. We consider three IR *tasks*: (a) adhoc news retrieval; (b) adhoc web search; and (c) diversified web search; for each task, we used two data sets to obtain pooled variance estimates. The adhoc/news data sets are from the TREC robust tracks, with "new" topics from each year [34, 35]. The web data sets are from the TREC web tracks [6, 7]. While we considered the measurement depths of $md=10,1000$ for adhoc/news, we considered $md=10$ only for the web tasks as we are interested in the quality of the *first* search engine result page.

The actual variance depends on the evaluation measure and conditions associated with it. Table 2 shows the evaluation measures considered in this study. For the adhoc/news and adhoc web tasks, we consider the binary Average Precision (AP), Q-measure (Q), normalised Discounted Cumulative Gain (nDCG) and normalised Expected Reciprocal Rank (nERR), all computed using the

Table 2: Evaluation measures used in this study.

task type	measure	used in tasks such as	tool
adhoc	AP	TREC adhoc/robust	`NTCIREVAL`
	Q	NTCIR CLIR/IR4QA/GeoTime	`NTCIREVAL`
	nDCG	TREC web adhoc	`NTCIREVAL`
	nERR	TREC web adhoc	`NTCIREVAL`
diversity	α-nDCG	TREC web diversity	`ndeval`
	nERR-IA	TREC web diversity	`ndeval`
	D-nDCG	NTCIR INTENT	`NTCIREVAL`
	D♯-nDCG	NTCIR INTENT	`NTCIREVAL`

Table 3: $\hat{\sigma}^2$ for different evaluation measures with measurement depth md, obtained from two-way ANOVA statistics.

			$\hat{\sigma}^2$			
(a1) task: adhoc/news ($md = 1000$)						
Data	m	n	AP	Q	nDCG	nERR
TREC03new	78	50	.0543	.0548	.0569	.1214
TREC04new	78	49	.0517	.0527	.0559	.1201
Pooled	-	-	**.0530**	**.0538**	**.0564**	**.1208**
(a2) task: adhoc/news ($md = 10$)						
Data	m	n	AP	Q	nDCG	nERR
TREC03new	78	50	.0970	.0711	.0785	.1282
TREC04new	78	49	.0824	.0668	.0779	.1260
Pooled	-	-	**.0898**	**.0690**	**.0782**	**.1271**
(b) task: adhoc/web ($md = 10$)						
Data	m	n	AP	Q	nDCG	nERR
TREC11w	37	50	.0912	.0499	.0571	.1061
TREC12w	28	50	.0840	.0275	.0360	.0762
Pooled	-	-	**.0876**	**.0387**	**.0466**	**.0912**
(c) task: diversity/web ($md = 10$)						
Data	m	n	α-nDCG	nERR-IA	D-nDCG	D♯-nDCG
TREC11wD	25	50	.0898	.0950	.0418	.0639
TREC12wD	20	50	.0768	.0843	.0331	.0453
Pooled	-	-	**.0833**	**.0897**	**.0375**	**.0546**

`NTCIREVAL` toolkit[2]. For the diversity/web task, we consider α-nDCG and Intent-Aware nERR (nERR-IA) computed using `ndeval`[3], as well as D-nDCG and D♯-nDCG computed using `NTCIREVAL`. When using `NTCIREVAL`, the gain value for each Lx-relevant document was set to $g(r) = 2^x - 1$: for example, the gain for an L3-relevant document is 7, while that for an L1-relevant document is 1. As for `ndeval`, the default settings were used: this program ignores per-intent graded relevance levels.

5.1 Variance Estimates

Table 3 shows the variance estimates we obtained for each task and evaluation measure, using the two-way ANOVA-based method described in Section 4.3. Figure 1 visualises the pooled variance estimates in this table, and compares them with the pooled estimates obtained using the 95%-percentile method (Section 4.1) and the one-way ANOVA-based method (Section 4.2). It can be observed that the ANOVA-based estimates are substantially larger than the 95%-percentile method; the two ANOVA-based methods yield very similar results, with the two-way ANOVA-based one giving marginally larger values. This subtle difference will only affect estimates for very large topic set sizes. Since larger variances imply larger topic sets, we use the pooled two-way ANOVA-based estimates for topic set size design, choosing to possibly "err on the side of over-sampling" as recommended by Ellis [10]. In the following sections, we discuss how our new variance estimates affect the results previously reported by Sakai [22, 23].

