Proceedings of the Seventh International Workshop on Evaluating Information Access (EVIA 2016),
a Satellite Workshop of the NTCIR-12 Conference, June 7, 2016 Tokyo Japan

On Estimating Variances for Topic Set Size Design

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

Lifeng Shang
Huawei Noah's Ark Lab, Hong Kong.
shang.lifeng@huawei.com

ABSTRACT

Topic set size design is a suite of statistical techniques for determining the appropriate number of topics when constructing a new test collection. One vital input required for these techniques is an estimate of the population variance of a given evaluation measure, which in turn requires a topic-by-run score matrix. Hence, to build a new test collection, a pilot data set is a prerequisite. Recently, we ran an IR task at NTCIR-12 where the number of topics was actually determined using topic set size design with an initial pilot data set based on only five similar runs; a test collection was then constructed accordingly by pooling 44 runs from 16 participating teams for 100 topics. In this study, we treat the new test collection with the associated runs as a more reliable pilot data set to investigate how many teams and topics are actually necessary in the pilot data for obtaining accurate variance estimates.

Categories and Subject Descriptors

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

Keywords

evaluation; measures; statistical significance; test collections; topics; variances

1. INTRODUCTION

Information Retrieval (IR) test collections are built every year at evaluation conferences such as TREC, NTCIR and CLEF. The traditional aproach is to prepare (say) $n = 50$ topics every year, collect runs from participating teams to form a document pool for each topic, and construct "qrels" (i.e., relevance assessments for each topic) from the pools. However, system comparison experiments with $n = 50$ topics can often lead to conclusions that do not hold for other data (e.g., [12]), and a proper justification of n was in order. In light of this, Sakai [8] released simple Excel tools that enable researchers to easily conduct *topic set size design*, a suite of statistical techniques for determining the appropriate number of topics (n) when constructing a new test collection. With topic set size design, test collection builders can justify a particular choice of n, as discussed in Section 3. However, Sakai's tools require an estimate of the population variance of a given evaluation measure, which in turn requires a topic-by-run score matrix. Hence, to build a new test collection, a pilot data set is a prerequisite.

Recently, we ran an IR task at NTCIR-12 (namely, the Chinese subtask of the new Short Text Conversation task [10]) where the number of topics was actually determined using topic set size design with an initial pilot data set based on only five similar runs [9]; a test collection was then constructed accordingly by pooling 44 runs from 16 participating teams for 100 topics. To our knowledge, we are the first to adopt this method for an actual task at an IR evaluation conference. In this study, we treat the new test collection with the associated runs as a more reliable pilot data set to investigate how many teams and topics are actually necessary in the pilot data for obtaining accurate variance estimates. We demonstrate that it is possible to obtain accurate variance estimates if we have runs from a few different teams for about $n' = 25$ topics, although the accuracy depends on the stability of the evaluation measure chosen.

2. RELATED WORK

The TREC Million Query Tracks that ran from 2007 to 2009 [1] was a bold move to go beyond the "just build 50 topics with depth-100 pools" tradition. The main idea behind the tracks was strategic selection of documents to judge so that a given set of runs can be discriminated from one another; the objective was to economise relevance assessments rather than to ensure test collection reusability. In contrast, the topic set size design approach that we adopt is applicable to traditional fixed-depth pooling, and utilises classical statistical techniques to determine the number of topics; it aims to ensure either high statistical power or narrow confidence intervals for the purpose of comparing any future systems [8].

Like topic set size design, *generalisability theory* [11] may also be used for discussing the appropriate number of topics n. However, to our knowledge, no simple tools such as the ones provided for topic set design [8] are readily available. Both of these techniques have suggested that having 50 topics in an IR test collection may not be sufficient for reliable evaluation.

Webber, Moffat and Zobel [13] addressed the problem of ensuring high statistical power for a test collection to be built, in which they estimated the population variance of evaluation score deltas by taking the 95th percentile of the sample variances from past data. However, given a topic-by-run score matrix, an unbiased estimate of the within-system variance can easily be obtained as discussed in Section 3. The delta variance can then be obtained as double the within-system variance [8] (perhaps minus the covariance).

