Designing Test Collections for Comparing Many Systems

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

ABSTRACT

A researcher decides to build a test collection for comparing her new information retrieval (IR) systems with several state-of-the-art baselines. She wants to know the number of topics (n) she needs to create in advance, so that she can start looking for (say) a query log large enough for sampling n good topics, and estimating the relevance assessment cost. We provide practical solutions to researchers like her using power analysis and sample size design techniques, and demonstrate its usefulness for several IR tasks and evaluation measures. We consider not only the paired t-test but also one-way analysis of variance (ANOVA) for significance testing to accommodate comparison of $m(\geq 2)$ systems under a given set of statistical requirements (α: the Type I error rate, β: the Type II error rate, and $minD$: the minimum detectable difference between the best and the worst systems). Using our simple Excel tools and some pooled variance estimates from past data, researchers can design statistically well-designed test collections. We demonstrate that, as different evaluation measures have different variances across topics, they inevitably require different topic set sizes. This suggests that the evaluation measures should be chosen at the test collection design phase. Moreover, through a pool depth reduction experiment with past data, we show how the relevance assessment cost can be reduced dramatically while freezing the set of statistical requirements. Based on the cost analysis and the available budget, researchers can determine the right balance betweeen n and the pool depth pd. Our techniques and tools are applicable to test collections for non-IR tasks as well.

Categories and Subject Descriptors

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

Keywords

effect sizes; evaluation; evaluation measures; power; sample sizes; statistical significance; test collections; variances

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661893 .

1. INTRODUCTION

Reliable experimentation is crucial to the progress of IR research. The present study concerns laboratory experiments with test collections, each consisting of a target document corpus, a set of topics and relevance assessments for each topic. More specifically, we address the following situation. A researcher decides to build a test collection for comparing her new information retrieval (IR) systems with several state-of-the-art baselines. She wants to know the number of topics (n) she needs to create in advance, so that she can start looking for (say) a query log large enough for sampling n good topics, and estimating the relevance assessment cost. We provide practical solutions to researchers like her using *power analysis* [11] and *sample size design* [15] techniques, and demonstrate its usefulness for several IR tasks and evaluation measures. We consider not only the paired t-test but also one-way analysis of variance (ANOVA) for significance testing to accommodate comparison of $m(\geq 2)$ systems under a given set of statistical requirements (α: the Type I error rate, β: the Type II error rate, and $minD$: the minimum detectable difference between the best and the worst systems). Using our simple Excel tools and some pooled variance estimates from past data, researchers can design statistically well-designed test collections. We demonstrate that, as different evaluation measures have different variances across topics, they inevitably require different topic set sizes. This suggests that the evaluation measures should be chosen at the test collection design phase. Moreover, through a pool depth reduction experiment with past data, we show how the relevance assessment cost can be reduced dramatically while freezing the set of statistical requirements. Based on the cost analysis and the available budget, researchers can determine the right balance betweeen n and the pool depth pd. Our techniques, as well as our Excel tools, are applicable to any *non-IR* tasks (e.g., question answering, summarisation, machine translation, recommendation etc.) as well, as long as the paired t-test or ANOVA is applicable to the task and test collection in question.

2. RELATED WORK

In some research disciplines, top-tier journals require reporting of *effect sizes* and *confidence intervals* along with p-values, as it is known that p-values are not informative enough [10, 11, 12, 20]. For example, suppose we have per-topic performance scores in terms of some measure M for systems X and Y with n topics: $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. Consider the test statistic t_0 for a paired t-test:

$$t_0 = \frac{\bar{d}}{\sqrt{V/n}} = \sqrt{n}\frac{\bar{d}}{\sqrt{V}} \quad (1)$$

where $d_j = x_j - y_j$, $\bar{d} = \sum_{j=1}^{n} d_j/n$ (the sample mean) and $V = \sum_{j=1}^{n}(d_j - \bar{d})^2/(n-1)$ (the unbiased estimate of the popula-

tion variance). Now, as t_0 becomes larger (i.e., more extreme), the corresponding p-value becomes smaller and therefore the between-system delta is more likely to be considered statistically significant. However, it is clear from Eq. 1 that a large t_0 may mean either (a) the sample size n is large; or (b) the sample *effect size* $\bar{d}/\sqrt{V}$, the difference between X and Y measured in standard deviation units, is large. In other words, p-values confound the effect of n and the magnitude of the "real" difference [11, 15, 20].

In the IR research discipline (and in related disciplines such as natural language processing), the aforementioned "statistical reform" is yet to happen [20]. Thus, unfortunately, it is not uncommon for IR researchers to discuss the Type I error rate (α) without heeding the Type II error rate (β) and the effect size. Exceptions in the IR literature include the work of Nelson [16] who pointed out the usefulness of statistical power analysis for IR research, and that of Carterette and Smucker [4] who investigated the relationship between the number of topics evaluated and the statistical power of the *sign test*, where *Average Precision* (AP) was computed with incomplete relevance assessments. In several ways, the present study extends the work of Webber, Moffat and Zobel [27] who proposed to design test collections based on power analysis. Our new contributions over their work are as follows:

- Webber *et al.* were primarily concerned with building a test collection *incrementally*, by adding topics with relevance assessments one by one while checking to see if the desired power is achieved and reestimating the population variance of the performance score differences. In contrast, the present study provides straight answers to questions like: "I want to build a new test collection that guarantees certain levels of Type I and Type II error rates (α and β). What is the number of topics (n) that I will have to prepare?" We also provide simple Excel tools that answer these questions.
- Webber *et al.* considered evaluating a given *pair* of systems and thus considered the t-test only. However, test collections are used to compare $m(\geq 2)$ systems in practice, and it is not advisable to conduct t-tests independently for every system pair. If this is done, the *family-wise* error rate (i.e., the probability of detecting at least one nonexistent between-system difference) amounts to $1-(1-\alpha)^{m(m-1)/2}$ [11][1]. In contrast, the present study computes the required topic set size n by considering both the t-test (for $m=2$) and one-way ANOVA (for $m \geq 2$), and examine the effect of m on the required n for a given set of statistical requirements[2].
- Webber *et al.* examined a few methods to estimate the population variance of the performance score deltas, which include taking the 95th percentile of the observed score delta variance from past data, and conducting pilot relevance assessments. However, a time-honoured method is available for estimating the population variance accurately by utilising the statistics obtained in ANOVA. The present study makes use of this technique. Morever, we *pool* variances from multiple existing data sets to obtain reliable estimates.
- Webber *et al.* considered AP only; we examine a variety of evaluation measures for ad hoc and diversified search, and demonstrate that some measures require many more topics than others under the same set of statistical requirements.

[1]Ellis [11] remarks that the Bonferroni correction to counter this problem "may be a bit like spending $1,000 to buy insurance for a $500 watch."

[2]One-way ANOVA is equivalent to the *unpaired* t-test generalised to the case of $m \geq 2$ systems.

Evaluation forums such as TREC, NTCIR and CLEF typically build (say) 50 topics every year for each IR task[3]. However, it is not appropriate to combine these topic sets in the hope of conducting an experiment with high statistical power unless each topic set is known to be reusable. Moreover, note that the choice of $n=50$ is arbitrary; topic set splitting tests have been used in the literature to answer retrospective questions such as "Was $n=50$ large enough for conducting reliable experiments?" (e.g. [26]). In contrast, the present study provides a method to systematically determine n for a new test collection, based on variance estimates from past data. Using this methodology, it is possible to *improve* the test collection design over past rounds of IR tasks.

