RD-003

Designing Test Collections That Provide Tight Confidence Intervals

Tetsuya Sakai*

1. Introduction

In many research disciplines such as *information retrieval* (IR) and *natural language processing* (NLP), systems are often compared using a standard *test collection*. For example, in IR, two search engines X and Y may be compared using a test collection from TREC[1] or NTCIR[2] that has n *topics* (i.e., search requests), using some evaluation measure. Based on the evaluation scores $\{x_i\}$ and $\{y_i\}$ $(i = 1, \ldots, n)$, researchers can discuss the statistical significance of the mean performance difference. Since we know that the scores x_i and y_i correspond to each other for any i, *paired* significance tests may be applied (typically the paired t-test). This evaluation setting applies to many tasks, such as recommendation, summarisation, question answering and machine translation.

One problem with current test collection building practices is that the *topic set size* n is chosen *arbitrarily* based on budget contraints and/or intuition. For example, a typical TREC test collection has 50 topics: while $n > 25$ seems good enough for the normality assumption required by parametric tests, it is not at all clear why n should be 50, or why there should be $n = 50$ topics while the *pool depth* (*pd*) should be 100[3]. In this study, we show a method for determining the topic set size n for researchers who are trying to build a new test collection, by requiring a tight *confidence interval* (CI) for the difference between any given pair of systems X and Y. By applying this method with estimates of the population variance from past data (i.e., existing test collections), more reliable test collections can be built for similar tasks. Specifically, we show that evaluation measures should be chosen at the test collection design phase, and that a tight CI can be achieved at a low cost by having many topics with shallow document pools.

2. Related Work

Theoretically-grounded methods for determining the topic set size n are not known widely in our research communities even though we routinely create and use test collections for comparing systems. For example, a TREC track typically creates 50 topics every year; although a number of *post hoc* analyses have been conducted to see whether $n = 50$ is enough for obtaining reliable conclusions (e.g., [11]), the heuristics used there cannot answer the question: "What is the right topic set size for obtaining reliable conclusions?" or more specifically, "How many topics should we prepare for our *next* test collection?".

Webber, Moffat and Zobel [12] addressed the question of guaranteeing high *statistical power* in IR experimentation. The main focus of the study by Webber *et al.* was on how to estimate the population variance accurately by incrementally adding topics with relevance assessments to achieve the desired power. Moreover, their study was limited to the case of evaluation with Average Precision.

In the present study, we provide a methodology for determining the topic set size n for a new test collection to be built by requiring a tight CI for the performance difference between any given pair of systems, using statistical techniques described by Nagata [6]. In several research disciplines such as medicine and psychology where statistical methods are used heavily, the reporting of CIs and *effect sizes* (i.e., the *magnitude* of the difference between systems) is preferred over classical significance tests [3, 4, 5, 7]: it is known that p-values are not informative enough as they reflect not only the effect size but also the sample size n[4]. CIs provide estimates and their precisions at the same time, and enable visualisation of substantial differences (or the lack thereof) between systems.

3. Method

Section 3.1 describes Nagata's technique for determining the topic set size n to achieve a given requirement on the *margin of error* (MOE) of a CI [6]. As this method requires an estimate of the population standard deviation, Section 3.2 describes our method for obtaining the estimates. While the present study concerns topic set sizes for a few *IR* tasks, note that the methods presented in this paper are applicable wherever there are paired performance scores $\{x_i\}$ and $\{y_i\}$ for comparing systems X and Y, given some past data.

3.1 Determining the Topic Set Size

To build a CI for the difference between systems X and Y, we model the performance scores (assumed independent) as follows:

$$x_i = \mu_X + \gamma_i + \varepsilon_{Xi} \ , \quad \varepsilon_{Xi} \sim N(0, \sigma_X^2) \ , \qquad (1)$$

$$y_i = \mu_Y + \gamma_i + \varepsilon_{Yi} \ , \quad \varepsilon_{Yi} \sim N(0, \sigma_Y^2) \qquad (2)$$

where γ_i represents the topic effect and $\mu_\bullet, \sigma_\bullet^2$ represent the population mean and variance for X, Y, respectively $(i = 1, \ldots, n)$. To cancel out γ_i, let

$$d_i = x_i - y_i = \mu_X - \mu_Y + \varepsilon_{Xi} - \varepsilon_{Yi} \qquad (3)$$

