The Effect of Score Standardisation on Topic Set Size Design

Tetsuya Sakai(✉)

Waseda University, Tokyo, Japan
tetsuyasakai@acm.org

Abstract. Given a topic-by-run score matrix from past data, *topic set size design* methods can help test collection builders determine the number of topics to create for a new test collection from a statistical viewpoint. In this study, we apply a recently-proposed *score standardisation* method called **std-AB** to score matrices before applying topic set size design, and demonstrate its advantages. For topic set size design, **std-AB** suppresses score variances and thereby enables test collection builders to consider realistic choices of topic set sizes, and to handle unnormalised measures in the same way as normalised measures. In addition, even discrete measures that clearly violate normality assumptions look more continuous after applying **std-AB**, which may make them more suitable for statistically motivated topic set size design. Our experiments cover a variety of tasks and evaluation measures from NTCIR-12.

1 Introduction

Given a topic-by-run score matrix from past data, *topic set size design* methods can help test collection builders determine the number of topics for a new test collection from a statistical viewpoint [8]. These methods enable test collection builders such as the organisers of evaluation conferences such as TREC, CLEF and NTCIR to improve the test collection design across multiple rounds of the tracks/tasks, through accumulation of topic-by-run score matrices and computation of better variance estimates.

In this study, we apply a recently-proposed *score standardisation* method called **std-AB** [7] to score matrices before applying topic set size design, and demonstrate its advantages. A standardised score for a particular topic means how different the system is from an "average" system in standard deviation units, and therefore enables cross-collection comparisons [14]. For topic set size design, **std-AB** suppresses score variances and thereby enables test collection builders to consider realistic choices of topic set sizes, and to handle unnormalised measures in the same way as normalised measures. In addition, even discrete measures that clearly violate normality assumptions look more continuous after applying **std-AB**, which may make them more suitable for statistically motivated topic set size design. Our experiments cover four different tasks from the recent NTCIR-12 conference[1]: MedNLP [1], MobileClick-2 [4], STC (Short Text Conversation) [11]

[1] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/index.html.

S. Ma et al. (Eds.): AIRS 2016, LNCS 9994, pp. 16–28, 2016.
DOI: 10.1007/978-3-319-48051-0_2

and QALab-2 [12], and some of the official evaluation measure scores from these tasks kindly provided by the task organisers.

2 Prior Art and Methods Applied

The present study demonstrates the advantages of our score standardisation method called **std-AB** [7] in the context of topic set size design, which determines the number of topics to be created for a new test collection [8]. This section situates these methods in the context of related work.

2.1 Power Analysis and Topic Set Size Design

Webber/Moffat/Zobel, CIKM 2008. Webber, Moffat and Zobel [15] proposed procedures for building a test collection based on power analysis. They recommend adding topics and conducting relevance assessments incrementally while examining the achieved *statistical power* (i.e., the probability of detecting a between-system difference that is real) and re-estimating the standard deviation σ_t of the between-system differences. They considered the comparison of two systems only and therefore adopted the t-test; they did not address the problem of the *family-wise error rate* [2,3]. Their experiments focused on Average Precision (AP), a binary-relevance evaluation measure. In order to estimate σ_t (or equivalently, the variance σ_t^2), they relied on empirical methods such as 95 %-percentile computation.

Sakai's Topic Set Size Design. Unlike the incremental approach of Webber *et al.* [15], Sakai's topic set size design methods seek to provide a straightforward answer to the following question: "I have a topic-by-run score matrix from past data and I want to build a new and statistically reliable test collection. How many topics should I create?" [8]. His methods cover not only the paired t-test but also one-way ANOVA for comparing more than two systems at the same time, as well as confidence interval widths. The present study focusses on the ANOVA-based approach, as it has been shown that the topic set sizes based on the other two methods can be deduced from ANOVA-based results. His ANOVA-based topic set size design tool[2] requires the following as input:

α, β**:** Probability of Type I error α and that of Type II error β.

m**:** Number of systems that will be compared ($m \geq 2$).

$minD$**:** *Minimum detectable range* [8]. That is, whenever the performance difference between the best and the worst systems is $minD$ or larger, we want to ensure $100(1-\beta)\%$ power given the significance level α.