Alternatives to classical significance testing include Killeen's p_{rep} [14] and the Baysian approach to hypothesis testing [2, 13], both of which are beyond the scope of this study. The Generalisability Theory (GT) has been shown to be useful for assessing the test collection reliability [1, 3, 23]: while both the GT approach and ours rely on variance estimates from past data, Urbano, Marrero and Martín [23] 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.

3. TWO SYSTEMS

This section discusses how to set the topic set size n when we want to compare $m=2$ systems (X and Y) in terms of some measure M using the two-sided paired t-test, where it is assumed that the per-topic performance scores $\{x_j\}$ and $\{y_j\}$ ($j=1,\ldots,n$) are independent and $x_j \sim N(\mu_X, \sigma_X^2), y_j \sim N(\mu_Y, \sigma_Y^2)$. The null and alternative hypotheses are: $H_0: \mu_X = \mu_Y$ (i.e., the population means of X and Y are identical) and $H_1: \mu_X \neq \mu_Y$.

3.1 Significance Criterion and Power

Let t be a random variable that obeys a t distribution with ϕ degrees of freedom; let $t(\phi; P)$ denote the two-sided critical t value for probability P (i.e., $Pr\{|t| \geq t(\phi;P)\} = P$). *Under* H_0, the test statistic t_0 (Eq. 1 in Section 2) obeys a t distribution with $\phi = n-1$ degrees of freedom. Given a *significance criterion* α, we reject H_0 if $|t_0| \geq t(\phi;\alpha)$. (The p-value is the probability of observing t_0 or something more extreme, $Pr\{|t| \geq t_0\}$, under H_0.) Thus, the probability of Type I error (i.e., "finding" a difference that does not exist) is exactly α by construction. Whereas, the probability of Type II error (i.e., missing a difference that actually exists) is denoted by β, and therefore the *statistical power* (i.e., the ability to detect a real difference) is given by $1-\beta$. Put another way, α is the probability of rejecting H_0 *when* H_0 *is true*, while the power is the probability of rejecting H_0 *when* H_1 *is true*. In either case, the probability of rejecting H_0 is given by

$$Pr\{t_0 \leq -t(\phi;\alpha)\} + Pr\{t_0 \geq t(\phi;\alpha)\}$$

$$= Pr\{t_0 \leq -t(\phi;\alpha)\} + 1 - Pr\{t_0 \leq t(\phi;\alpha)\}\ . \quad (2)$$

Under H_0, Eq. 2 amounts to α, where t_0 (Eq. 1) obeys a (central) t distribution as mentioned above. *Under* H_1, Eq. 2 represents the power ($1-\beta$), where t_0 obeys a *noncentral* t distribution with $\phi = n-1$ degrees of freedom and a *noncentrality parameter* $\lambda_t = \sqrt{n}\Delta_t$. Here, Δ_t is a simple form of *effect size*, given by:

$$\Delta_t = \frac{\mu_X - \mu_Y}{\sigma_t} = \frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}} \quad (3)$$

[3]Exceptions include the TREC Million Query track that was designed specifically to construct a "minimal" test collection for a given set of systems and a particular evaluation measure (AP) [3].

where $\sigma_t^2 = \sigma_X^2 + \sigma_Y^2$ is the population variance of the score differences. Thus, Δ_t quantifies the difference between X and Y in *standard deviation units*, regardless of the evaluation measure used.

While computations involving a noncentral t distribution can be complex, a normal approximation is available: let t' denote a random variable that obeys the aforementioned noncentral t distribution; let u denote a random variable that obeys $N(0, 1^2)$. Then:

$$Pr\{t' \leq w\} \approx Pr\{u \leq \frac{w(1 - 1/4\phi) - \lambda_t}{\sqrt{1 + w^2/2\phi}}\} . \quad (4)$$

Hence, given the topic set size n, the effect size Δ_t and the significance criterion α, the power can be computed from Eqs. 2 and 4 as [15]:

$$1 - \beta \approx Pr\{u \leq \frac{(-w)(1-1/4(n-1)) - \sqrt{n}\Delta_t}{\sqrt{1+(-w)^2/2(n-1)}}\} + 1 - Pr\{u \leq \frac{w(1-1/4(n-1)) - \sqrt{n}\Delta_t}{\sqrt{1+w^2/2(n-1)}}\} \quad (5)$$

where $w = t(n-1; \alpha)$. But what we are more interested in is: given $(\alpha, \beta, \Delta_t)$, what is the required n?

3.2 How to Determine the Topic Set Size

Under H_0, we know that $\Delta_t = 0$ (See Eq. 3). However, under H_1, all we know is that $\Delta_t \neq 0$. In order to require that an experiment has a statistical power of $1 - \beta$, a *minimum detectable effect* $min\Delta_t$ must be specified in advance: we correctly reject H_0 with $100(1 - \beta)\%$ confidence whenever $|\Delta_t| \geq min\Delta_t$. That is, we should not miss a real difference if its effect size is $min\Delta_t$ or larger. Cohen calls $min\Delta_t = 0.2$ a *small* effect, $min\Delta_t = 0.5$ a *medium* effect, and $min\Delta_t = 0.8$ a *large* effect [9, 11][4].

Let z_P denote the one-sided critical z value of $u(\sim N(0, 1^2))$ for probability P (i.e., $Pr\{u \geq z_P\} = P$). Given $(\alpha, \beta, min\Delta_t)$, it is known that the required n can be approximated by [15]:

$$n \approx (\frac{z_{\alpha/2} - z_{1-\beta}}{min\Delta_t})^2 + \frac{z_{\alpha/2}^2}{2} . \quad (6)$$

For example, if we let $(\alpha, \beta, min\Delta_t) = (.05, .20, .50)$ (i.e., *Cohen's five-eighty convention* [9, 11] with Cohen's *medium* effect),

$$n \approx (\frac{1.960 - (-.842)}{.50})^2 + \frac{1.960^2}{2} = 33.3 . \quad (7)$$

As this is only an approximation, we need to check that the desired power is actually achieved with an integer n close to 33.3. Suppose we let $n = 33$. Then, by substituting $w = t(33 - 1; .05) = 2.037$ and $\Delta_t = min\Delta_t = .50$ to Eq. 5, we obtain:

$$1 - \beta \approx Pr\{u \leq -4.742\} + 1 - Pr\{u \leq -.825\} = .795 \quad (8)$$

which means that the desired power of 0.8 is not quite achieved. So we let $n = 34$, and the achieved power can be computed similarly: $1 - \beta = .808$. Therefore $n = 34$ is the topic set size we want.

Our Excel tool `samplesizeTTEST` (See Section 7) automates the above procedure for any given combination of $(\alpha, \beta, min\Delta_t)$. Table 1 shows the required topic set sizes for the paired t-test for some typical combinations. For example, under Cohen's five-eighty convention ($\alpha = .05, \beta = .20$)[5], if we want the minimum detectable effect to be $min\Delta_t = .2$ (i.e., one-fifth of the score-difference standard deviation), we need $n = 199$ topics.

[4]Strictly speaking, Cohen's criteria are for *unpaired* tests; effect sizes for paired and unpaired tests are not directly comparable [17].

[5]Note that this convention, which implies that a Type I error is four times as serious as a Type II error, is only a convention [11]. Researchers should consider whether this is appropriate for their experiments, and should not follow it blindly.

Table 1: Topic set sizes for $(\alpha, \beta, min\Delta_t)$.