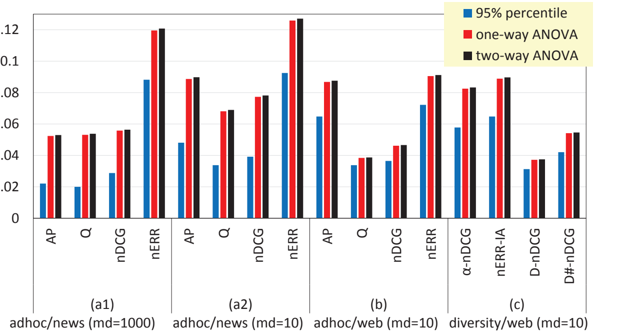

Figure 1: Comparison of the three variance methods with the pooled variances.

5.2 Results based on Power Analysis

Tables 4-7 show the required topic set sizes for different IR tasks under different power-based requirements for $m = 10, 100$ systems, based on the two-way ANOVA-based pooled variances shown in Table 3. These tables can be compared with Tables 8-11 from Sakai [22] who used one-way ANOVA-based estimates. The **boldface** values are those under *Cohen's five-eighty convention* [8], i.e., $(\alpha, \beta) = (0.05, 0.20)$. From Tables 4 and 5 (adhoc/news with $md = 1000, 10$), we can observe that:

- As nERR is substantially less stable than AP, Q and nDCG (See Table 3(a1) and (a2)), it requires many more topics than the other measures. For example, Table 4(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $md = 1000$, nERR requires 975 topics while AP, Q and nDCG require only 428, 435, 456 topics, respectively[4]. Throughout Table 4, nERR is more than twice as expensive as AP, Q and nDCG.

- As reducing md causes higher variances (Compare Table 3(a1) and (a2)), this also means we need more topics. More importantly, while the advantage of utilising the graded relevance assessments with Q and nDCG is not clear when $md = 1000$ (a typical TREC ad hoc setting), it is clear when $md = 10$. For example, Table 5(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $md = 10$, AP requires 725 topics, while Q and nDCG require only 557 and 631 topics, respectively. As for nERR, it requires 1,026 topics[5].

- In Table 4(I), the topic set sizes for AP, Q and nDCG under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.20, 10)$ are 42, 43, and 45, respectively[6]. Thus, a typical TREC adhoc/news test collection with $n = 50$ topics is good enough for guaranteeing a minimum detectable range of 0.20 in terms of AP, Q, and nDCG for comparing $m = 10$ systems under Cohen's five-eighty convention.

From Table 6 (adhoc/web with $md = 10$), we can observe that:

- As Q and nDCG are substantially more stable than AP and nERR for this task (See Table 3(b)), they require substantially fewer topics. For example, Table 6(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $md = 10$, AP and nERR require 707 and 736 topics, while Q and

[2] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`. For computing AP and Q, we follow Sakai and Song [27] and divide by $\min(md, R)$ rather than by R (i.e., the number of relevant documents) in order to properly handle small measurement depths.

[3] `http://trec.nist.gov/data/web/12/ndeval.c`

[4] The corresponding one-way ANOVA-based estimates of n are 965, 423, 429, and 451 [22].

[5] The corresponding one-way ANOVA-based estimates of n are 716, 550, 624, and 1,016 [22].

[6] The corresponding one-way ANOVA-based estimates of n are 42, 42, 44 [22].

Table 4: Topic set size table ($m = 10, 100$) for adhoc/news ($md = 1000$) with AP/Q/nDCG/nERR.

α	$minD$	$\beta = .10$	$\beta = .20$
(I) $m = 10$			
.01	.02	6920/7024/7364/15771	5659/5745/6022/12898
	.05	1108/1125/1179/2524	906/920/964/2065
	.10	278/282/295/632	227/231/242/517
	.20	70/71/75/159	58/59/61/130
	.25	45/46/48/102	37/38/40/84
.05	.02	5257/5336/5594/11981	**4127/4190/4392/9406**
	.05	842/854/894/1917	**661/671/703/1506**
	.10	211/214/224/480	**166/168/176/377**
	.20	53/54/57/120	**42/43/45/95**
	.25	34/35/36/77	**27/28/29/61**
(II) $m = 100$			
.01	.02	16492/16741/17550/37588	14000/14211/14898/31909
	.05	2639/2679/2809/6015	2241/2275/2384/5106
	.10	660/670/703/1504	561/569/597/1277
	.20	166/168/176/377	141/143/150/320
	.25	106/108/113/241	90/92/96/205
.05	.02	13040/13237/13876/29720	**10688/10849/11374/24360**
	.05	2087/2118/2221/4756	**1711/1737/1820/3898**
	.10	522/530/556/1189	**428/435/456/975**
	.20	131/133/139/298	**108/109/114/244**
	.25	84/85/89/191	**69/70/74/157**

Table 5: Topic set size table ($m = 10, 100$) for adhoc/news ($md = 10$) with AP/Q/nDCG/nERR.