so that $d_i \sim N(\mu, \sigma_t^2), \mu = \mu_X - \mu_Y, \sigma_t^2 = \sigma_X^2 + \sigma_Y^2$. It then follows that $t = \frac{\bar{d} - \mu}{\sqrt{V/n}} \sim t(n-1)$ (i.e., t distribution with $(n-1)$ degrees of freedom), where $\bar{d} = \sum_{i=1}^n d_i/n$ and $V = \sum_{i=1}^n (d_i - \bar{d})^2/(n-1)$. Hence, for a given significance criterion α, the following holds:

$$Pr[-t(n-1;\alpha) \leq t \leq t(n-1;\alpha)] = 1 - \alpha \qquad (4)$$

where $t(\phi; P)$ is the two-sided critital t value with ϕ degrees of freedom for probability P. (With Microsoft Excel, `TINV`(α, ϕ) or `T.INV.2T`(α, ϕ).) Hence,

$$Pr[\bar{d} - MOE \leq \mu \leq \bar{d} + MOE] = 1 - \alpha \qquad (5)$$

*Waseda University tetsuyasakai@acm.org

[1] http://trec.nist.gov/

[2] http://research.nii.ac.jp/ntcir/

[3] The *relevance assessments* of modern test collections are often collected by taking, for each topic, the top *pd* documents returned by each of the m participating systems. The total assessment cost is roughly proportional to $n * pd$ [8]: it is generally not directly proportional to m as the documents pooled from each system tend to overlap heavily.

[4] With a sufficiently large n, the difference between *any* two systems will be statistically significant [6].

Table 1: TREC test collections and runs used for estimating σ_t. The web track relevance grades [2] 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	528,155 (disks 4+5 minus
TREC04new	2004 robust	49 (651-700 minus 672)	78∗	100	L0-L2	the Congressional Record)
(b) task: adhoc/web						
TREC11w	2011 web - adhoc	50	37	25	L0-L3	approx. one billion
TREC12w	2011 web - adhoc	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)

∗This is the set of 78 runs used by Webber, Moffat and Zobel [12].

where

$$MOE = t(n-1;\alpha)\sqrt{V/n}\ . \tag{6}$$

Eq. 5 shows that the $100(1-\alpha)\%$ CI for the difference in population means ($\mu = \mu_X - \mu_Y$) is given by $[\bar{d} - MOE, \bar{d} + MOE]$. This much is very well known.

In this study, we determine the topic set size n by *requiring* that $2MOE \leq \delta$: that is, *the CI of the difference between X and Y should be no larger than some constant δ*. This ensures that experiments using the test collection will be conclusive whereover possible: for example, note that a wide CI that includes zero implies that we are very unsure as to whether systems X and Y actually differ. Since MOE (Eq. 6) contains a random variable (V), we actually impose the above requirement on the *expectation* of $2MOE$:

$$E(2MOE) = 2t(n-1;\alpha)\frac{E(\sqrt{V})}{\sqrt{n}} \leq \delta\ . \tag{7}$$

Now, it is known that

$$E(\sqrt{V}) = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}\sigma_t \tag{8}$$

where $\sigma_t = \sqrt{\sigma_X^2 + \sigma_Y^2}$ and $\Gamma(\bullet)$ is the gamma function[5]. (With Microsoft Excel, `GAMMA`($\bullet$).) By substituting Eq. 8 to Eq. 7, the requirement can be rewritten as:

$$\frac{t(n-1;\alpha)\Gamma(\frac{n}{2})}{\sqrt{n(n-1)}\Gamma(\frac{n-1}{2})} \leq \frac{\delta}{2\sqrt{2}\sigma_t}\ . \tag{9}$$

In order to find the smallest n that satisfies Eq. 9, we first consider an "easy" case where the population variance σ_t^2 is *known*. In this case, the MOE is given by $MOE_z = z_{\alpha/2}\sqrt{\sigma_t^2/n}$ (*cf.* Eq. 6), where z_P denotes the one-sided critical z value for probability P. (With Microsoft Excel, z_P = `NORMINV`$(1-P, 0, 1)$ = `NORM.S.INV`$(1-P)$.) By requiring that $2MOE_z \leq \delta$, we can obtain a tentative topic set size n':

$$n' \geq \frac{4z_{\alpha/2}^2\sigma_t^2}{\delta^2}\ . \tag{10}$$

First, the smallest integer that satisfies Eq. 10 can be tested to see if it satisfies Eq. 9; n' is incremented until it does. The resultant $n = n'$ is the topic set size we want.

We have devised a simple Excel file that automatically executes the above procedure to find the required sample size n, for any given combination of $(\alpha, \delta, \hat{\sigma}_t^2)$[6].