α	$min\Delta_t$	$\beta = .10$	$\beta = .20$
.01	.1	1492	1172
	.2	376	296
	.5	63	51
	1.0	19	16
.05	.1	1053	**787**
	.2	265	**199**
	.5	44	**34**
	1.0	13	**10**

The above approach starts by requiring a $min\Delta_t$, which is independent of the evaluation method (i.e., the measure, pool depth and the measurement depth). However, researchers may want to require a minimum detectable *absolute difference* $minD_t$ in terms of a particular evaluation measure M instead (e.g., "I want high power guaranteed whenever the true absolute difference in mean AP is 0.05 or larger."). In this case, instead of setting a minimum ($min\Delta_t$) for Eq. 3, we can set a minimum ($minD_t$) for the *numerator* of Eq. 3: we guarantee a power of $1 - \beta$ whenever $|\mu_X - \mu_Y| \geq minD_t$. To do this, we need an estimate of $\sigma_t^2 (= \sigma_X^2 + \sigma_Y^2)$, which we denote by $\hat{\sigma}_t^2$, so that we can convert $minD_t$ to $min\Delta_t = minD_t / \sqrt{\hat{\sigma}_t^2}$ and follow the aforementioned procedure for finding the right n. The `samplesizeTTEST` tool has a separate sheet for computing n from $(\alpha, \beta, minD_t, \hat{\sigma}_t^2)$; we shall discuss how to obtain $\hat{\sigma}_t^2$ from past data in Section 5.

4. MORE THAN TWO SYSTEMS

This section discusses how to set the topic set size n when we assume that there are $m \geq 2$ systems to be compared using one-way ANOVA. Let x_{ij} denote the score of the i-th system for topic j in terms of some measure M; we assume that $\{x_{ij}\}$ are independent and that $x_{ij} \sim N(\mu_i, \sigma^2)$. Note the *homoscedasticity* assumption: the variance σ^2 is assumed to be common across systems. (We did not assume this when we discussed the paired t-test.) We define the population grand mean μ and the i-th *system effect* a_i as follows:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} \mu_i , \quad a_i = \mu_i - \mu \quad (9)$$

where $\sum_{i=1}^m a_i = \sum_{i=1}^m (\mu_i - \mu) = \sum_{i=1}^m \mu_i - m\mu = 0$. The null hypothesis for the ANOVA is $H_0 : \mu_1 = \cdots = \mu_m$ (or $a_1 = \cdots = a_m = 0$) while the alternative hypothesis H_1 is that at least one of the system effects is not zero. The basic statistics that we compute for the ANOVA are as follows. Let $\bar{x}_i = \frac{1}{n}\sum_{j=1}^n x_{ij}$ (sample system mean) and $\bar{x} = \frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n x_{ij}$ (sample grand mean); the total variation $S_T = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2$ can be decomposed into between-system and within-system variations S_A and S_E (i.e., $S_T = S_A + S_E$), where

$$S_A = n\sum_{i=1}^{m} (\bar{x}_i - \bar{x})^2 , \quad S_E = \sum_{i=1}^{m}\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 . \quad (10)$$

The corresponding degrees of freedom are $\phi_A = m - 1$, $\phi_E = m(n-1)$. Also, let $V_A = S_A/\phi_A, V_E = S_E/\phi_E$ for later use.

4.1 Significance Criterion and Power

Let F be a random variable that obeys an F distribution with (ϕ_A, ϕ_E) degrees of freedom; let $F(\phi_A, \phi_E; P)$ denote the critical F value for probability P (i.e., $Pr\{F \geq F(\phi_A, \phi_E; P)\} = P$). *Under* H_0, the test statistic F_0 defined below obeys a (central) F distribution with (ϕ_A, ϕ_E) degrees of freedom:

$$F_0 = \frac{V_A}{V_E} = \frac{m(n-1)S_A}{(m-1)S_E} . \quad (11)$$

Table 2: Linear approximation of λ, the noncentrality parameter of a noncentral χ^2 distribution [15].

α	β	formula
.01	.10	$\lambda = 10.439 + 5.213\sqrt{\phi_A}$
.01	.20	$\lambda = 7.736 + 4.551\sqrt{\phi_A}$
.05	.10	$\lambda = 7.049 + 4.244\sqrt{\phi_A}$
.05	.20	$\lambda = 4.860 + 3.584\sqrt{\phi_A}$

Given a significance criterion α, we reject H_0 if $F_0 \geq F(\phi_A, \phi_E; \alpha)$. (The p-value is given by $Pr\{F \geq F_0\}$ under H_0.) From Eq. 11, it can be observed that H_0 is rejected if the between-system variation S_A is large compared to the within-system variation S_E, or simply if the sample size n is large. Again, the p-value does not tell us which is the case.

The probability of rejecting H_0 is given by

$$Pr\{F_0 \geq F(\phi_A, \phi_E; \alpha)\} = 1 - Pr\{F_0 \leq F(\phi_A, \phi_E; \alpha)\} . \quad (12)$$

Under H_0, Eq. 12 amounts to α by construction, where F_0 obeys a (central) F distribution as mentioned above. *Under* H_1, Eq. 12 represents the power $(1-\beta)$, where F_0 obeys a *noncentral* F distribution with (ϕ_A, ϕ_B) degrees of freedom and a *noncentrality parameter* $\lambda = n\Delta$, where

$$\Delta = \frac{\sum_{i=1}^{m} a_i^2}{\sigma^2} = \frac{\sum_{i=1}^{m} (\mu_i - \mu)^2}{\sigma^2} . \quad (13)$$

Thus Δ measures the total system effects in *variance units*.

While computations involving a noncentral F distribution can be complex, a normal approximation is available: let F' denote a random variable that obeys the aforementioned noncentral F distribution; let $u \sim N(0, 1^2)$. Then:

$$Pr\{F' \leq w\} \approx Pr\{u \leq \frac{\sqrt{\frac{w}{\phi_E}}\sqrt{2\phi_E - 1} - \sqrt{\frac{c_A}{\phi_A}}\sqrt{2\phi_A^* - 1}}{\sqrt{\frac{c_A}{\phi_A} + \frac{w}{\phi_E}}}\} \quad (14)$$

where

$$c_A = \frac{m - 1 + 2n\Delta}{m - 1 + n\Delta} , \quad \phi_A^* = \frac{(m - 1 + n\Delta)^2}{m - 1 + 2n\Delta}. \quad (15)$$

Hence, given (n, Δ, α), the power $(1-\beta)$ can be computed from Eqs. 12-15 as [15]:

$$1 - Pr\{u \leq \frac{\sqrt{\frac{w}{m(n-1)}}\sqrt{2m(n-1) - 1} - \sqrt{\frac{c_A}{m-1}}\sqrt{2\phi_A^* - 1}}{\sqrt{\frac{c_A}{m-1} + \frac{w}{m(n-1)}}}\} \quad (16)$$

where $w = F(m-1, m(n-1); \alpha)$. But what we are more interested in is: given (α, β, Δ), what is the required n?

4.2 How to Determine the Topic Set Size

Under H_0, we know that $\Delta = 0$ (See Eq. 13). However, under H_1, all we know is that $\Delta \neq 0$. In order to require that an experiment has a statistical power of $1-\beta$, a minimum detectable delta $min\Delta$ must be specified in advance. Let us require that we correctly reject H_0 with $100(1-\beta)\%$ confidence whenever the *range* of the population means ($D = \max_i a_i - \min_i a_i$) is at least as large as a specified value ($minD$), and let $min\Delta = \frac{minD^2}{2\sigma^2}$. Since $\sum_{i=1}^{m} a_i^2 \geq \frac{D^2}{2}$ holds[6], it follows that

$$\Delta = \frac{\sum_{i=1}^{m} a_i^2}{\sigma^2} \geq \frac{D^2}{2\sigma^2} \geq \frac{minD^2}{2\sigma^2} = min\Delta . \quad (17)$$

That is, Δ is bounded below by $min\Delta$ as defined above. Hence, although specifying $minD$ does not uniquely determine Δ (as Δ depends on systems other than the best and the worst ones), we can plug in $\Delta = min\Delta$ to Eqs. 15 and 16 to obtain the worst-case estimate of the power.