α	$minD$	$\beta = .10$	$\beta = .20$
(I) $m = 10$			
.01	.02	11724/9009/10210/16593	9588/7367/8350/13570
	.05	1877/1442/1634/2656	1535/1180/1337/2172
	.10	470/361/409/665	385/296/335/544
	.20	118/91/103/167	97/75/85/137
	.25	76/59/66/107	62/48/55/88
.05	.02	8906/6843/7756/12605	**6993/5373/6090/9897**
	.05	1426/1096/1242/2017	**1120/860/975/1584**
	.10	357/274/311/505	**281/216/244/397**
	.20	90/69/78/127	**71/55/62/100**
	.25	58/44/50/81	**46/35/40/64**
(II) $m = 100$			
.01	.02	27942/21470/24333/39548	23720/18226/20656/33573
	.05	4471/3436/3894/6328	3796/2917/3306/5372
	.10	1118/860/974/1583	950/730/827/1344
	.20	280/215/244/396	238/183/207/337
	.25	180/138/156/254	153/117/133/216
.05	.02	22093/16976/19240/31270	**18109/13915/15770/25630**
	.05	3535/2717/3079/5004	**2898/2227/2524/4101**
	.10	884/680/770/1251	**725/557/631/1026**
	.20	222/170/193/313	**182/140/158/257**
	.25	142/109/124/201	**117/90/102/165**

nDCG require only 313 and 377 topics, respectively[7]. Throughout Table 6, AP and nERR are more than twice as expensive as Q.

As noted by Sakai [22], the advantage of Q and nDCG over AP as demonstrated in both Table 5 (adhoc/news) and Table 6 (adhoc/web) strongly suggests the importance of utilising graded relevance assessments when the measurement depth md is small. On the other hand, when md is large, how many relevant documents have been retrieved, and at what positions, probably outweigh whether each document is highly or partially relevant.

From Table 7 (diversity/web with $md = 10$), we can observe that:

- As D-nDCG and D♯-nDCG are substantially more stable than α-nDCG and nERR-IA (See Table 3(c)), they require substantially fewer topics. For example, Table 7(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $md = 10$, D-nDCG and D♯-nDCG require only 303 and 441 topics, while α-nDCG and nERR-IA require as many as 673 and 724 topics, respectively[8].

Table 6: Topic set size table ($m = 10, 100$) for adhoc/web ($md = 10$) with AP/Q/nDCG/nERR.

α	$minD$	$\beta = .10$	$\beta = .20$
(I) $m = 10$			
.01	.02	11437/5053/6084/11907	9353/4133/4976/9738
	.05	1831/809/974/1906	1497/662/797/1559
	.10	458/203/244/477	375/166/200/391
	.20	115/51/62/120	95/42/51/98
	.25	74/33/40/77	61/28/33/63
.05	.02	8688/3839/4622	**6821/3014/3629/7102**
	.05	1391/615/740/1448	**1092/483/581/1137**
	.10	348/154/186/362	**274/121/146/285**
	.20	88/39/47/91	**69/31/37/72**
	.25	56/25/30/59	**44/20/24/46**
(II) $m = 100$			
.01	.02	27258/12042/14501/28378	23139/10223/12310/24090
	.05	4362/1927/2321/4541	3703/1636/1970/3855
	.10	1091/482/581/1136	926/410/493/964
	.20	273/121/146/285	232/103/124/242
	.25	175/78/94/182	149/66/80/155
.05	.02	21552/9522/11465/22438	**17665/7805/9398/18391**
	.05	3449/1524/1835/3591	**2827/1249/1504/2943**
	.10	863/381/459/898	**707/313/377/736**
	.20	216/96/115/225	**177/79/95/185**
	.25	139/62/74/144	**114/51/61/118**

Table 7: Topic set size table ($m = 10, 100$) for diversity/web ($md = 10$) with α-nDCG/nERR-IA/D-nDCG/D♯-nDCG.