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

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

Sakai recommends estimating within-system variances σ^2 for topic set size design using the sample residual variance V_E which can easily be obtained as a by-product of one-way ANOVA; it is known that V_E is an unbiased estimate of σ^2. Let x_{ij} denote the performance score for the i-th system with topic j ($i = 1, \ldots, m'$ and $j = 1, \ldots, n'$); let $\bar{x}_{i\bullet} = \frac{1}{n'} \sum_{j=1}^{n'} x_{ij}$ (sample system mean) and $\bar{x} = \frac{1}{m'n'} \sum_{i=1}^{m'} \sum_{j=1}^{n'} x_{ij}$ (sample grand mean). Then:

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

If there are more than one topic-by-run matrices available from past data, a pooled variance may be calculated to improve the accuracy of the variance estimate [8]. However, this is beyond the scope of the present study, as we are interested in obtaining a future topic set size based on a single matrix from NTCIR-12 for each measure in each task.

The present study uses the above method with *existing* NTCIR test collections and propose topic set sizes for the next NTCIR rounds. Sakai and Shang [9] considered the problem of topic set size design for a *new* task, where we can only assume the availability of a small pilot topic-by-run matrix rather than a complete test collection. Based on reduced versions of the NTCIR-12 STC official Chinese subtask topic-by-run matrices, they conclude that accurate variance estimates for topic set size design can be obtained if there are about $n' = 25$ topics and runs from only a few different teams.

2.2 Score Standardisation

Webber/Moffat/Zobel, SIGIR 2008. Webber, Moffat and Zobel [14] proposed *score standardization* for information retrieval evaluation with multiple test collections. Given m' runs and n' topics, a topic-by-run raw score matrix $\{raw_{ij}\}$ ($i = 1, \ldots, m'$, $j = 1, \ldots, n'$) is computed for a given evaluation measure. For each topic, let the sample mean be $mean_{\bullet j} = \frac{1}{m'} \sum_i raw_{ij}$, and the sample standard deviation be $sd_{\bullet j} = \sqrt{\frac{1}{m'-1} \sum_i (raw_{ij} - mean_{\bullet j})^2}$. The standardised score is then given by

$$std_{ij} = \frac{raw_{ij} - mean_{\bullet j}}{sd_{\bullet j}}, \qquad (2)$$

which quantifies how different a system is from the "average" system in standard deviation units. Using standardised scores, researchers can compare systems across different test collections without worrying about topic hardness (since, for every j, the mean $mean_{\bullet j}$ across runs is subtracted from the raw score) or normalisation (since the standardised scores, which are in the $[-\infty, \infty]$ range, are later mapped to the $[0, 1]$ range as described below). In practice, runs that participated in the pooling process for relevance assessments (*pooled systems*) can also serve as the runs for computing the *standardisation factors* ($mean_{\bullet j}, sd_{\bullet j}$) for each topic (*standardising systems*) [14]. The same standardisation factors are then used also for evaluating new runs.

In order to map the standardised scores into the $[0, 1]$ range, Webber *et al.* chose to employ the cumulative density function (CDF) of the standard normal distribution. The main reason appears to be that, after this transformation, a score of 0.5 means exactly "average" and that outlier data points are suppressed.

Our Method: Std-AB. Recently, we proposed to replace the aforementioned CDF transformation of Webber *et al.* [14] by a simple linear transformation [7]:

$$lin_{ij} = A * std_{ij} + B = A * \frac{raw_{ij} - mean_{\bullet j}}{sd_{\bullet j}} + B, \quad (3)$$

where A and B are constants. By construction, the sample mean and the standard deviation of std_{ij} over the known systems are 0 and 1, respectively ($j = 1, \ldots, n'$). It then follows that the sample mean and the standard deviation of lin_{ij} are B and A, respectively ($j = 1, \ldots, n'$). Regardless of what distribution raw_{ij} follows, Chebyshev's inequality guarantees that at least 89 % of the transformed scores lin_{ij} fall within $[-3A, 3A]$. In the present study, we let $B = 0.5$ as we want to assign a score of 0.5 to "average" systems, and let $A = 0.15$ so that the 89 % score range will be $[0.05, 0.95]$. Furthermore, in order to make sure that even outliers fall into the $[0, 1]$ range, we apply the following *clipping* step:

```
    **if** $lin_{ij} > 1$ **then** $lin_{ij} = 1$
    **else if** $lin_{ij} < 0$ **then** $lin_{ij} = 0$;
```

This means that *extremely* good (bad) systems relative to others are all given a score of 1 (0). Note that if A is too small, the achieved range of **std-AB** scores would be narrower than the desired $[0, 1]$; if it is too large, the above clipping would be applied to too many systems and we would not be able to distinguish among them. The above approach of using A and B with standardisation is quite common for comparing students' scores in educational research: for example, SAT (Scholastic Assessment Test) and GRE (Graduate Record Examinations) have used $A = 100, B = 500$ [5]; the Japanese *hensachi* ("standard score") uses $A = 10, B = 50$.