Unfortunately, no closed formula similar to Eq. 6 is available for ANOVA. However, since the worse-case estimate of n can be obtained as $n = \lambda/min\Delta = 2\sigma^2\lambda/minD^2$, we can first estimate λ as follows. (How to obtain $\hat{\sigma}^2$, the estimate of σ^2, is discussed in Section 5.) Recall that, under H_1, Eq. 12 represents the power $(1-\beta)$ where F_0 obeys a noncentral F distribution with (ϕ_A, ϕ_E) degrees of freedom and the noncentrality parameter λ. By letting $\phi_E = m(n-1) \approx \infty$, the power can be approximated by:

$$Pr\{F_0 \geq F(\phi_A, \infty; \alpha)\} = Pr\{\chi'^2 \geq \chi^2(\phi_A; \alpha)\} \quad (18)$$

where χ'^2 is a random variable that obeys a *noncentral* χ'^2 distribution with ϕ_A degrees of freedom whose noncentrality parameter is λ, and $\chi^2(\phi; P)$ is the critical χ^2 value for probability P of a random variable that obeys a (central) χ^2 distribution with ϕ degrees of freedom (i.e., $Pr\{\chi^2 \geq \chi^2(\phi; P)\} = P$). For noncentral χ^2 distributions, some linear approximations of λ are available, as shown in Table 2 [15]. Hence an initial estimate of n given $(\alpha, \beta, minD, \hat{\sigma}^2, m)$ can be obtained as shown below.

Suppose we let $(\alpha, \beta, minD, m) = (.05, .20, .5, 3)$ and that we obtained $\hat{\sigma}^2 = .5^2$ from past data so that $min\Delta = \frac{minD^2}{2\sigma^2} = .5^2/(2 * .5^2) = .5$. Then $\phi_A = m - 1 = 2$ and $\lambda = 4.860 + 3.584 * \sqrt{2} = 9.929$ and hence $n = \lambda/min\Delta = 19.9$. If we let $n = 19$, then $\phi_E = 3(19-1) = 54$, $w = F(2, 54; .05) = 3.168$. From Eq. 15, $c_A = 1.826, \phi_A^* = 6.298$, and from Eq. 16, the achieved power is $1 - Pr\{u \leq -.809\} = .791$, which does not quite satisfy the desired power of 80%. On the other hand, if $n = 20$, the achieved power can be computed similarly as .813. Hence $n = 20$ is what we want. Our Excel tool `samplesizeANOVA` (See Section 7)[7] automates the above procedure for given $(\alpha, \beta, minD, \hat{\sigma}^2, m)$.

5. ESTIMATING THE DELTA VARIANCE

5.1 Estimation Method

Recall that for our topic set size design based on the paired t-test, we need an estimate of $\sigma_t^2 = \sigma_X^2 + \sigma_Y^2$ to compute $min\Delta_t$ (Section 3.2). Similarly, for our topic set size design based on one-way ANOVA, we need an estimate of σ^2 to compute $min\Delta$ (Section 4.2). There are time-honoured methods for estimating the population variances from ANOVA statistics; for one-way ANOVA, the following estimate is available [17]:

$$\hat{\sigma}^2 = \frac{m' - 1}{m'n'}(V_A' - V_E') + V_E' \quad (19)$$

where the symbols used represent n, m, V_A and V_E as already defined, except that here we emphasise that they are computed from

[6] Let $A = \max_i a_i$ and $a = \min_i a_i$. Then $D^2/2 = (A^2 + a^2 - 2Aa)/2 \leq A^2 + a^2 \leq \sum_{i=1}^{m} a_i^2$. The equality holds when $A = D/2, a = -D/2$ and $a_i = 0$ for all other systems.

[7] While `samplesizeTTEST` handles arbitrary values of (α, β), `samplezieANOVA` can only handle the four combinations shown in Table 2.

Table 3: TREC test collections and runs used for estimating σ^2. The web track relevance grades [7, 8] 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 [27])

Table 4: 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 5: $\hat{\sigma}^2$ for different evaluation measures with measurement depth l.

			$\hat{\sigma}^2$			
(a1) task: adhoc/news ($l = 1000$)						
Data	m	n	AP	Q	nDCG	nERR
TREC03new	78	50	.0537	.0542	.0563	.1201
TREC04new	78	49	.0510	.0520	.0553	.1190
Pooled	-	-	**.0524**	**.0531**	**.0558**	**.1196**
(a2) task: adhoc/news ($l = 10$)						
Data	m	n	AP	Q	nDCG	nERR
TREC03new	78	50	.0958	.0702	.0775	.1268
TREC04new	78	49	.0815	.0660	.0770	.1249
Pooled	-	-	**.0887**	**.0681**	**.0773**	**.1259**
(b) task: adhoc/web ($l = 10$)						
Data	m	n	AP	Q	nDCG	nERR
TREC11w	37	50	.0902	.0494	.0566	.1052
TREC12w	28	50	.0834	.0273	.0357	.0757
Pooled	-	-	**.0868**	**.0384**	**.0462**	**.0905**
(c) task: diversity/web ($l = 10$)						
Data	m	n	α-nDCG	nERR-IA	D-nDCG	D♯-nDCG
TREC11wD	25	50	.0890	.0942	.0415	.0634
TREC12wD	20	50	.0760	.0835	.0329	.0449
Pooled	-	-	**.0825**	**.0889**	**.0372**	**.0542**

past data. The first term is an estimate of the population between-system variance σ_A^2; the second term is an estimate of the population within-system variance σ_E^2. These estimates are also used for computing accurate effect size estimates for ANOVA [17].

Given an n'-by-m' topic-by-system score matrix for a particular evaluation measure M from past data, we can easily obtain $\hat{\sigma}^2$ by substituting the ANOVA statistics to Eq. 19, under the homoscedasticity assumption. Now, if we introduce homoscedasticity to the paired t-test as well, it seems reasonable to obtain the required variance estimate for the score *differences* as $\hat{\sigma}_t^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 = 2\hat{\sigma}^2$.

5.2 Pooling Variances across Data

To enhance the reliability of the variance estimates, it is possible to pool variances across data sets for a given IR task. Let C denote a past data set, and let n'_C and $\hat{\sigma}_C^2$ denote the number of topics in C and the variance estimate obtained from C using Eq. 19. We then use the following *pooled* variance estimate:

$$\hat{\sigma}^2 = \sum_C (n'_C - 1)\hat{\sigma}_C^2 / \sum_C (n'_C - 1) \ . \qquad (20)$$

Table 3 provides some statistics of the past data that we used for obtaining $\hat{\sigma}^2$'s. We considered 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 [24, 25]. (We also conducted some experiments with the "old" topics of the TREC 2004 robust track, following Webber, Moffat and Zobel [27], and the results obtained were similar to the ones reported in this paper. However, the data set is not ideal for our pool depth reduction experiment (Section 6.3), as the relevance assessment pools for the old topics come from old TREC ad hoc runs, not the TREC 2004 robust track runs. Moreover, the relevance assessments for the old topics are binary, even though we are primarily interested in graded-relevance measures.) The web data sets are from the TREC web tracks [7, 8]. While we considered the measurement depths of $l = 10, 1000$ for adhoc/news, we considered only $l = 10$ 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 4 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 `NTCIREVAL` toolkit[8]. For the diversity/web task, we consider α-nDCG and Intent-Aware nERR (nERR-IA) computed using `ndeval`[9], 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.