[5]Note that $\sqrt{V}$ is *not* an unbiased estimate of σ_t while V is an unbiased estimate of σ_t^2 (i.e., $E(V) = \sigma_t^2$) [6, 7].

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

Here, $\hat{\sigma}_t$ is an *estimate* of the population standard deviation σ_t obtained for a given IR task and an evaluation measure, using past data. The estimate is used to compute Eqs. 9 and 10. The next section describes, as case studies, how we obtained the $\hat{\sigma}_t$ values from past TREC data.

3.2 Estimating σ_t from Past Data

Given an existing data set C with n_C topics and a set of m *runs* (i.e., system output files) with their per-topic scores in terms of some evaluation measure, we can compute, for each of the $k = m(m-1)/2$ run pairs, the unbiased estimate of the population variance V_C^b ($b = 1, \ldots, k$) (computed similarly to V from Section 3.1). Following Webber *et al.* [12], we then take the *95th percentile* of the k variances as the estimate of the population variance, which we denote by $\hat{\sigma}_{t,C}^2$. Elsewhere, we shall discuss a more theoretically sound method for obtaining variance estimates, which utilises Analysis of Variance (ANOVA) statistics [7].

Given multiple existing data sets that represent a single IR *task* (e.g., adhoc news retrieval), we first obtain $\hat{\sigma}_{t,C}^2$ from each data set C, and then estimate the population variance via pooled variances as follows:

$$\hat{\sigma}_t^2 = \sum_C (n_C - 1)\hat{\sigma}_{t,C}^2 / \sum_C (n_C - 1)\ . \tag{11}$$

Table 1 shows some statistics of the six TREC data sets that we used for estimating the σ_t: the IR tasks we consider are (a) *adhoc news retrieval* (TREC robust tracks [10]); (b) *adhoc web search*; and (c) *adhoc diversity search* (TREC web tracks; *adhoc* and *diversity* [2]). For the adhoc tasks (a) and (b), we consider four evaluation measures: *Average Precision* (AP), *Q-measure* (Q), *normalised Discounted Cumulative Gain* (nDCG) and *normalised Expected Reciprocal Rank* (nERR): the `NTCIREVAL` toolkit is used for computing the measures[7]. For the diversity task (c), we consider four measures especially designed to balance relevance and diversity: *α-nDCG*, *ERR-IA*, *D-nDCG* and *D♯-nDCG*. The first two measures are computed using an official evaluation script from TREC[8], and the other two measures used at the *NTCIR INTENT task* [9] are computed using `NTCIREVAL`. The exact formulae for computing the evaluation measures can be found elsewhere (e.g., [8]); for the purpose of this study, it suffices to note here that Expected Reciprocal Rank (ERR) (which forms the basis of nERR and nERR-IA) is an evaluation measure suitable for *navigational* search intents: once a relevant

[7]`http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`

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

Table 2: Estimated σ_t for different evaluation measures with measurement depth l.

(a1) task: adhoc/news ($l = 1000$)				
Data	AP	Q	nDCG	nERR
TREC03new	.21	.20	.23	.41
TREC04new	.20	.20	.24	.43
Pooled	**.21**	**.20**	**.24**	**.42**
(a2) task: adhoc/news ($l = 10$)				
Data	AP	Q	nDCG	nERR
TREC03new	.32	.26	.27	.42
TREC04new	.30	.27	.29	.44
Pooled	**.31**	**.26**	**.28**	**.43**
(b) task: adhoc/web ($l = 10$)				
Data	AP	Q	nDCG	nERR
TREC11w	.34	.28	.29	.39
TREC12w	.37	.23	.25	.38
Pooled	**.36**	**.26**	**.27**	**.38**
(c) task: diversity/web ($l = 10$)				
Data	α-nDCG	nERR-IA	D-nDCG	D♯-nDCG
TREC11wD	.35	.37	.26	.31
TREC12wD	.32	.34	.25	.27
Pooled	**.34**	**.36**	**.25**	**.29**

Table 3: Topic set sizes for achieving $E(2MOE) \leq \delta$ at $\alpha = 0.05$.