α	$minD$	$\beta = .10$	$\beta = .20$
(I) $m = 10$			
.01	.02	10875/11711/4896/7129	8894/9577/4005/5830
	.05	1741/1874/784/1141	1424/1533/642/934
	.10	436/469/197/286	357/384/161/234
	.20	110/118/50/72	90/97/41/59
	.25	70/76/32/47	58/62/27/38
.05	.02	8262/8896/3720/5415	**6487/6985/2921/4252**
	.05	1322/1424/596/867	**1039/1118/468/681**
	.10	331/357/149/217	**260/280/118/171**
	.20	83/90/38/55	**66/71/30/43**
	.25	54/58/24/35	**42/46/20/28**
(II) $m = 100$			
.01	.02	25920/27911/11669/16990	22004/23694/9906/14423
	.05	4148/4466/1868/2719	3521/3792/1586/2308
	.10	1037/1117/467/680	881/949/397/578
	.20	260/280/117/171	221/238/100/145
	.25	167/179/75/109	142/152/64/93
.05	.02	20494/22069/9226/13433	**16798/18089/7563/11011**
	.05	3280/3532/1477/2150	**2688/2895/1211/1762**
	.10	820/883/370/538	**673/724/303/441**
	.20	206/221/93/135	**169/182/76/111**
	.25	132/142/60/87	**108/116/49/71**

- In Table 7(II), the number of topics required required by D-nDCG is $n = 49$ under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.25, 100)$. Thus, a typical TREC diversity/web test collection with $n = 50$ topics is good enough for guaranteeing a minimum detectable range of 0.25 in terms of D-nDCG for comparing $m = 100$ systems under Cohen's convention. On the other hand, α-nDCG, nERR-IA and D♯-nDCG do not pass the test as the required topic set sizes are 108, 116, and 71, respectively[9].

Figure 2 visualises the relationship between the required topic set size (n) and the number of systems to be compared (m) under $(\alpha, \beta, minD) = (0.05, 0.20, 0.10)$. For example, Figure 2(b) shows that, if we expect to compare $m = 200$ adhoc/web sys-

[7] The corresponding one-way ANOVA-based estimates of n are 701, 731, 310 and 373 [22].

[8] The corresponding one-way ANOVA-based estimates of n are 301, 438, 666, and 718.

[9] The corresponding one-way ANOVA-based estimates of n are 49, 107, 115, and 71[22].

tems[10] under the above requirements, AP and nERR would require 964 and 1,004 topics, while Q and nDCG would require only 426 and 513 topics. Similarly, Figure 2(c) shows that, if we expect to compare $m = 200$ diversity/web systems under the above requirements, α-nDCG and nERR-IA would require 917 and 987 topics, while D-nDCG and D♯-nDCG would require only 413 and 601 topics.