Table 5 shows the variance estimates obtained for each evaluation measure and for each task, using Eqs. 19 and 20. It can be observed that nERR is highly unstable for Tasks (a1), (a2) and (b). Due to its *diminishing return* property [5, 19], (n)ERR basically ignores most retrieved relevant documents except for the ones retrieved at the very top. Relying on fewer data points hurts statistical stability, and hence calls for a very large topic set size, as we shall see later. As for AP, it is much more unstable than Q and nDCG for Task (b) (adhoc/web), which suggests that the use of graded relevance is important in evaluating this task. Note also that AP is less stable than Q and nDCG for Task (a2), that is, adhoc/news with a small measurement depth. Finally, for Task (c) (diversity/web), it can be observed that α-nDCG and nERR-IA have high variances compared to D-nDCG and D♯-nDCG: this is

[8] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`. For computing AP and Q, we follow Sakai and Song [21] and divide by $\min(l, R)$ rather than by R in order to properly handle small measurement depths.

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

Table 6: Topic set size table for $(\alpha, \beta) = (0.01, 0.10)$ with $m = 2$: t-test vs. one-way ANOVA.

$minD_{(t)}$	t-test	ANOVA
(a1) adhoc/news, $l = 1000$ (AP/Q/nDCG/nERR)		
.02	3902/3954/4155/8902	3850/3901/4099/8785
.05	628/636/668/1427	617/625/657/1407
.10	160/162/170/360	155/157/165/352
.20	43/43/45/93	40/40/42/89
.25	29/29/30/61	26/26/27/57
(a2) adhoc/news, $l = 10$ (AP/Q/nDCG/nERR)		
.02	6603/5070/5755/9370	6516/5003/5678/9248
.05	1060/814/924/1502	1043/801/909/1481
.10	268/206/234/378	262/201/228/371
.20	70/55/61/98	66/51/58/94
.25	46/36/41/64	43/33/37/60
(b) adhoc/web, $l = 10$ (AP/Q/nDCG/nERR)		
.02	6461/2861/3441/6737	6376/2821/3394/6648
.05	1037/461/554/1081	1021/452/544/1065
.10	262/118/141/273	256/114/137/267
.20	68/32/38/71	65/29/35/68
.25	45/22/26/47	42/19/23/44
(c) adhoc/diversity, $l = 10$ (α-nDCG/nERR-IA/D-nDCG/D♯-nDCG)		
.02	6142/6618/2771/4036	6060/6530/2733/3982
.05	986/1062/447/649	971/1046/438/638
.10	249/268/115/165	243/262/110/160
.20	65/70/32/44	62/66/28/41
.25	43/46/22/30	40/43/19/27

Table 7: Topic set size table for $(\alpha, \beta) = (0.05, 0.20)$ with $m = 2$: t-test vs. one-way ANOVA.

$minD_{(t)}$	t-test	ANOVA
(a1) adhoc/news, $l = 1000$ (AP/Q/nDCG/nERR)		
.02	**2059/2086/2192/4696**	**2006/2033/2136/4578**
.05	**331/336/353/753**	**322/326/343/733**
.10	**85/86/90/190**	**81/82/86/184**
.20	**23/23/24/49**	**21/21/22/47**
.25	**16/16/17/33**	**14/14/15/30**
(a2) adhoc/news, $l = 10$ (AP/Q/nDCG/nERR)		
.02	**3483/2675/3036/4943**	**3395/2607/2959/4819**
.05	**559/430/488/793**	**544/418/474/772**
.10	**142/109/124/200**	**137/105/119/194**
.20	**37/29/33/52**	**35/27/30/49**
.25	**25/20/22/34**	**23/18/20/32**
(b) adhoc/web, $l = 10$ (AP/Q/nDCG/nERR)		
.02	**3409/1509/1816/3554**	**3322/1470/1769/3464**
.05	**547/244/293/571**	**532/236/284/555**
.10	**139/63/75/144**	**134/60/72/139**
.20	**37/18/21/38**	**34/16/19/36**
.25	**24/12/14/25**	**22/10/12/23**
(c) adhoc/diversity, $l = 10$ (α-nDCG/nERR-IA/D-nDCG/D♯-nDCG)		
.02	**3240/3491/1462/2129**	**3158/3403/1424/2075**
.05	**520/561/236/343**	**506/545/229/333**
.10	**132/142/61/88**	**127/137/58/84**
.20	**35/37/17/24**	**32/35/15/22**
.25	**23/25/12/16**	**21/23/10/14**

because α-nDCG and nERR-IA possess the *per-intent* diminishing return property; it is known that these two measures behave similarly [5, 6, 22]. While D-nDCG does not have the diminishing return property, D♯-nDCG compensates for this by averaging D-nDCG with *intent recall* (a.k.a. *subtopic recall*) [19, 21].

Note that within each IR task, the variance estimates are very similar. Henceforth, we shall use the pooled variance estimates (shown in bold in Table 5) to compute the required topic set sizes n based on ANOVA. As for the case with the t-test, we use the same pooled estimates to obtain $\hat{\sigma}_t^2 = 2\hat{\sigma}^2$.

6. RESULTS AND DISCUSSIONS

6.1 Two Systems: t-test vs. ANOVA

Table 6 compares the required topic set sizes computed based on the t-test and one-way ANOVA, when we want to design a test collection for comparing a *pair* of systems ($m = 2$) under $(\alpha, \beta) = (0.01, 0.10)$. The t-test row uses `samplesizeTTEST` (See Section 3.2) while the ANOVA row uses `samplesizeANOVA` (See Section 4.2), using the pooled variance estimates shown in Table 5. Recall that while $minD_t$ is the minimum detectable difference between two systems being compared with the t-test, $minD$ is the minimum detectable difference between the best and the worst systems when m systems are being compared with ANOVA. Hence $minD$ is equivalent to $minD_t$ when $m = 2$. Table 7 provides similar information under a less demanding condition of $(\alpha, \beta) = (0.05, 0.20)$, i.e., Cohen's five-eighty convention. (Throughout this paper, we use **boldface** whenever we show topic set size estimates under Cohen's convention within tables.) For example, Table 7(a1) (adhoc/news, $l = 1000$) says that, if we require the minimum detectable difference of $minD_t = 0.05$ under Cohen's five-eighty convention, t-tests with AP, Q, nDCG and nERR would require 331, 336, 353, 753 topics, respectively; similarly, ANOVA ($m = 2$) with AP, Q, nDCG and nERR would require 322, 326, 343, 733 topics, respectively. It can be observed that the t-test and ANOVA results are very similar, and that the ANOVA-based estimates are slightly smaller, despite the fact that the t-test exploits the paired data information (i.e., that the per-topic score x_i corresponds to y_i). This may be because the $\hat{\sigma}^2$ values are overestimates: if this is the case, the error is doubled when we compute $\hat{\sigma}_t^2 = 2\hat{\sigma}^2$ for the t-test-based sample size design. Clearly, if the variance is overestimated, the required topic set size will also be overestimated. However, since we want to *guarantee* low probabilities of Type I and Type II errors, it seems appropriate to "err on the side of oversampling" as suggested by Ellis [11]. Henceforth, we use `samplesizeANOVA` to discuss the general case of $m \geq 2$.

6.2 Topic Set Sizes for Comparing Many Systems

Tables 8-11 show the required topic set sizes for different IR tasks under different statistical requirements, for $m = 10, 100$. Again, the pooled variance estimates shown in Table 5 were used for the computation. The interested reader can use `samplesizeANOVA` to easily reproduce our results or try other parameter settings, with her own evaluation measures and variance estimates.