In our previous work [7], we demonstrated the advantages **std-AB** over the CDF-based method of Webber *et al.*: **std-AB** ensures pairwise system comparisons that are more consistent across different data sets, and is arguably more convenient for designing a new test collection from a statistical viewpoint. More specifically, using a small value of A ensures that the variance estimates $\hat{\sigma}^2$ will be small, which facilitates test collection design, as we shall demonstrate later. Moreover, as score *normalisation* becomes redundant if we apply standardisation [14], we can handle unnormalised measures (i.e., those that do not lie between 0 and 1). Furthermore, even discrete measures (i.e., those that only have a few possible values), which clearly violate the normality assumptions, look more continuous after applying **std-AB**. While our previous work was limited to the discussion of TREC robust track data and normalised ad hoc IR evaluation measures, the present study extends the work substantially by experimenting with four different NTCIR tasks with a variety of evaluation measures, including unnormalised and discrete ones for the first time.

3 NTCIR-12 Tasks Considered in the Present Study

The core subtask of the *MedNLPDoc* task is *phenotyping*: given a medical record, systems are expected to identify possible disease names by means of ICD (International Classification of Diseases) codes [1]. Systems are evaluated based on recall and precision of ICDs. MedNLPDoc provided us with a *precision* matrix with $n' = 78$ topics (i.e., medical records) and $m' = 14$ runs, as well as a *recall* matrix with $n' = 76$ topics and $m' = 14$ runs.

The *MobileClick-2* task evaluates search engines for smartphones. Systems are expected to output a two-layered textual summary in response to a query [4]. The basic evaluation unit is called *iUnit*, which is an atomic piece of factual information that is relevant to a given query. In the *iUnit ranking* subtask, systems are required to rank given iUnits by importance, and are evaluated by *nDCG* (normalised discounted cumulative gain) and *Q-measure*. In the *iUnit summarisation* subtask, systems are required to construct a two-layered summary from a given set of iUnits. The systems are expected to minimise the reading effort of users with different search intents; for this purpose the subtask employs a variant of the *intent-aware U-measure* [6], called *M-measure* [4], which is an unnormalised measure. MobileClick-2 provided us with 12 topic-by-run matrices in total: six from the English results and six from the Japanese results. While the variances of the unnormalised M-measure are too large for the topic set size design tool to handle, we demonstrate that the problem can be solved by applying **std-AB**.

The *STC* (Short Text Conversation) task requires systems to return a human-like response given a tweet (a Chinese Weibo post or a Japanese twitter post) [11]. Rather than requiring systems to generate natural language responses, however, STC makes them search a repository of past responses (posted in response to some other tweet in the past) and rank them. The STC Chinese subtask provided us with three matrices, representing the official results in *nG@1* (normalised gain at 1), *P+* (a variant of Q-measure), and *nERR@10* (normalised expected reciprocal rank at 10), all of which are navigational intent measures [10].

The *QALab-2* task tackles the problem of making machines solve university entrance exam questions. From the task organisers, we received two matrices based on National Center Test multiple choice questions, one for Phase-1 (where question types are provided to the system) and one for Phase-3 (where question types are not provided). As each topic is a multiple choice question, the evaluation measure is “Boolean” (either 0 or 1).

nG@1 for STC takes only three values: 0, 1/3 or 1 [10], and Boolean for QAlab-2 takes only two values: 0 or 1. These clearly violate the normality assumptions behind ANOVA: $x_{ij} \sim N(\mu_i, \sigma^2)$ for each system i. Thus, it should be noted that, when we apply topic set size design using the variances of these *raw* measures, what we get are topic set sizes for some normally distributed measure M that happens to have the same variance as that discrete measure, rather than topic set sizes for that measure per se. Whereas, if we apply **std-AB**, these measures behave more like continuous measures, as we shall demonstrate later.

Table 1. Columns (e) and (f) show within-system variance estimates $\hat{\sigma}^2$ based on the NTCIR-12 topic-by-run matrices and their **std-AB** versions. The values in bold are those plugged into the topic set design tool in this study. Column (g) compares the system rankings before and after applying **std-AB** in terms of Kendall's τ, with 95 % confidence intervals.