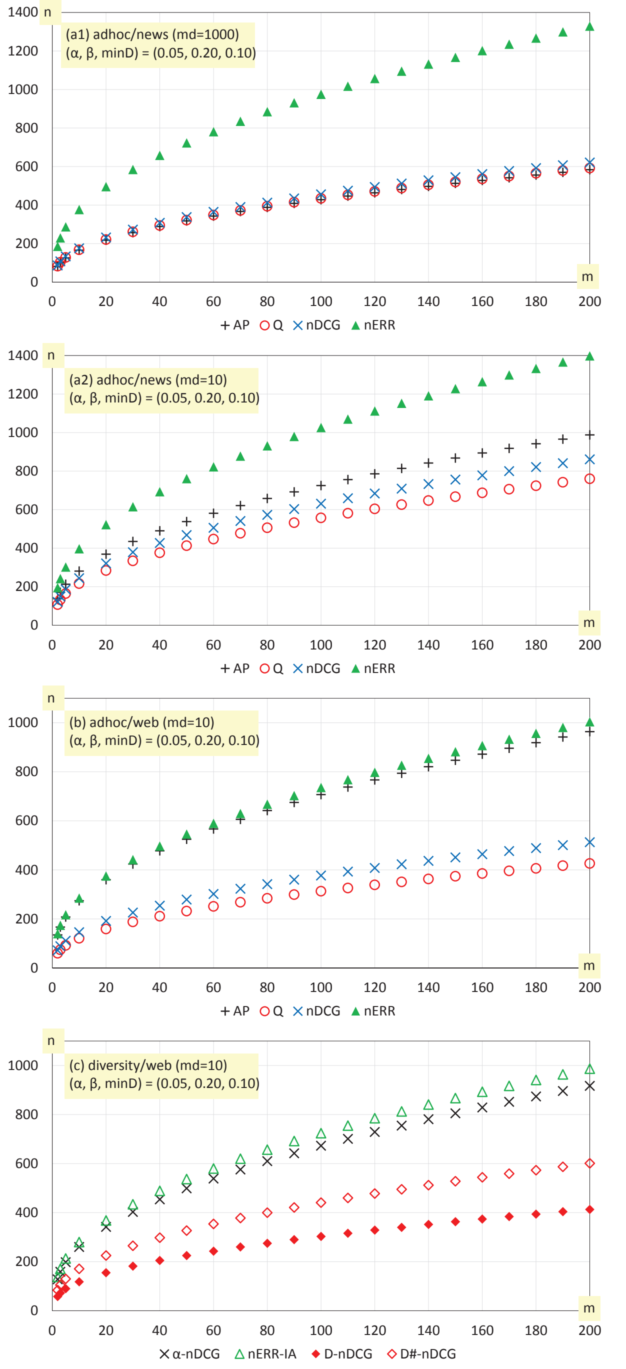

Figure 2: Required number of topics n against the number of systems m, with $(\alpha, \beta, minD) = (0.05, 0.20, 0.10)$.

Table 8: Topic set sizes for achieving tight CIs at $\alpha = 0.05$.

(a1) task: adhoc/news ($md = 1000$)				
δ	AP	Q	nDCG	nERR
.10	165	168	176	-
.15	75	76	79	167
.20	43	44	46	95
.25	29	29	30	62
(a2) task: adhoc/news ($md = 10$)				
δ	AP	Q	nDCG	nERR
.10	278	214	243	-
.15	125	97	109	176
.20	71	55	63	100
.25	47	36	41	65
(b) task: adhoc/web ($md = 10$)				
δ	AP	Q	nDCG	nERR
.10	272	121	146	283
.15	122	55	66	127
.20	70	32	38	73
.25	46	22	25	47
(c) task: diversity/web ($md = 10$)				
δ	α-nDCG	nERR-IA	D-nDCG	D♯-nDCG
.10	258	278	118	170
.15	116	125	54	77
.20	66	71	31	44
.25	43	47	21	29

5.3 Results based on Confidence Intervals

Table 8 shows the topic set sizes required to ensure tight CIs, based on two-way ANOVA variance estimates. The table can be compared with Table 3 from Sakai [23] who used 95% percentile estimates. For example, $\delta = 0.10$ means that the CI of any between-system difference is given by $\bar{d} \pm 0.05$ or something narrower, where $\bar{d}$ denotes the mean difference. For a few cells, we could not compute the gamma function with the Excel tool as n was too large ($n > 343$). It can be observed that:

- Q outperforms the other measures in Tasks (a2) and (b) in terms of the required topic set size, while nERR consistently underperforms the others. For example, in (b) (adhoc/web), given $(\alpha, \delta) = (0.05, 0.10)$, Q requires only 121 topics, while AP and nERR require 272 and 283 topics, respectively[11].
- In Task (c), D-nDCG requires substantially fewer topics than α-nDCG and nERR-IA. For example, given $(\alpha, \delta) = (0.05, 0.10)$, D-nDCG requires only 118 topics, while α-nDCG and nERR-IA require 258 and 278 topics, respectively[12].

Figure 3 shows the relationship between our power-based and CI-based topic set size design methods. Note that Sakai [22, 23] did not discuss this. In each graph, the CI upperbound δ and the minimum detectable range $minD$ for the ANOVA-based power analysis are plotted against the required topic set size n. The top graph uses $\hat{\sigma}^2 = .0690$, which is the estimated variance for Q in Table 3(a2); the bottom graph uses $\hat{\sigma}^2 = 0.375$, which is the estimated variance for D-nDCG in Table 3(c). These are the smallest variances in each IR task. It can be observed that the power-based curve with $m = 10$ almost completely overlaps with the CI-based one. That is, requiring that $(\alpha, \beta, minD, m) = (0.05, 0.20, c, 10)$ based on the power-based method is equivalent to requiring that $(\alpha, \delta) = (0.05, c)$ based on the CI-based method, for any c. Also, the bottom graph shows, for stable measures such as D-nDCG, that when we have $n = 50$ topics (as in a TREC diversity task), the

[10] This setting is not unrealistic. For example, the TREC 2011 Microblog Track received 184 runs from 59 participating teams [18].