Tables 8 and 9 are the topic set size tables for the adhoc/news task with $l = 1000$ and $l = 10$. We can observe that:

- As nERR is substantially less stable than AP, Q and nDCG (See Tables 5(a1) and (a2)), it requires many more topics than the other measures under the same condition. For example, Table 8(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $l = 1000$, nERR requires 965 topics while AP, Q and nDCG require only 423, 429, 451 topics, respectively. Throughout Table 8, nERR is more than twice as expensive as AP, Q and nDCG.

- As reducing the measurement depth l causes higher variances (Compare Tables 5(a1) and (a2)), this also means we need more topics. More importantly, however, while the advantage of utilising the graded relevance assessments with Q and nDCG is not clear when $l = 1000$ (a typical TREC ad hoc setting), it is clear for a shallow measurement depth of $l = 10$. For example, Table 9(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $l = 10$, AP requires 716 topics, while Q and nDCG require only 550 and 624 topics, respectively. (nERR requires 1,016 topics.)

Table 8: Topic set size table ($m = 10, 100$) for adhoc/news ($l = 1000$) with AP/Q/nDCG/nERR.

α	$minD$	$\beta = .10$	$\beta = .20$
		(I) $m = 10$	
.01	.02	6842/6933/7285/15614	5595/5670/5958/12770
	.05	1095/1110/1166/2499	896/908/954/2044
	.10	275/278/292/625	225/228/239/512
	.20	69/70/74/157	57/58/61/129
	.25	45/45/48/101	37/37/39/83
.05	.02	5197/5267/5534/11862	**4081/4135/4345/9313**
	.05	832/843/886/1898	**654/662/696/1491**
	.10	209/211/222/475	**164/166/175/373**
	.20	53/53/56/119	**42/42/44/94**
	.25	34/34/36/77	**27/27/29/60**
		(II) $m = 100$	
.01	.02	16305/16523/17363/37215	13842/14027/14740/31592
	.05	2609/2644/2779/5955	2215/2245/2359/5055
	.10	653/662/695/1489	554/562/590/1264
	.20	164/166/174/373	139/141/148/317
	.25	105/106/112/239	89/91/95/203
.05	.02	12892/13064/13729/29425	**10567/10708/11253/24118**
	.05	2063/2091/2197/4709	**1691/1714/1801/3859**
	.10	516/523/550/1178	**423/429/451/965**
	.20	130/131/138/295	**106/108/113/242**
	.25	83/84/88/189	**68/69/73/155**

Table 9: Topic set size table ($m = 10, 100$) for adhoc/news ($l = 10$) with AP/Q/nDCG/nERR.

α	$minD$	$\beta = .10$	$\beta = .20$
		(I) $m = 10$	
.01	.02	11580/8891/10092/16437	9471/7271/8254/13442
	.05	1854/1423/1615/2631	1516/1164/1321/2152
	.10	464/357/405/658	380/292/331/539
	.20	117/90/102/165	96/74/84/135
	.25	75/58/65/106	62/48/54/87
.05	.02	8797/6754/7667/12486	**6907/5303/6019/9803**
	.05	1408/1081/1227/1998	**1106/849/964/1569**
	.10	353/271/307/500	**277/213/242/393**
	.20	89/68/77/126	**70/54/61/99**
	.25	57/44/50/81	**45/35/39/64**
		(II) $m = 100$	
.01	.02	27600/21190/24053/39175	23430/17989/20419/33256
	.05	4417/3391/3849/6269	3749/2879/3268/5322
	.10	1105/848/963/1568	938/720/818/1331
	.20	277/213/241/392	235/181/205/333
	.25	177/136/155/251	151/116/132/214
.05	.02	21823/16755/19018/30975	**17887/13733/15588/25388**
	.05	3492/2681/3043/4956	**2863/2198/2495/4063**
	.10	874/671/761/1240	**716/550/624/1016**
	.20	219/168/191/310	**180/138/157/255**
	.25	140/108/122/199	**115/89/100/163**

- In Table 8(I), the number of topics required by AP is $n = 42$ under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.20, 10)$. Thus, a typical TREC adhoc/news test collection with $n = 50$ topics is good enough for guaranteeting a minimum detectable range of 0.20 in terms of AP for $m = 10$ systems under Cohen's five-eighty convention.

Table 10 is the topic set size table for the adhoc/web task with $l = 10$. It can be observed that:

- As Q and nDCG are substantially more stable than AP and nERR for this task (See Table 5(b)), they require substantially fewer topics under the same condition. For example, Table 10(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $l = 10$, AP and nERR require 701 and 731 topics, while Q and nDCG require only 310 and 373 topics, respectively. Throughout Table 10, AP and nERR are more than twice as expensive as Q.

- In Table 10(II), the number of topics required by Q is $n = 50$ under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.25, 100)$. Thus, a typical TREC adhoc/web test collection with $n = 50$ topics is just good enough for guaranteeting a minimum detectable range of 0.25 in terms of Q for $m = 100$ systems under Cohen's convention. On the other hand, AP, nDCG and nERR do not pass the test as the required topic set sizes are 113, 60 and 118, respectively.

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

α	$minD$	$\beta = .10$	$\beta = .20$
		(I) $m = 10$	
.01	.02	11332/5014/6032/11815	9268/4104/4933/9663
	.05	1814/803/966/1891	1484/657/790/1547
	.10	454/201/242/473	372/165/198/388
	.20	114/51/61/119	94/42/50/98
	.25	73/33/39/77	60/27/33/63
.05	.02	8609/3809/4582/8976	**6759/2991/3598/7047**
	.05	1378/610/734/1437	**1082/479/576/1128**
	.10	345/153/184/360	**271/120/145/283**
	.20	87/39/46/90	**68/31/37/71**
	.25	56/25/30/58	**44/20/24/46**
		(II) $m = 100$	
.01	.02	27009/11949/14376/28160	22928/10144/12204/23905
	.05	4322/1912/2301/4506	3669/1624/1953/3826
	.10	1081/479/576/1127	918/407/489/957
	.20	271/120/144/282	230/102/123/240
	.25	174/77/93/181	148/66/79/154
.05	.02	21355/9448/11367/22266	**17504/7744/9317/18250**
	.05	3417/1512/1819/3563	**2801/1240/1491/2921**
	.10	855/379/455/891	**701/310/373/731**
	.20	214/95/114/223	**176/78/94/183**
	.25	137/61/73/143	**113/50/60/118**

Table 11: Topic set size table ($m = 10, 100$) for diversity/web ($l = 10$) with α-nDCG/nERR-IA/D-nDCG/D♯-nDCG.

α	$minD$	$\beta = .10$	$\beta = .20$
		(I) $m = 10$	
.01	.02	10771/11606/4857/7077	8809/9492/3973/5787
	.05	1724/1858/778/1133	1410/1520/637/927
	.10	432/465/195/284	353/381/160/233
	.20	109/117/49/72	89/96/41/59
	.25	70/75/32/46	57/62/26/38
.05	.02	8182/8817/3690/5376	**6424/6923/2897/4221**
	.05	1310/1411/591/861	**1029/1108/464/676**
	.10	328/353/148/216	**258/278/117/170**
	.20	82/89/38/54	**65/70/30/43**
	.25	53/57/24/35	**42/45/19/28**
		(II) $m = 100$	
.01	.02	25671/27662/11576/16865	21792/23483/9827/14317
	.05	4108/4427/1853/2699	3487/3758/1573/2291
	.10	1028/1107/464/675	872/940/394/573
	.20	257/277/116/169	219/236/99/144
	.25	165/178/75/109	140/151/64/92
.05	.02	20298/21872/9153/13335	**16637/17927/7502/10930**
	.05	3248/3500/1465/2134	**2662/2869/1201/1749**
	.10	812/875/367/534	**666/718/301/438**
	.20	204/219/92/134	**167/180/76/110**
	.25	131/141/59/86	**107/115/49/71**

The advantage of Q and nDCG over AP as demonstrated in both Table 9 (adhoc/news) and Table 10 (adhoc/web) strongly suggests the importance of utilising graded relevance assessments when the measurement depth is shallow. On the other hand, when the measurement depth is large (e.g., $l = 1000$), how many relevant documents have been retrieved, and at what positions, probably outweigh whether each document is highly or partially relevant.