3. TOPIC SET SIZE DESIGN

Sakai's ANOVA-based tool[1] employs Nagata's *sample size design* techniques [5], and enables us to determine the number of topics n for a test collection to be built in order to ensure high statistical power for comparing m systems [8]. The tool requires the following as input:

α: Probability of *Type I error* (detecting a difference that does not exist).

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

β**:** Probability of *Type II error* (missing a difference that actually exists).

m**:** Number of systems that will be compared in one-way ANOVA ($m \geq 2$).

$minD$**:** *Minimum detectable range*. That is, whenever the performance difference between the best and the worst systems is $minD$ or higher, we want to ensure a *statistical power* of $(1-\beta)$ (i.e., the probability of detecting a difference that actually exists) given the significance level α.

$\hat{\sigma}^2$**:** Estimated variance of a system's performance, under the *homoscedasticity* assumption [8]. That is, it is assumed that the scores of the i-th system obey $N(\mu_i, \sigma^2)$, where μ_i's differ while σ^2 is common to all systems. This variance known to be heavily dependent on the evaluation measure.

Suppose that, using some pilot data and a particular evaluation measure, we have obtained an $n' \times m'$ topic-by-run matrix of scores x_{ij} ($i = 1, \ldots, m',\ j = 1, \ldots, n'$). The σ^2 for the evaluation measure can be estimated from ANOVA calculations as [8]:

$$\hat{\sigma}^2 = E(\sigma^2) = V_E = \frac{\sum_{i=1}^{m'} \sum_{j=1}^{n'} (x_{ij} - \bar{x}_{i\bullet})^2}{m'(n'-1)}, \qquad (1)$$

where $\bar{x}_{i\bullet} = \frac{1}{n'} \sum_{j=1}^{n'} x_{ij}$ (sample mean for the i-th run).

4. THE NTCIR-12 STC TASK

The Short Text Conversation (STC) Task was introduced at NTCIR-12; details of the task, including the official evaluation results, can be found in the task overview paper [10]. STC consists of Chinese and Japanese subtasks, but the present study discusses the Chinese subtask only. Below, we focus on those features of STC that are directly relevant to the present study.

4.1 Task Design and Data

STC is an IR task where, given a microblog *post* as the input query, the system searches a microblog repository and returns a ranked list of responses, or *comments*, that may serve as a valid human response to the input post. The retrieved comments are manually labelled $L0$ ("not acceptable"), $L1$ ("possibly acceptable") or $L2$ ("acceptable") by considering the following four aspects: *coherence*, *topical relevance*, *context independence* and *non-repetitiveness* [10]. The long-term goal of the task is to build and evaluate artificial intelligence systems that can respond sensibly and usefully to any microblog post; as the first step, we are evaluating the IR ("reuse-an-old-comment") approach to this problem rather than natural language response generation.

Participating teams are provided with the following: (a) a microblog repository (i.e., target corpus); (b) a set of test topics, which are microblog posts sampled from outside the above repository; and (c) a training data set that consists of post-comment pairs (p, c), where p is a post sampled from *outside* the repository (just like a test topic) and c is a "(possibly) acceptable" comment from *within* the repository. Thus, each pair is a positive example, meaning: "given a new post p, a past comment c can be reused to serve as a valid response."

Table 1 provides some statistics of the STC Chinese data set. In our previous work [9], we used the training data comprising $n' = 225$ topics with a set of $m' = 5$ baseline runs to create a 225×5 score matrix for each of our official evaluation measures. We then applied topic set size design (Section 3) and decided to have $n = 100$ test topics for the task. 16 participating teams submitted a total of 44 runs, each returning no more than 10 comments per topic. We set the pool depth to 10, thereby including all submitted comments in our pools. Multiple relevance assessors examined each pooled comment; details can be found in the task overview [10].

Table 1: The STC Chinese data set.