(a) Task/subtask	(b) Measure	(c) m'	(d) n'	(e) $\hat{\sigma}^2$ (raw scores)	(f) $\hat{\sigma}^2$ (**std-AB**)	(g) τ [95 %CI]
MedNLPDoc	precision	14	78	.0597	.0139	.978 [.585, 1.371]
	recall	14	76	**.0601**	**.0127**	.956 [.563, 1.349]
MobileClick	Q-measure	25	100	.0023	.0211	.867 [.587, 1.147]
iUnit ranking	nDCG@3	25	100	**.0259**	**.0215**	.720 [.440, 1.000]
(English)	nDCG@5	25	100	.0198	.0214	.713 [.433, .993]
	nDCG@10	25	100	.0141	.0212	.773 [.493, 1.053]
	nDCG@20	25	100	.0077	.0211	.853 [.573, 1.133]
(Japanese)	Q-measure	12	100	.0189	.0155	.970 [.537, 1.403]
	nDCG@3	12	100	**.0570**	**.0176**	.970 [.537, 1.403]
	nDCG@5	12	100	.0466	.0173	.909 [.476, 1.342]
	nDCG@10	12	100	.0355	.0163	.970 [.537, 1.403]
	nDCG@20	12	100	.0276	.0159	1 [.567,1.433]
MobileClick iUnit summarisation						
(English)	M-measure	16	100	44.3783	**.0072**	.983 [.620, 1.346]
(Japanese)	M-measure	13	100	93.5109	**.0077**	.949 [.537, 1.361]
STC (Chinese)	nG@1	44	100	**.1144**	**.0193**	.884 [.679, 1.089]
	P+	44	100	.0943	.0186	.962 [.757, 1.167]
	nERR@10	44	100	.0867	.0182	.947 [.742, 1.152]
QALab Phase-1	Boolean	27	41	.2124	.0191	.892 [.624, 1.160]
Phase-3	Boolean	34	36	**.2130**	**.0204**	.964 [.728, 1.200]

4 Results and Discussions

4.1 Results Overview

Table 1 Columns (e) and (f) show the variance estimates obtained by applying Eq. 1 to the aforementioned topic-by-run matrices, before and after performing **std-AB** as defined by Eq. 3. It can be observed that the variances are substantially smaller after applying **std-AB**. This means that the required topic set sizes will be smaller, provided that the tasks take up the habit of using **std-AB** measures. For each subtask (and language), we selected the *largest* raw score variance, shown in bold in Column (e), and plugged into the topic set size design tool (except for the unnormalised M-measure, whose variances were too large for the tool to handle); that is, we focus on the least stable measures to obtain topic set sizes that are reliable enough for all evaluation measures. We then used the variances of the corresponding **std-AB** measures, shown in bold in Column (f).

Currently, there is no task at NTCIR that employs score standardisation. Now, how would **std-AB** actually affect the official results? Table 1 Column (g) compares the run rankings before and after applying **std-AB** in terms of Kendall's τ for each evaluation measure in each subtask. The 95 % confidence intervals show that the two rankings are statistically equivalent for all cases, except for nDCG@5 in MobileClick English iUnit ranking whose 95 % CI is [.433, .993]. These results suggest that, by and large, **std-AB** enables cross-collection comparisons without affecting within-collection comparisons.

Table 2. Recommended topic set sizes for four NTCIR-12 Tasks ($\alpha = 0.05, \beta = 0.80$).

(I) MedNLPDoc	(a) raw recall ($\hat{\sigma}^2 = \mathbf{.0601}$)				(b) **std-AB** recall ($\hat{\sigma}^2 = \mathbf{.0127}$)			
$m \downarrow minD \rightarrow$	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20
2	2301	369	93	24	487	79	20	6
10	4680	750	188	48	990	159	40	11
20	6159	986	247	62	1302	209	53	14
30	7262	1163	291	73	1535	246	62	16
50	8986	1438	360	91	1899	305	77	20

	iUnit Ranking								iUnit Summarisation			
(II) MobileClick English	(a) raw nDCG@3 ($\hat{\sigma}^2 = \mathbf{.0259}$)				(b) **std-AB** nDCG@3 ($\hat{\sigma}^2 = \mathbf{.0215}$)				(c) **std-AB** M-measure ($\hat{\sigma}^2 = \mathbf{.0072}$)			
$m \downarrow minD \rightarrow$	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20
2	992	159	41	11	824	133	34	9	276	45	12	4
10	2017	323	82	21	1675	269	68	18	561	91	23	6
20	2655	425	107	27	2204	353	89	23	739	119	30	8
30	3130	501	126	32	2598	416	105	27	871	140	36	9
50	3873	620	156	39	3215	515	129	33	1077	173	44	12
(III) MobileClick Japanese	(a) raw nDCG@3 ($\hat{\sigma}^2 = \mathbf{.0570}$)				(b) **std-AB** nDCG@3 ($\hat{\sigma}^2 = \mathbf{.0176}$)				(c) **std-AB** M-measure ($\hat{\sigma}^2 = \mathbf{.0077}$)			
$m \downarrow minD \rightarrow$	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20
2	2182	350	88	23	674	109	28	8	296	48	13	4
10	4439	711	178	45	1371	220	56	15	600	97	25	7
20	5842	935	234	59	1804	289	73	19	790	127	32	9
30	6887	1103	276	70	2127	341	86	22	931	150	38	10
50	8522	1364	342	86	2632	422	106	27	1152	185	47	12