Table 11 is the topic set size table for the diversity/web task with $l = 10$; note that this table discusses four *diversity* measures. It can be observed that:

- As D-nDCG and D♯-nDCG are substantially more stable than α-nDCG and nERR-IA (See Table 5(c)), they require substantially fewer topics under the same condition. For example, Table 11(II) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$ with $l = 10$, D-nDCG and D♯-nDCG

require only 301 and 438 topics, while α-nDCG and nERR-IA require as many as 666 and 718 topics, respectively. Throughout Table 11, α-nDCG and nERR-IA are more than twice as expensive as D-nDCG.

- In Table 11(II), the number of topics 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 guaranteeting a minimum detectable range of 0.25 in terms of D-nDCG for $m = 100$ 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 107, 115 and 71, respectively.

As was mentioned in Section 5.2, the statistical stability of D-nDCG arises from the fact that it lacks the per-intent diminshing return property: unlike nERR-IA and α-nDCG, it pays attention to every relevant document returned for each intent. Note that the statistical stability of an evaluation measure does not imply that the measure measures "what we want to measure."[10].

Figures 1-3 visualise the relationships between the required topic set size (n) and the number of systems to be compared (m) under $(\alpha, \beta, minD) = (0.05, 0.20, 0.05)$. For example, Figure 2 shows that, if we expect to compare $m = 200$ adhoc/web systems[11] under the above set of requirements, AP and nERR would require 3,819 and 3,982 topics, while Q and nDCG would require only 1,690 and 2,033 topics, respectively. Similarly, Figure 3 shows that, if we expect to compare $m = 200$ diversity/web systems under the above set of requirements, α-nDCG and nERR-IA would require 3,630 and 3,911 topics, while D-nDCG and D♯-nDCG would require only 1,637 and 2,385 topics, respectively.

From all of the results we have reported so far, it seems advisable to choose evaluation measures at the test collection design phase, as different evaluation measures have different variances and therefore require different topic set sizes under the same set of statistical requirements. Note that our method provides a method to compare evaluation measures in terms of *practical significance* [11], which in our case means the assessment cost. For example, while nERR has an intuitive user model (i.e., the diminishing return property, which says that the user does not value "redundant" documents), it is important to see beforehand that it can be twice as costly as some of the other alternatives.

6.3 Assessment Cost

The analysis in Section 6.2 covered adhoc/news, adhoc/web and diversity/web search tasks, but assumed that the pool depth was a given. In this section, we focus our attention to the adhoc/news task (with $l = 1000$), where we have depth-100 and depth-125 pools (See Table 3), which gives us the option of reducing the pool depth. Hence we can discuss the total assessment cost by multiplying n by the average number of documents that need to be judged per topic for a given pool depth pd.

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 variance estimates were obtained and pooled as described in Section 5.

Table 12 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, its 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 $pd = 10$. In our analysis discussed below, 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.

Figures 4-5 plot the required topic set size n against the average number of documents judged per topic, for $minD = 0.05, 0.10$ and $m = 10, 100$ under Cohen's five-eighty convention. Note that while the plots in the four figures look identical (as they should), the y-axis scales are very different. For example, Figure 5 (bottom) shows that, under $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.10, 100)$:

- From a purely statistical point of view, constructing $n = 569$ topics with pool depth $pd = 10$ is equivalent to constructing $n = 423$ topics with $pd = 100$ if AP is going to be used. The first option requires $569*96 = 54,624$ judgments, while the second option (i.e., the traditional TREC practice) requires $423 * 731 = 309,213$ judgments, which is 5.7 times as expensive.

- Because the variance of nERR is high regardless of the choice of pd (See Table 12), the plots for nERR form a near-horizontal line (i.e., n does not change with pd), and the required n is about twice as many as the other three measures.

It is a well-known fact that it is better to have many topics with few judgments per topic than to have few topics with many judgments per topic (e.g., [3, 4, 27]). Our present analysis confirms this through a simple visualisation with a theoretical underpinning by means of sample size design for ANOVA. Similar results can be obtained using sample size design for the t-test for $m = 2$. We encourage researchers who plan to build a test collection to use some past data and the expected number of systems to compare (m) and conduct an analysis such as the one we have demonstrated: then they can choose the combination of n and pd depending on the budget. For example, a researcher with a budget for 150,000 relevance assessments may look at Figure 5 (bottom) and decide to go with $pd = 30$ ($n = 476$) in order to use AP: the actual cost in this case would be about $476 * 253 = 120,428 < 150,000$ (See Table 12). While this is 2.2 times as expensive as the aforementioned case with $n = 569$, $pd = 10$, having more judgments is of course desirable if the test collection is going to be reused later.

7. CONCLUSIONS

We demonstrated how a researcher can design a new test collection for comparing $m(\geq 2)$ systems using power analysis and sample size design techniques with variance estimates from past data. While both t-test-based and ANOVA-based methods are available, the t-test approach suffers from the family-wise error rate problem and does not provide a completely sound solution to the evaluation of $m(> 2)$ systems. We thus recommend the use of our ANOVA-based method for designing new test collections based on statistical requirements and the expected number of systems to compare. Our

[10] Sakai and Song [22] reported that D-nDCG and D♯-nDCG outperform nERR-IA and α-nDCG in terms of the *concordance test*, that is, how often they agree with straightforward measures like precision and intent recall when two ranked lists are being compared.

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

Table 12: Number of relevance assessments and pooled $\hat{\sigma}^2$ for reduced pool depths with adhoc/news (measurement depth $l = 1000$).

Pool depth pd	TREC03new #judged for 50 topics	TREC04new #judged for 49 topics	Average judged/topic	Pooled $\hat{\sigma}^2$ AP	Q	nDCG	nERR
125	47,932	-	-	-	-	-	-
100	37,605	34,792	731	.0523	.0530	.0556	.1196
70	27,816	24,491	528	.0539	.0544	.0557	.1196
50	20,839	18,612	398	.0554	.0555	.0559	.1196
30	13,045	11,968	253	.0589	.0582	.0564	.1198
10	4,905	4,581	96	.0705	.0650	.0588	.1194

experiments with several IR tasks suggest that as different evaluation measures have different variances, test collection builders should carefully choose evaluation measures at the test collection design phase. We also showed how a cost analysis can be conducted through a pool depth reduction experiment using past test collections and runs. Although it is possible to reduce the assessment cost dramatically while preserving the statistical reliability by having many topics with shallow pools, the actual design should probably be determined based on the available budget if the test collection will be reused later. The relationship between statistical reliability and reusability will be examined in our future work.

Our methods can also be used to compare evaluation measures in terms of practical significance: "How much assessment cost will each of the candidate evaluation measures require under the same set of statistical requirements?" While *discriminative power* [19] is often used to compare the statistical reliability of evaluation measures, our methods can translate the reliability into actual cost.