[11] The corresponding 95%-percentile estimates of n are 106, 202 and 224 [23].

[12] The corresponding 95%-percentile estimates of n are 98, 180 and 202 [23].

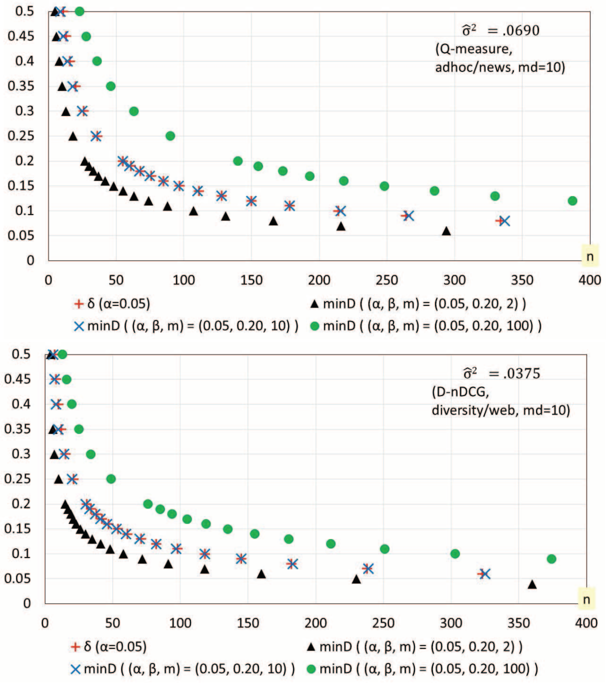

Figure 3: The relationship between δ and $minD$.

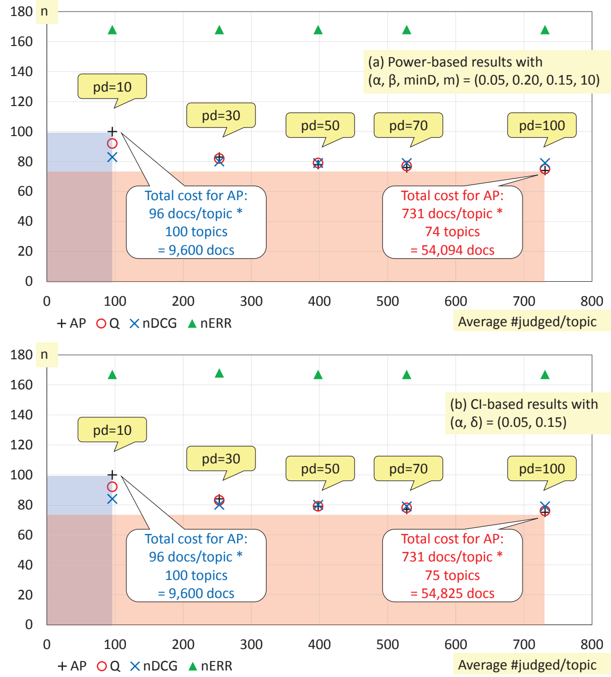

Figure 4: Required number of topics n against the average number of documents judged per topic for a given pool depth pd, for adhoc/news ($md = 1000$).

minimum detectable range $minD$ for $m = 2$ systems (i.e., the minimum detectable *difference* between two systems [22]) is about 0.10; the CI upperbound δ and the $minD$ for $m = 10$ systems are both about 0.15; and the $minD$ for $m = 100$ systems is about 0.25, i.e., one-quarter of the range of a normalised evaluation measure.

5.4 Cost Analysis via Pool Depth Reduction

Previous sections discussed adhoc/news, adhoc/web and diversity/web search tasks, but assumed that the pool depth pd was a given. In this section, we focus our attention on the adhoc/news task with $md = 1000$, where we have depth-125 and depth-100 pools (See Table 1), which enables us to consider shallower pools [22, 23]. From the original TREC03new and TREC04new relevance assessments, we created depth-pd ($pd = 100, 90, 70, 50, 30, 10$) versions of the relevance assessments by filtering out all topic-document pairs that were not contained in the top pd documents of any run. Using each set of the depth-pd relevance assessments, we re-evaluated all runs using AP, Q, nDCG and nERR. Then, using these new topic-by-run matrices, new pooled variance estimates were obtained using the two-way ANOVA-based method.