(IV) STC	(a) raw nG@1 ($\hat{\sigma}^2 = \mathbf{.1144}$)				(b) **std-AB** nG@1 ($\hat{\sigma}^2 = \mathbf{.0193}$)			
$m \downarrow minD \rightarrow$	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20
2	4379	701	176	45	739	119	30	8
10	8908	1426	357	90	1504	241	61	16
20	11724	1876	470	118	1979	317	80	21
30	13822	2212	554	139	2333	374	94	24
50	17104	2737	685	172	2886	462	116	30

(V) QALab	(a) raw Boolean ($\hat{\sigma}^2 = \mathbf{.2130}$)				(b) **std-AB** Boolean ($\hat{\sigma}^2 = \mathbf{.0204}$)			
$m \downarrow minD \rightarrow$	0.02	0.05	0.10	0.20	0.02	0.05	0.10	0.20
2	8152	1305	327	82	782	126	32	9
10	16585	2654	664	167	1589	255	64	17
20	21828	3493	874	219	2091	335	84	22
40	28992	4639	1160	291	2777	445	112	29
50	31845	5096	1275	319	3051	489	123	31

Table 2 shows the recommended topic set sizes with $\alpha = 0.05, \beta = 0.20$ (Cohen's five-eighty convention [3]), for several values of m (i.e., number of systems to be compared) and $minD$ (i.e., minimum detectable range), based on the variances shown in bold in Table 1. It should be noted first, that the values of $minD$ are not comparable across Parts (a) and (b). For example, a $minD$ of 0.02 with raw scores and a $minD$ of 0.02 with **std-AB** scores are not equivalent, because **std-AB** applies score standardisation (Eq. 2) followed by a linear transformation (Eq. 3). Nevertheless, it can be observed that, after applying **std-AB**, the choices of topic set sizes look more realistic. For example, let us consider the $m = 2$ row in Table 2(I). If we want to guarantee 80 % power whenever the difference between the two systems is $minD = 0.05$ (i.e., 5 % of the score range) or larger in raw recall, we would require 369 topics. Whereas, if we want to guarantee 80 % power whenever the difference between the two systems is $minD = 0.05$ (i.e., 5 % of the score range) or larger in **std-AB** recall, we would require only 79 topics. Although the above two settings of *minD* mean different things, the latter is much more practical. In other words, while ensuring 80 % power for a *minD* of 0.05 in raw recall is not realistic, ensuring the same power for a *minD* of 0.05 in **std-AB** is.

Figure 1 visualises the per-topic scores before and after applying **std-AB** for some of our data. Below, we discuss the effect of **std-AB** on recommended topic set sizes for each task in turn.

4.2 Recommendations for MedNLPDoc

The effect of **std-AB** on the recall scores from MedNLPDoc can be observed by comparing Fig. 1(a) and (a'). Note that while many of the raw recall values are 0's, all values are positive after applying **std-AB**. Moreover, there are fewer 1's after applying **std-AB**.

From Table 2(I), a few recommendations for a future MedNLPDoc test collection would be as follows. If the task is continuing to use raw recall, then:

- Create 100 topics: this guarantees 80 % power for comparing any $m = 2$ systems with a *minD* of 0.10 (93 topics are sufficient), and for comparing any $m = 50$ systems with a *minD* of 0.20 (91 topics are sufficient);
- Create 50 topics: this guarantees 80 % power for comparing $m = 10$ systems with a *minD* of 0.20 (48 topics are sufficient).

Whereas, if the task adopts **std-AB** recall, then:

- Create 80 topics: this guarantees 80 % power for comparing $m = 2$ systems with a *minD* of 0.05 (79 topics are sufficient), and for comparing $m = 50$ systems with a *minD* of 0.10 (77 topics are sufficient).