The experiments reported in this paper are reproducible: the topic set size computation tools and all topic-by-run performance matrices used in this study are available from our website[12]. We encourage the interested reader to try computing topic set sizes for their own new test collections and evaluation measures, with their own variance estimates. We believe that improving test collection design based on past experience is important: perhaps it is time to stop producing (say) $n = 50$ topics every year without heeding statistical power. Note that, if the research community shares the basic ANOVA statistics used in Eq. 19, then variance estimates can easily be obtained from past data, and the estimation accuracy can be improved as we accumulate more test collections. This is much easier than sharing a repository of the actual run data, so we hope to put this into practice at evaluation venues such as TREC and NTCIR.

Finally, we stress again that our techniques and tools are applicable to any *non-IR* tasks (e.g., question answering, summarisation, machine translation, recommendation etc.) as well, as long as the paired t-test or ANOVA is applicable with the task and test collection in question.

Acknowledgement

This research is a part of Waseda University's project "Taxonomising and Evaluating Web Search Engine User Behaviours," supported by Microsoft Research.

8. 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, 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.

[12] `http://www.f.waseda.jp/tetsuya/tools.html` and `http://www.f.waseda.jp/tetsuya/data.html`

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

[5] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[6] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of ACM WSDM 2011*, pages 75–84, 2011.

[7] 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.

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

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

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

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

[12] 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.

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

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

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

[22] T. Sakai and R. Song. Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. *Information Retrieval*, 16(4):504–529, 2013.

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

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

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

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

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

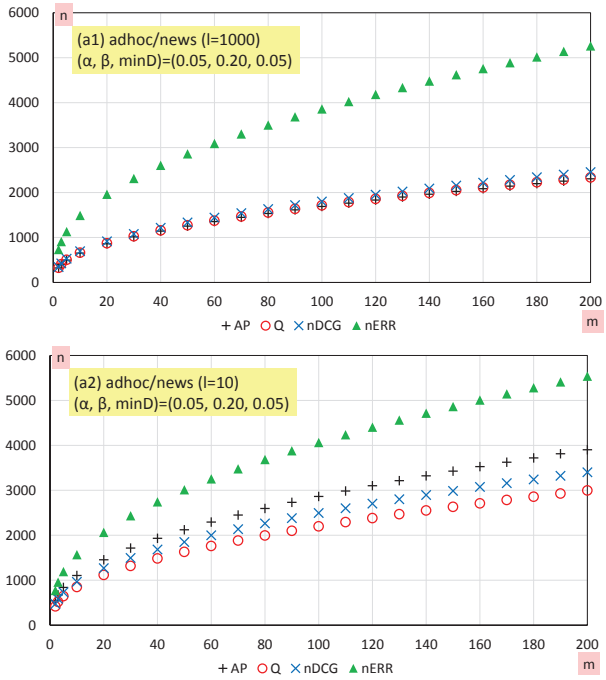

Figure 1: Required number of topics n against the number of systems (m), for adhoc/news with $(\alpha, \beta, minD) = (0.05, 0.20, 0.05)$.

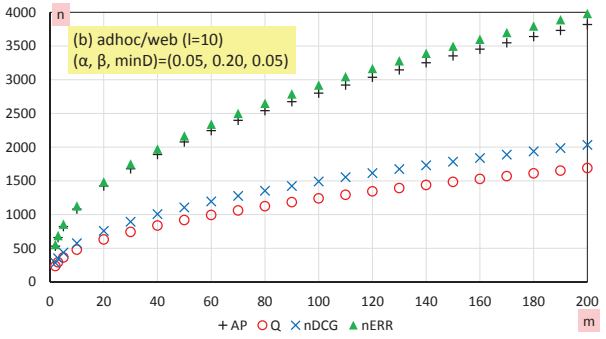

Figure 2: Required number of topics n against the number of systems (m), for adhoc/web with $(\alpha, \beta, minD) = (0.05, 0.20, 0.05)$.

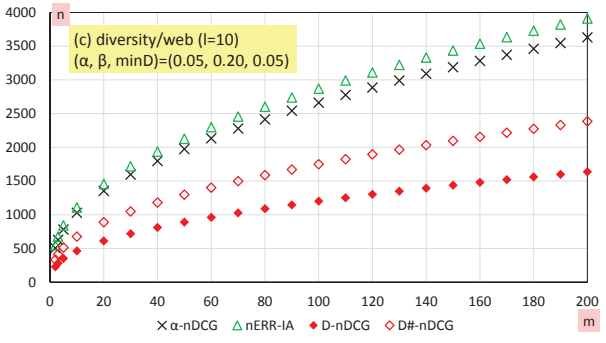

Figure 3: Required number of topics n against the number of systems (m), for diversity/web with $(\alpha, \beta, minD) = (0.05, 0.20, 0.05)$.

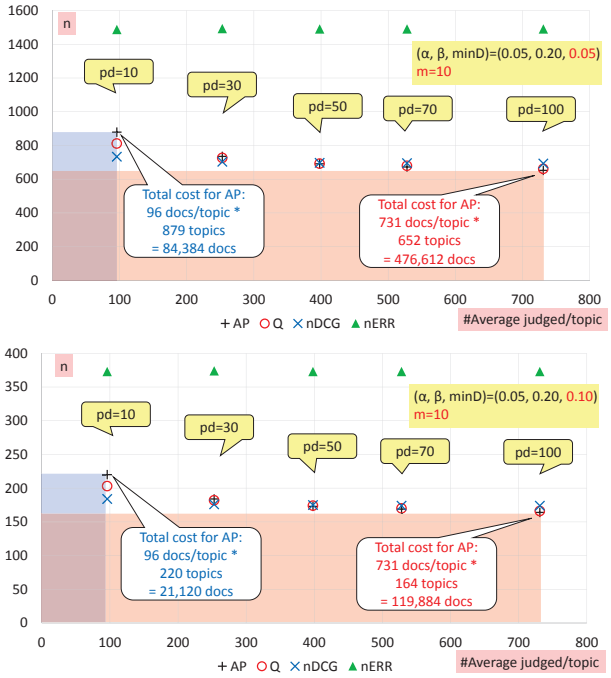

Figure 4: Required number of topics n against the average number of documents judged for a given pool depth, for adhoc/news ($l = 1000$) with $(\alpha, \beta) = (0.05, 0.20)$, $m = 10$.

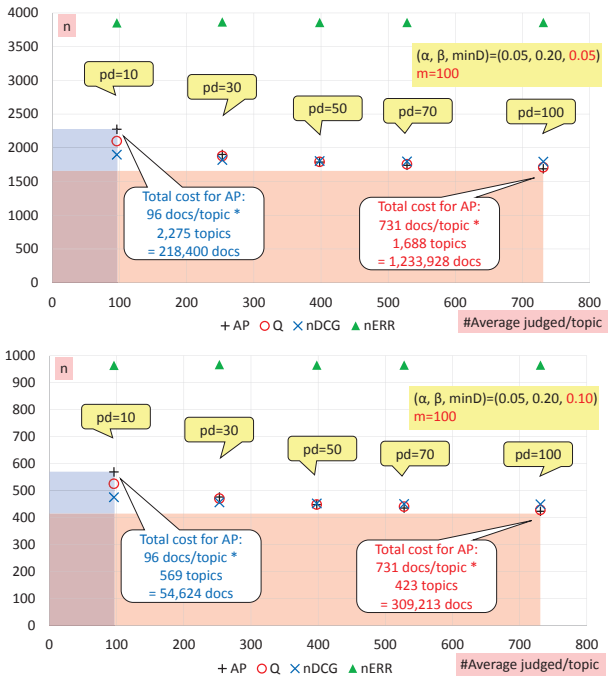

Figure 5: Required number of topics n against the average number of documents judged for a given pool depth, for adhoc/news ($l = 1000$) with $(\alpha, \beta) = (0.05, 0.20)$, $m = 100$.