Table 9 shows the pooled variance estimates obtained from the depth-pd versions of the TREC03new and TREC04new relevance assessments. It also shows the average number of documents judged per topic for each pd. For example, while the original depth-125 relevance assessments for TREC03new contain 47,932 topic-document pairs, the depth-100 version has 37,605 pairs across 50 topics; the original TREC04new depth-100 relevance assessments have 34,792 pairs across 49 topics. Hence, on average, $(37,605 + 34,792)/(50 + 49) = 731$ documents are judged per topic when $pd = 100$. Similarly, $(4,905+4,581)/(50+49) = 96$ documents are judged per topic when $pd = 10$. We assume that the average number of documents judged is a constant for a given pd, though in reality it depends on the number and the diversity of runs besides pd.

Based on the power-based method, Figure 4(a) plots the required topic set size n against the average number of documents judged per topic, for $minD = 0.15$ with $m = 10$ under Cohen's five-eighty convention. Similarly, based on the CI-based method, Figure 4(b) plots n against the average number of documents judged per topic, for $\delta = 0.15$. Again, it can be observed that when $m = 10$, setting $minD$ for the power-based method is almost equivalent to setting δ for the CI-based method, and therefore that the two graphs are almost identical. The two graphs show that, if we construct a topic set with $n = 100$ topics with depth-10 pools and use AP, this ensures a minimum detectable range $minD$ of 0.15 when we compare up to $m = 10$ systems, as well as CI widths no greater than $\delta = 0.15$. This design is statistically equivalent to having $n = 75$ (or $n = 74$) topics with depth-100 pools. On average, the first design would require $96 * 100 = 9,600$ relevance judgments, while the second design would require $731 * 75 = 54,825$ relevance judgments, which is 5.7 times as expensive. Thus, these graphs visualise what is well-known: it is better to have many topics with few judgments than to have few topics with many judgments (e.g. [4, 5, 38]). It can also be observed that, because the variance of nERR is very high regardless of the pool depth, it is always about twice as costly as the other evaluation measures.

6. CONCLUSIONS

We reran the experiments reported by Sakai [22, 23] using variance estimates based on two-way ANOVA [17], which turned out to be slightly larger than their one-way ANOVA counterparts and substantially larger than the percentile-based counterparts. If researchers should choose to "err on the side of over-sampling" as recommened by Ellis [10], the variance estimation method based on two-way ANOVA and the results reported in this paper are probably the ones researchers should adopt. We demonstrated that, using variance estimates from existing data, the topic set size n

Table 9: Number of relevance assessments and pooled $\hat{\sigma}^2$ for reduced pool depths with adhoc/news (measurement depth $md = 1000$).

Pool depth	TREC03new #judged	TREC04new #judged	Average	Pooled $\hat{\sigma}^2$			
pd	for 50 topics	for 49 topics	judged/topic	AP	Q	nDCG	nERR
125	47,932	-	-	-	-	-	-
100	37,605	34,792	731	.0530	.0537	.0562	.1208
70	27,816	24,491	528	.0546	.0550	.0563	.1208
50	20,839	18,612	398	.0561	.0562	.0565	.1208
30	13,045	11,968	253	.0596	.0589	.0570	.1209
10	4,905	4,581	96	.0714	.0658	.0594	.1206

can be determined to ensure high power and/or tight CIs; as different evaluation measures have different variances, evaluation measures should be chosen at the test collection design phase [22, 23]. Our pool depth reduction experiments with the power-based and CI-based topic set size design methods have shown that having $n = 75$ topics with depth-100 pools is about 5.7 times as costly as having $n = 100$ topics with depth-10 pools, even though these two designs are statistically equivalent. In addition, by comparing the power-based and CI-based results, we showed that requiring $(\alpha, \beta, minD, m) = (0.05, 0.20, c, 10)$ based on the power-based method is equivalent to requiring that $(\alpha, \delta) = (0.05, c)$ based on the CI-based method, for any c. In practice, researchers can choose from a set of statistically equivalent test collection designs based on the available budget, in order to maximise reusability.