Note that MedNLPDoc actually had 76–78 topics (Table 1(d)), and therefore that the above recommendation is quite practical.

4.3 Recommendations for MobileClick-2

The effect of **std-AB** on the nDCG@3 scores from MobileClick-2 iUnit ranking (English) can be observed by comparing Fig. 1(b) and (b'). It can be observed that, after applying **std-AB**, the scores are more evenly distributed within the $[0, 1]$ range. Similarly, the effect of **std-AB** on the unnormalised M-measure from MobileClick-2 iUnit summarisation (English) can be observed by comparing Fig. 1(c) and (c'). Note that the scale of the y-axis for Fig. 1(c) is very different

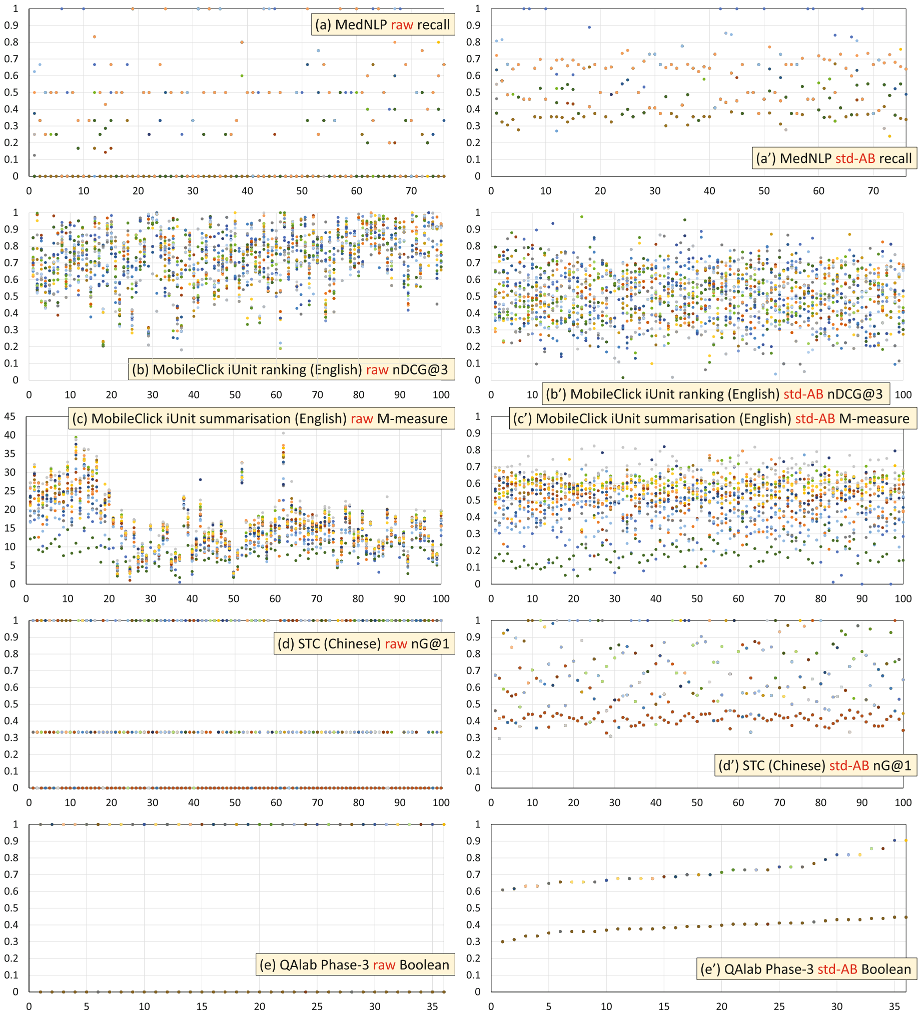

Fig. 1. Per-topic raw and **std-AB** scores for selected NTCIR-12 tasks. The horizontal axes represent topics. Different colours represent different runs (best viewed in colour).

from others. Despite this, Fig. 1(c') shows that **std-AB** transforms the scores into the $[0, 1]$ range without any problems. In this way, **std-AB** can handle any unnormalised measure. Put another way, if we take up the habit of using **std-AB** scores, normalisation becomes no longer necessary.

Since MobileClick-2 is a multilingual task, let us discuss topic set sizes that work for both English and Japanese. Moreover, since the topic set is shared across the iUnit ranking and summarisation subtasks, we want topic set sizes that work across these two subtasks. From Table 2(II) and (III), a few recommendations for future MobileClick test collections would be as follows. If the task is continuing to use raw nDCG@3, then:

- Create 90 topics: this guarantees 80 % power for comparing any $m = 10$ English iUnit ranking systems with a *minD* of 0.10 (82 topics are sufficient), and for comparing any $m = 2$ Japanese iUnit ranking systems with a *minD* of 0.10 (88 topics are sufficient).