Repository	#posts	196,395
	#real post-comment pairs	5,648,128
Training data	#posts	225
	#labelled post-comment pairs	6,017 ($L2$: 974; $L1$: 3,028; $L0$: 2,015)
Test data	#posts	100
	#labelled post-comment pairs	26,096 ($L2$: 2,261; $L1$: 3,043; $L0$: 20,792)

4.2 Evaluation Measures

Our ultimate goal is to build artificial intelligence that can return *one* human-like response to a given post. For this reason, the official evaluation measures for the STC task are those designed for *navigational* search intents: *normalised gain* at 1 (nG@1), *P+* and *normalised Expected Reciprocal Rank* at 10 (nERR@10) [7, 9]. nG@1 is also known as nDCG@1, but called thus since neither discounting nor cumulation is applied at rank 1. P+ is a variant of the more well-known *Q-measure* (or just Q); both utilise the *blended ratio* which combines the binary precision and the *normalised cumulative gain* [4]. The difference is that while Q assumes a uniform stopping probability distribution over *all* relevant documents (just like Average Precision does), P+ assumes a uniform distribution over relevant documents ranked at or above the *preferred rank* r_p: this is the rank of the document with the highest relevance level in the list that is nearest to the top [6]. nERR is a measure with a unique property known as *diminishing return*: the value of a relevant document depends on the values of the other relevant documents ranked above it [2].

The above evaluation measures for navigational intents are known to be statistically less stable than those for informational intents such as nDCG (with a large cutoff) and Q. To be more specific, the within-system variances are relatively large for these official measures. Since the present study concerns estimation of population variances from known data, we also consider the above two measures in our experiments for comparison. The `NTCIREVAL` tool[2] was used for computing all evaluation measures, using the gain value of 3 for each $L2$ document and 1 for each $L1$ document. We use the "Microsoft version" of nDCG [7].

Table 2 compares the run rankings with different evaluation measures in terms of Kendall's τ, with 95% confidence intervals (CIs). Some of the τ values are in bold because the upper confidence limit exceeds 1: the two rankings are statistically equivalent. It can be observed, for example, that the rankings by P+ and nERR@10 (i.e., two official measures that consider the entire ranked list) are very similar, and so are those by Q@10 and nDCG@10 (i.e., two additional measures for informational intents).

5. EXPERIMENTS

We utilised our new test collection with the associated runs, with $n' = 100$ topics and $m' = 44$ runs from 16 teams, to investigate how many teams and topics are actually necessary as pilot data for obtaining accurate variance estimates.

5.1 Leaving Out k Teams

First, starting with the $n' = 100$ topics with the original qrels and the $m' = 44$ runs from the 16 teams, we gradually reduced the

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

Table 2: Run ranking similarity across the five measures: Kendall's τ values with 95% CIs.

	nG@1	P+	nERR@10	nDCG@10
P+	**.854** [.649, 1.059]	-	-	-
nERR @10	**.848** [.643, 1.053]	**.926** [.721, 1.131]	-	-
nDCG @10	.782 [.577, .987]	.784 [.579, .989]	**.841** [.636, 1.046]	-
Q @10	.757 [.552, .962]	.751 [.546, .956]	**.803** [.598, 1.008]	**.945** [.740, 1.150]

number of participating teams and examined its effect on the variance estimates as follows. From the set of 16 teams, we select k teams at random (for $k = 1, \ldots, 15$); we then remove their *unique contributions* from the official qrels [7], re-evaluate each run from the $(16 - k)$ teams with the new qrels, form a new topic-by-run matrix and obtain a variance estimate from it for each evaluation measure using Eq. 1. For each k, the above procedure is done 10 times to obtain 95% confidence intervals. When $k = 15$, note that we rely on exactly one team to estimate the variances; for two of the 10 trials with $k = 15$, we happened to rely on a team that submitted only one run ($m' = 1$). For these two cases, Eq. 1 reduces to a sample variance of a vector, but we did not exclude these extreme cases as they did not result in outliers. Also, removing teams sometimes meant that we also lost some *topics*, whenever all runs failed completely for these topics. For example, when $k = 15$, the *actual* number of topics n' used for estimating the population variance of nG@1 varied between 17 and 87 across the 10 trials.