(a1) task: adhoc/news ($l = 1000$)				
δ	AP	Q	nDCG	nERR
.05	273	248	-	-
.10	70	64	91	273
.15	33	30	42	123
.20	19	18	25	70
.25	13	12	17	46
(a2) task: adhoc/news ($l = 10$)				
δ	AP	Q	nDCG	nERR
.05	-	-	-	-
.10	150	106	123	287
.15	68	49	56	129
.20	39	28	33	73
.25	26	19	22	48
(b) task: adhoc/web ($l = 10$)				
δ	AP	Q	nDCG	nERR
.05	-	-	-	-
.10	202	106	114	224
.15	91	49	52	101
.20	52	28	30	58
.25	34	19	20	38
(c) task: diversity/web ($l = 10$)				
δ	α-nDCG	nERR-IA	D-nDCG	D♯-nDCG
.05	-	-	-	-
.10	180	202	98	132
.15	81	91	45	60
.20	47	52	26	35
.25	31	34	18	23

document is found, it basically ignores other relevant ones. This *diminishing return* property is intuitive but is known to make the measure unstable [8].

Table 2 shows the values of $\hat{\sigma}_t$ that we obtained using the above methods, for different evaluation measures, with the *measurement depths* [8] $l = 1000$ and $l = 10$ for adhoc news (Task (a)) and with $l = 10$ for the web tasks (Tasks (b) and (c)). Note that adhoc news retrieval typically evaluates the top $l = 1000$ documents, while web search evaluation generally focusses on the top ranked documents (e.g., the first search engine result page). Hereafter, we use the estimates based on the *pooled* variances shown in Table 2 to compute appropriate topic set sizes for each of the IR tasks (a1), (a2), (b) and (c).

4. Results and Discussions

4.1 Main Results

Table 3 summarises our experimental results for Tasks (a1)-(c), with the significance criterion $\alpha = 0.05$ and $\delta = .05, .10, .15, .20, .25$. Recall, for example, that $\delta = 0.10$ means that the CI of the difference between any two system means is given by $\bar{d} \pm 0.05$, or something narrower. For some cells, we could not compute the gamma function with Excel as n was too large ($n > 343$). We can observe the following:

Choice of δ: For some IR tasks, it may not be realistic to require $\delta = 0.05$. For Task (a1) (adhoc/news with $l = 1000$), Q-measure can achieve this goal with $n = 248$ topics. For the other three tasks with $l = 10$, the variances are too large and therefore the required topic set sizes will be over 343. Whereas, $\delta = 0.10$ seems like an achievable goal for all tasks.

Choice of measure: For the three adhoc tasks (a1), (a2) and (b), Q-measure requires fewer topics than AP and nDCG, while nERR clearly underperforms these three measures. For example, in Task (b) (adhoc/web), given $(\alpha, \delta) = (0.05, 0.10)$, Q requires only 106 topics, while AP and nERR require 202 and 224 topics, respectively. For Task (c) (diversity/web), D-nDCG requires far fewer topics than α-nDCG and nERR-IA: for example, given $(\alpha, \delta) = (0.05, 0.10)$, D-nDCG requires only 98 topics, while α-nDCG and nERR-IA require 180 and 202 topics, respectively.

It is clear that when designing a test collection, one should carefully consider which evaluation measures to use. Unstable measures have large variances and therefore require more topics.

4.2 Comparison of Judgement Costs

The results discussed in Section 4.1 were based on the official relevance assessment data: the pool depths used by TREC were treated as a given. However, the cost of constructing a test collection mainly arises from relevance assessment, whose cost is roughly proportional not only to the number of topics n but also to the pool depth pd. In this section, we focus on the adhoc/news retrieval task where the pool depth is typically much larger than that for web search tasks, and explore the balance between n and pd. To this end, we constructed *shallow-pool* versions of the TREC03new and TREC04new relevance assessment data: for example, from the original TREC04new relevance assessments that are based on depth-100 pools (See Table 1), we filtered out all topic-document pairs that were not contained in the top $pd(< 100)$ documents of any run to form depth-pd ($pd = 70, 50, 30, 10$) relevance assessment data. All the runs were then re-evaluated using the new assessment data.

Table 4 shows the total number of relevance assessments (i.e., number of topic-document pairs judged) for each pool depth considered; note that the original pool depth for TREC03new was 125 but the full assessments are not used in the present analysis. The "Average" row combines the statistics from the two data sets to compute the average number of documents judged for each pool depth. The table also shows the estimated standard deviations for AP, Q, nDCG and nERR, computed based on the shallow-pool versions of the relevance assessments. Naturally, standard deviations tend to increase as more documents are filtered out and the uncertainty (i.e. the number of "unjudged" documents) increases.