However, the above setting cannot guarantee anything for the iUnit summarisation task, due to the use of the unnormalised M-measure. In contrast, if the tasks adopts **std-AB** nDCG@3 and **std-AB** M-measure, then:

- Create 100 topics: this guarantees 80 % power for comparing any $m = 20$ English iUnit ranking systems with a *minD* of 0.10 (89 topics are sufficient), and for comparing any $m = 30$ Japanese iUnit ranking systems with a *minD* of 0.10 (86 topics are sufficient), and for comparing any $m = 10$ English iUnit summarisation systems with a *minD* of 0.05 (91 topics are sufficient), and for comparing any $m = 10$ Japanese iUnit summarisation systems with a *minD* of 0.05 (97 topics are sufficient).

Thus being able to handle unnormalised measures just like normalised measures seems highly convenient. Also, recall that MobileClick-2 actually had 100 topics.

4.4 Recommendations for STC

The effect of **std-AB** on the nG@1 scores from STC (Chinese) can be observed by comparing Fig. 1(d) and (d'). It can be verified from Fig. 1(d) that nG@1 indeed take only three values: 0, 1/3 and 1. Whereas, Fig. 1(d') shows that **std-AB** nG@1 is more continuous, and that there are fewer 1's, and no 0's.

From Table 2(IV), a few recommendations for a future STC test collection would be as follows. If the task is continuing to use raw nG@1, then:

- Create 120 topics: this guarantees 80 % power for comparing any $m = 20$ systems with a *minD* of 0.20 (118 topics are sufficient);
- Create 90 topics: this guarantees 80 % power for comparing any $m = 10$ systems with a *minD* of 0.20 (exactly 90 topics are needed).

But note that, strictly speaking, the above recommendations are for normally distributed measures that have a variance similar to that of nG@1, since nG@1 takes only three values. Whereas, if the tasks adopts **std-AB** nG@1, then:

- Create 100 topics: this guarantees 80 % power for comparing any $m = 30$ systems with a *minD* of 0.10 (94 topics are sufficient).

The STC task actually had 100 topics; this was actually a decision based on topic set size design with raw evaluation measures and pilot data [10].

4.5 Recommendations for QALab

The effect of **std-AB** on the Boolean scores from QALab Phase-3 can be observed by comparing Fig. 1(e) and (e'). It can be observed that **std-AB** transforms the raw Boolean scores (0's and 1's) into something a little more continuous, but that the resultant scores still fall into two distinct score ranges; hence our topic set size design results for QALab should be taken with a large grain of salt even after applying **std-AB** as the scores are clearly not normally distributed. The reason why the **std-AB** scores are monotonically increasing from left to right is just that the QALab organisers sorted the topics by the number of systems that correctly answered them before providing the matrices to the present author. This is equivalent to sorting the topics by $mean_{\bullet j}$ (in decreasing order, i.e., easy topics first).

From Table 2(V), a few recommendations for a future STC test collection would be as follows. If the task is continuing to use raw Boolean, then:

- Create 90 topics: this guarantees 80 % power for comparing any $m = 2$ systems with a *minD* of 0.20 (82 topics are sufficient).

Whereas, if the tasks adopts **std-AB** Boolean, then:

- Create 40 topics: this guarantees 80 % power for comparing any $m = 2$ systems with a *minD* of 0.10 (32 topics are sufficient), or any $m = 50$ systems with a *minD* of 0.20 (31 topics are sufficient).

But recall that the above recommendations are for normally distributed measures whose variances happen to be similar to those of the Boolean measures.

QALab-2 Phase-3 actually had 36 topics only. Note that $n = 36$ is not satisfactory in any of the settings shown in Table 2(V)(a); $n = 36$ does not even satisfy the suggested setting shown above for (a normally distributed equivalent of) **std-AB** Boolean. These results suggest that the QALab task should have more topics to ensure high statistical power.