Figure 1 visualises the results of our leave-k-out experiments. The x axes represent k (i.e., how many teams were removed), and the y axes represent $\hat{\sigma}^2$; error bars for $k = 1, \ldots, 15$ visualise 95% CIs. At $k = 0$, where the full 100×44 matrices with the official qrels are utilised, the point estimates are 0.114, 0.094 and 0.087 for nG@1, P+ and nERR@10, respectively. These are the most reliable estimates we have for the population variances. Whereas, the variance estimates we obtained in our previous work [9] using the 225×5 initial pilot data matrices were 0.152, 0.064 and 0.064, respectively; the discrepancies suggest that we cannot obtain highly accurate estimates if we rely on only one team (with only five similar runs), even if we have many topics. As discussed below, this is consistent with our observation from Figure 1, where our new test collection is now regarded as pilot data.

It can be observed that the 95% CIs keep widening as we remove more and more teams. In particular, if we rely on only one team ($k = 15$), the variance estimates vary considerably depending on exactly which team to rely on. However, as long as we rely on multiple teams, the variance estimates are quite robust to the reduction of the number of teams. For example, in Figure 1(a), the 95% CI for nG@1 misses the best point estimate (0.114) for the first time when $k = 9$ (i.e., using only seven teams); similarly, in (b) and (c), the 95% CI for P+ misses the best estimate (0.094) and the 95% CI for nERR@10 misses the best estimate (0.087) for the first time when $k = 14$ (i.e., using only two teams). It can also be observed that the CIs of P+ and nERR@10 are tighter than those of nG@1, and furthermore that the CIs of nDCG@10 and Q@10 (i.e., the measures for informational intents) are tighter than those of the three official measures. These differences across evaluation measures reflect how many data points (i.e., ranks of relevant documents) are taken into account in each measure. The 95% CI for nDCG@10 misses the best estimate (0.034) for the first time when $k = 10$, but is still quite accurate ($[0.035, 0.040]$); similarly, the 95% CI for Q@10 misses the best estimate (0.029) for the first time when $k = 11$, but is still quite accurate ($[0.030, 0.034]$).

Table 3: Effect on point estimates (σ^2's) with the full 16 teams as the pilot data topic set size n' is reduced (in bold). The numbers in parentheses are the corresponding new topic set sizes n for $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.15, 10)$.

measure	$n' = 100$	$n' = 75$	$n' = 50$	$n' = 25$	$n' = 10$
nG@1	**.114**	**.114**	**.118**	**.121**	**.113**
	(159)	(159)	(164)	(168)	(157)
P+	**.094**	**.094**	**.095**	**.097**	**.095**
	(131)	(131)	(132)	(135)	(132)
nERR@10	**.087**	**.086**	**.089**	**.090**	**.091**
	(121)	(120)	(124)	(125)	(127)
nDCG@10	**.034**	**.032**	**.034**	**.035**	**.043**
	(48)	(45)	(48)	(49)	(60)
Q@10	**.029**	**.028**	**.030**	**.029**	**.041**
	(41)	(40)	(42)	(41)	(58)

In summary, provided that we have enough pilot topics and reasonably stable evaluation measures, we can obtain highly accurate variance estimates by relying on a small number of teams. If only one or two teams are used, however, we may overestimate the variances, which will force us to create more topics than necessary.

5.2 Leaving Out l Topics and k Teams

Next, in order to explore minimal pilot data sizes for topic set design, we consider reducing the number of pilot topics (n') in addition to reducing the number of teams. This was achieved as follows. From each topic-by-run matrix discussed earlier, we randomly removed the vectors for 25, 50, 75, 90 topics to form new matrices with exactly $n' = 75, 50, 25, 10$ topics. For example, a matrix with 10 topics is a subset of another with 25 topics. New variance estimates were obtained from the reduced matrices; again, 10 trials were done for each k ($k = 1, \ldots, 15$) to obtain 95% CIs.