All of our experiments are reproducible: the topic-by-run matrices are available at `http://www.f.waseda.jp/tetsuya/data.html` (Use the `CIKM2014PACK`), and the topic set size design tools are available at `http://www.f.waseda.jp/tetsuya/tools.html`. Our techniques can easily be applied to non-IR tasks as well, starting with some data that are equivalent to topic-by-run matrices.

Acknowledgements

This research was supported by Waseda University Grants for Special Research Projects (2014A-026, 2014B-181, 2014S-077) and by Microsoft Research (Waseda University's project name: "Taxonomising and Evaluating Web Search Engine User Behaviours").

7. REFERENCES

[1] D. Bodoff and P. Li. Test theory for assessing IR test collections. In *Proceedings of ACM SIGIR 2007*, pages 367–374, 2007.

[2] B. Carterette. Model-based inference about IR systems. In *ICTIR 2011 (LNCS 6931)*, pages 101–112, 2011.

[3] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1), 2012.

[4] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proceedings of ACM SIGIR 2008*, pages 651–658, 2008.

[5] B. Carterette and M. D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of ACM CIKM 2007*, pages 643–652, 2007.

[6] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 web track. In *Proceedings of TREC 2011*, 2012.

[7] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 web track. In *Proceedings of TREC 2012*, 2013.

[8] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (Second Edition)*. Lawrence Erlbaum Associates, 1988.

[9] G. Cumming. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge, 2012.

[10] P. D. Ellis. *The Essential Guide to Effect Sizes*. Cambridge University Press, 2010.

[11] F. Fidler, C. Geoff, B. Mark, and T. Neil. Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics*, 33:615–630, 2004.

[12] E. Kanoulas and J. A. Aslam. Empirical justification of the gain and discount function for nDCG. In *Proceedings of ACM CIKM 2009*, pages 611–620, 2009.

[13] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[14] P. R. Killeen. An alternative to null hypothesis significance tests. *Psychological Science*, 16:345–353, 2005.

[15] Y. Nagata. *How to Design the Sample Size (in Japanese)*. Asakura Shoten, 2003.

[16] M. J. Nelson. Statistical power and effect size in information retrieval experiments. In *Proceedings of CAIS/ASCI'98*, pages 393–400, 1998.

[17] M. Okubo and K. Okada. *Psychological Statistics to Tell Your Story: Effect Size, Confidence Interval (in Japanese)*. Keiso Shobo, 2012.

[18] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 microblog track. In *Proceedings of TREC 2011*, 2012.

[19] S. E. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. In *Proceedings of ACM SIGIR 2010*, pages 603–610, 2010.

[20] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532, 2006.

[21] T. Sakai. On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, 43:531–548, 2007.

[22] T. Sakai. Designing test collections for comparing many systems. In *Proceedings of ACM CIKM 2014*, 2014.

[23] T. Sakai. Designing test collections that provide tight confidence intervals. In *Forum on Information Technology 2014 (Volume 2) RD-003*, pages 15–18, 2014.

[24] T. Sakai. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pages 116–163, 2014.

[25] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

[26] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of ACM SIGIR 2013*, pages 473–482, 2013.

[27] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of ACM SIGIR 2011*, pages 1043–1042, 2011.

[28] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33(4):495–512, 1997.

[29] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of ACM CIKM 2007*, pages 623–632, 2007.

[30] M. D. Smucker and C. L. A. Clarke. Modeling user variance in time-biased gain. In *Proceedings of HCIR 2012*, 2012.

[31] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proceedings of ACM SIGIR 2012*, pages 95–104, 2012.

[32] J. Urbano, M. Marrero, and D. Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proceedings of ACM SIGIR 2013*, pages 925–928, 2013.

[33] J. Urbano, M. Marrero, and D. Martín. On the measurement of test collection reliability. In *Proceedings of ACM SIGIR 2013*, pages 393–402, 2013.

[34] E. M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proceeings of TREC 2003*, 2004.

[35] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceeings of TREC 2004*, 2005.

[36] E. M. Voorhees. Topic set size redux. In *Proceedings of ACM SIGIR 2009*, pages 806–807, 2009.

[37] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of ACM SIGIR 2002*, pages 316–323, 2002.

[38] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proceedings of ACM CIKM 2008*, pages 571–580, 2008.

[39] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of ACM SIGIR 1998*, pages 307–314, 1998.