5 Conclusions and Future Work

Using topic-by-run score matrices from the recent NTCIR-12 MedNLPDoc, MobileClick-2, STC and QALab tasks, we conducted topic set design experiments with and without score standardisation and demonstrated the advantages of employing **std-AB** in this context. It is clear from our results that **std-AB** suppresses score variances and thereby enables test collection builders to consider realistic choices of topic set sizes, and that it can easily handle even unnormalised measures such as M-measure. Other unnormalised measures such as Time-Biased Gain [13], U-measure [6] and those designed for diversified search may be handled similarly. Furthermore, we have demonstrated that discrete measures such as nG@1, which clearly violate the normality assumptions, can be "smoothed" to some extent by applying **std-AB**. Recall that topic set size design assumes that the scores are indepent and identically distributed: that the scores for system i obey $N(\mu_i, \sigma^2)$. While this is clearly a crude assumption especially for unnormalised and discrete measures, **std-AB** makes it a little more believable at least, as shown in the right half of Fig. 1.

In our previous work [7], we performed a preliminary investigation into the robustness of standardisation factors $mean_{\bullet j}, sd_{\bullet j}$ for handling unknown runs (i.e., those that contributed to neither pooling nor the computation of standardising factors). However, our experiments were limited to handling unknown runs from the *same* round of TREC. Hence, to examine the longevity of standardisation factors over technological advances, we have launched a new web search task at NTCIR, which we plan to run for several years[3]. The standardisation factors obtained from the first round of this task will be compared to those obtained from the last round: will the initial standardisation factors hold up against the latest, more advanced systems?

Acknowledgement. We thank the organisers of the NTCIR-12 MedNLPDoc, QALab-2, MobileClick-2, and STC tasks, in particular, Eiji Aramaki, Hideyuki Shibuki, and Makoto P. Kato, for providing us with their topic-by-run matrices of the official results prior to the NTCIR-12 conference.

References

1. Aramaki, E., Morita, M., Kano, Y., Ohkuma, T.: Overview of the NTCIR-12 MedNLPDoc task. In: Proceedings of NTCIR-12, pp. 71–75 (2016)
2. Carterette, B.: Multiple testing in statistical analysis of systems-based information retrieval experiments. ACM TOIS **30**(1) (2012). Article No. 4
3. Ellis, P.D.: The Essential Guide to Effect Sizes. Cambridge University Press, Cambridge (2010)
4. Kato, M.P., Sakai, T., Yamamoto, T., Pavlu, V., Morita, H., Fujita, S.: Overview of the NTCIR-12 MobileClick task, pp. 104–114 (2016)
5. Lodico, M.G., Spaulding, D.T., Voegtle, K.H.: Methods in Educational Research, 2nd edn. Jossey-Bass, San Francisco (2010)
6. Sakai, T.: How intuitive are diversified search metrics? Concordance test results for the diversity U-measures. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 13–24. Springer, Heidelberg (2013)
7. Sakai, T.: A simple and effective approach to score standardisation. In: Proceedings of ACM ICTIR 2016 (2016)
8. Sakai, T.: Topic set size design. Inf. Retr. **19**(3), 256–283 (2016)
9. Sakai, T., Shang, L.: On estimating variances for topic set size design. In: Proceedings of EVIA 2016 (2016)
10. Sakai, T., Shang, L., Lu, Z., Li, H.: Topic set size design with the evaluation measures for short text conversation. In: Zuccon, G., Geva, S., Joho, H., Scholer, F., Sun, A., Zhang, P. (eds.) AIRS 2015. LNCS, vol. 9460, pp. 319–331. Springer, Heidelberg (2015). doi:10.1007/978-3-319-28940-3_25

[3] http://www.thuir.cn/ntcirwww/.

11. Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., Miyao, Y.: Overview of the NTCIR-12 short text conversation task. In: Proceedings of NTCIR-12, pp. 473–484 (2016)
12. Shibuki, H., Sakamoto, K., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T., Kando, N.: Overview of the NTCIR-12 QA Lab-2 task. In: Proceedings of NTCIR-12, pp. 392–408 (2016)
13. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Proceedings of ACM SIGIR 2012, pp. 95–104 (2012)
14. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: Proceedings of ACM SIGIR 2008, pp. 51–58 (2008)
15. Webber, W., Moffat, A., Zobel, J.: Statistical power in retrieval experimentation. In: Proceedings of ACM CIKM 2008, pp. 571–580 (2008)

http://www.springer.com/978-3-319-48050-3

Information Retrieval Technology
12th Asia Information Retrieval Societies Conference,
AIRS 2016, Beijing, China, November 30 – December 2,
2016, Proceedings
Ma, S.; Wen, J.-R.; Liu, Y.; Dou, Z.; Zhang, M.; Chang, Y.;
Zhao, X. (Eds.)
2016, XIII, 372 p. 61 illus., Softcover
ISBN: 978-3-319-48050-3