Table 3 shows the variance estimates (in bold) when n' is reduced gradually but the full 16 teams are kept. It can be observed that, except perhaps for the unstable nG@1, the estimates are quite accurate even with just $n' = 25$ topics. Whereas, when $n' = 10$, the variances for nDCG@10 and Q@10 are slightly overestimated. The numbers in parentheses show how these differences in variance estimates translate to the actual number of topics (n) recommended for a future test collection, under a particular set of statistical requirements $(\alpha, \beta, minD, m) = (0.05, 0.20, 0.15, 10)$ as considered in our previous work [9]. This setting means that we want to ensure 80% statistical power with 5% significance level (i.e., *Cohen's five-eighty convention* [3]) when comparing $m = 10$ systems where the best and the worst systems differ by at least $minD = 0.15$ for a given evaluation measure. For example, for Q@10, a variance of 0.029 (our best estimate with $n' = 100$) requires 41 topics, while an overestimate 0.041 ($n' = 10$) requires 58 topics (Table 3 bottom): hence the cost of overestimation would be relevance assessments for 17 topics.

Figure 2 visualises our leave-k-out results with only $n' = 10$ topics in a way similar to Figure 1 which used the full $n' = 100$ topics (minus those lost along with the removed teams). It can be observed that the variance estimates are now less stable, but that they are still quite robust to the reduction of the number of teams.

6. CONCLUSIONS

Topic set size design for a new test collection requires a variance estimate, which in turn requires a topic-by-run matrix with some pilot data. Our experimental results with the new Chinese STC

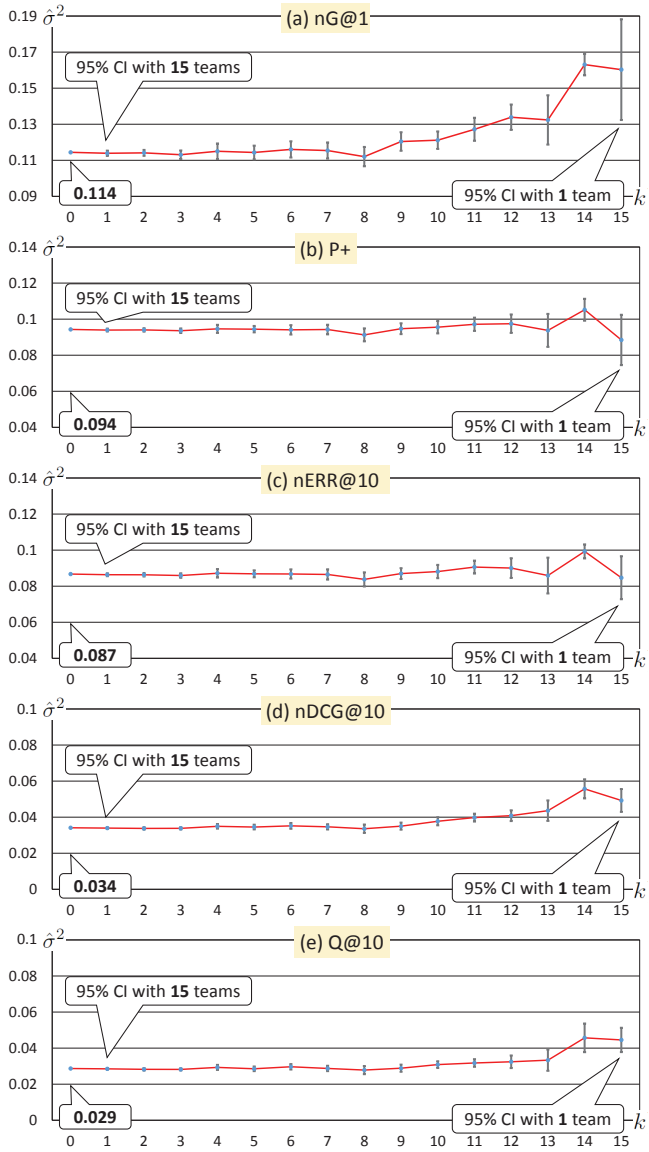

Figure 1: Results of leaving out k teams on the variance estimates, starting with 100 topics.