It should be noted that the $\hat{\sigma}_t$ for nERR stays at 0.42 while the pool depth is reduced from 100 to 10; similarly, the $\hat{\sigma}_t$ for nDCG stays at 0.24 while the pool depth is reduced from 100 to 30. Thus, from a purely statistical point of view, having a large pool depth to evaluate systems with these particular measures is a waste of considerable cost and effort.

For $(\alpha, \delta) = (0.05, 0.10)$, Figure 1 plots the required topic set size against the average number of documents

Table 4: Number of relevance assessments and pooled $\hat{\sigma}_t$ for reduced pool depths.

Pool depth pd	TREC03new #judged for 50 topics	TREC04new #judged for 49 topics	Average judged/topic	Pooled $\hat{\sigma}_t$ ($l = 1000$)			
				AP	Q	nDCG	nERR
125	47,932	-	-	-	-	-	-
100	37,605	34,792	731	.21	.20	.24	.42
70	27,816	24,491	528	.22	.21	.24	.42
50	20,839	18,612	398	.22	.22	.24	.42
30	13,045	11,968	253	.24	.23	.24	.42
10	4,905	4,581	96	.26	.24	.26	.42

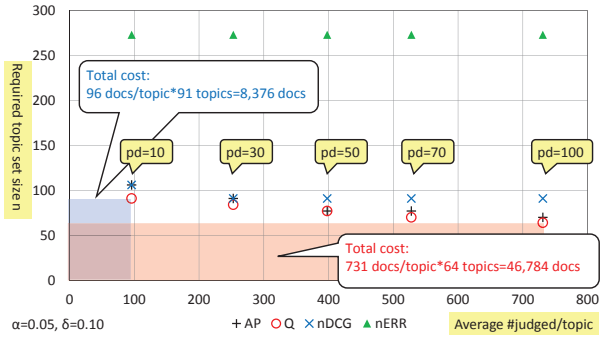

Figure 1: Required topic set size n against average #judged documents per topic based on standard deviation estimates shown in Table 4 ($\alpha = 0.05, \delta = 0.10$).

judged, based on the data shown in Table 4. For example, the leftmost data point for Q-measure is plotted as follows: when $pd = 10$ (i.e., #judged per topic is 96), $\hat{\sigma}_t = .24$, so we enter $(\alpha, \delta, \hat{\sigma}_t^2) = (0.05, 0.10, 0.24^2)$ into the aforementioned Excel tool and obtain $n = 91$. Thus, as the leftmost baloon in the figure shows, the total number of topic-document pairs that need to be assessed in this setting is $96 * 91 = 8,376$. Whereas, as the rightmost baloon shows, the number of topics required when when the pool depth is 100 (i.e., #judged per topic is 731) is $n = 64$ ($\hat{\sigma}_t = .24$), so the assessments amount to $731 * 64 = 46,784$; the cost is 5.6 times as high as the first case! While it has been known that it is better to have many topics with few judgments than to have few topics with many topics (e.g., [1, 12]), our approach provides a simple approach to quantifying the phenomenon from a particular point of view, namely, the requirement of a tight CI.

The triangles in Figure 1 represent nERR. For nERR, due to the aforementioned lack of sensitivity of its variance to pool depth changes, the required number of topics n stays at 273 for $pd = 10, \ldots, 100$. As for nDCG, n stays at 91 for $pd = 30, \ldots, 100$. Again, this is a lot of cost and effort wasted from a statistical point of view: for example, if nDCG is to be used with $n = 91$ topics for evaluating adhoc news retrieval, our results show that the same MOE will be obtained whether depth-100 pools are assessed or only depth-30 pools are assessed.

5. Conclusions

We showed a theoretically-grounded approach to determining the topic set size n by requiring a tight CI for comparing any given pair of systems. Our experiments (in a few IR contexts) suggest that test collections should be designed with specific evaluation measures in mind, and that cost analysis should be performed based on past data. We showed that $E(2MOE) \leq 0.10$ is an achievable requirement, and that a tight CI can be achieved at a low cost by having many topics with shallow document pools. Our approach thus enables *improving* test collection designs based on past experiences, and is applicable to any task involving paired evaluation scores $\{x_i\}, \{y_i\}$.

While this paper discussed how to determine the topic set size n for a new test collection by requiring a tight CI, it is also possible to systematically set n by requiring specific levels of *statistical power* and *effect size* [4, 6] in the context of the paired t-test or ANOVA. This alternative approach, together with a better method for estimating σ_t^2 from ANOVA statistics, will be discussed elsewhere.

Acknowledgement

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

References

[1] 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.

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

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

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

[5] 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.

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

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

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

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

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

[11] 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.

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