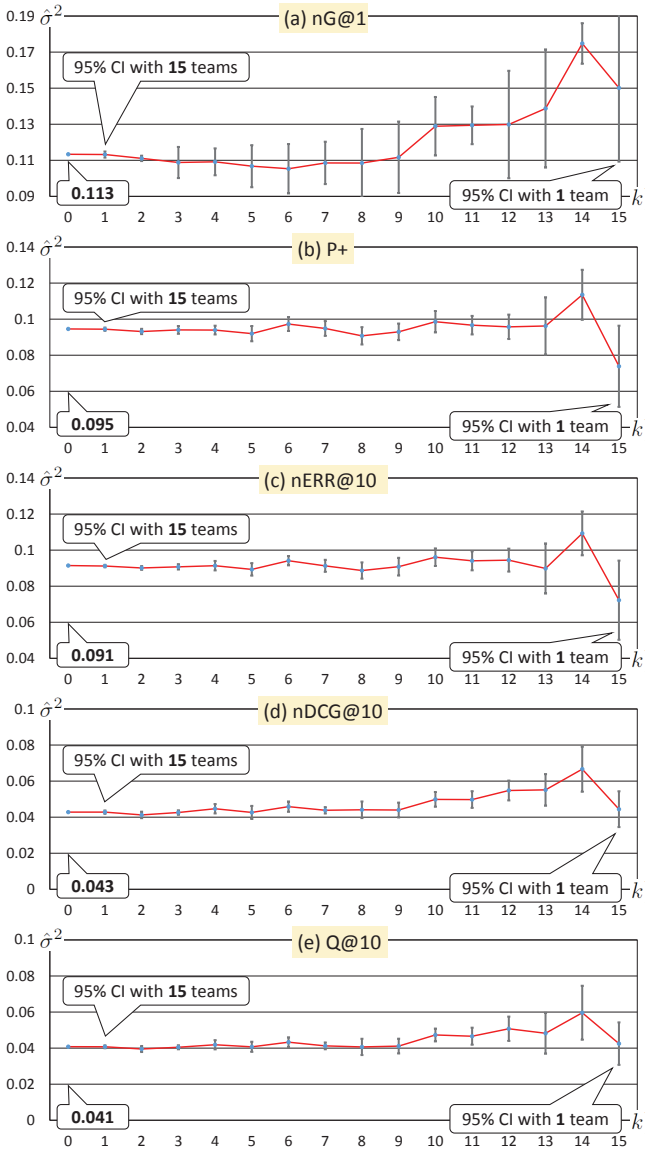

Figure 2: Results of leaving out k teams on the variance estimates, with 10 topics.

data suggest that we can obtain accurate variance estimates if we have runs from a few different teams for about $n' = 25$ topics, although the accuracy depends on the stability of the evaluation measure chosen. Having constructed a test collection, the new and better variance estimates obtained from it can be used for designing a test collection for the next round of the task.

7. REFERENCES

[1] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Million query track 2009 overview. In *Proceedings of TREC 2009*, 2010.

[2] 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.

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

[4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, 2002.

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

[6] T. Sakai. Bootstrap-based comparisons of IR metrics for finding one relevant document. In *AIRS 2006 (LNCS 4182)*, pages 374–389, 2006.

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

[8] T. Sakai. Topic set size design. *Information Retrieval Journal*, 2015.

[9] T. Sakai, L. Shang, Z. Lu, and H. Li. Topic set size design with the evaluation measures for short text conversation. In *Proceedings of AIRS 2015 (LNCS 9460)*, pages 319–331, 2015.

[10] L. Shang, T. Sakai, Z. Lu, H. Li, R. Higashinaka, and Y. Miyao. Overview of the ntcir-12 short text conversation task. In *Proceedings of NTCIR-12*, 2016.

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

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

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