A New Perspective on Score Standardization

Julián Urbano
Delft University of Technology
The Netherlands
urbano.julian@gmail.com

Harlley Lima
Delft University of Technology
The Netherlands
h.a.delima@tudelft.nl

Alan Hanjalic
Delft University of Technology
The Netherlands
a.hanjalic@tudelft.nl

ABSTRACT

In test collection based evaluation of IR systems, score standardization has been proposed to compare systems across collections and minimize the effect of outlier runs on specific topics. The underlying idea is to account for the difficulty of topics, so that systems are scored relative to it. Webber et al. first proposed standardization through a non-linear transformation with the standard normal distribution, and recently Sakai proposed a simple linear transformation. In this paper, we show that both approaches are actually special cases of a simple standardization which assumes specific distributions for the per-topic scores. From this viewpoint, we argue that a transformation based on the empirical distribution is the most appropriate choice for this kind of standardization. Through a series of experiments on TREC data, we show the benefits of our proposal in terms of score stability and statistical test behavior.

KEYWORDS

Evaluation, test collection, score standardization, statistical testing

ACM Reference Format:
Julián Urbano, Harlley Lima, and Alan Hanjalic. 2019. A New Perspective on Score Standardization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331315

1 INTRODUCTION

In the traditional Cranfield paradigm for test collection based evaluation in Information Retrieval (IR), systems are evaluated and compared by assessing their effectiveness on the set of topics contained in a collection. Specifically, an effectiveness measure like Average Precision is used to score every system with every topic, and the per-system mean scores over topics are often used as the single indicator of performance to rank systems [4]. It is well known in the IR literature that this paradigm does not allow to compare the performance of systems tested on different collections. The main reason for this is the very large variability we find in topic difficulty [1, 6, 8]. A system with a good score on one collection may very well achieve a low score on another. Even when comparing systems using the same collection, not all topics contribute equally to the final score because of their differences in difficulty (see for instance the bottom-left plot in Figure 1). Therefore, the observed differences in mean scores may be disproportionately due to a few topics in the collection [2, 9].

To mitigate this problem, Webber et al. [9] proposed a two-step standardization process to look at scores relative to the difficulty of the topic. First, given a raw effectiveness score x of some system on some topic, a traditional z-score is computed

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ and σ are the mean and standard deviation of the system scores for the topic. The effect is twofold: whether the topic is easy or hard (high or low μ), the distribution of z-scores is centered at zero; and whether systems perform similarly for the topic or not (low or high σ), the z-scores have unit variance. Thanks to this first step, all topics contribute equally to the final scores.

The second step is a transformation of the z-score so that the final standardized score y is bounded between 0 and 1, as is customary in IR measures. Webber et al. [9] propose to use the cumulative distribution function (*cdf*) of the standard normal distribution, which naturally maps z-scores on to the unit interval:

$$y = \Phi(z). \tag{2}$$

Recently, Sakai [3] proposed a simple linear transformation of the z-score instead of the non-linear transformation applied by Φ:

$$y = Az + B. \tag{3}$$

On the grounds of Chebyshev's inequality, they further suggested $A = 0.15$ and $B = 0.5$ so that at least 89% of the scores will fall within $[0.05, 0.95]$. Furthermore, and to ensure that standardized scores always stay within the unit interval, they proposed to simply censor y between 0 and 1, computing $y = \max(\min(1, Az + B), 0)$ in reality.

In this paper we show that the standardizations by Webber et al. [9] and Sakai [3] are actually special cases of a general class of standardizations consisting in assuming a specific distribution for the per-topic scores, and that they differ only in what distribution they assume. From this new perspective, we argue that the empirical distribution is actually the most appropriate choice because of its properties. We also carry out two experiments on TREC data that show how our proposal behaves better than both raw scores and the previous standardization schemes.

2 SCORE STANDARDIZATION

Let F be the distribution of scores by some population of systems on some particular topic and according to some specific measure like *AP*. If we knew this distribution, we could standardize a raw score x simply by computing $y = F(x) = P(X \leq x)$, which naturally bounds y between 0 and 1. The reasoning is that the *cdf* actually tells us where x is with respect to the rest of scores that one may expect for the topic, precisely computing the fraction of systems with lower

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331315

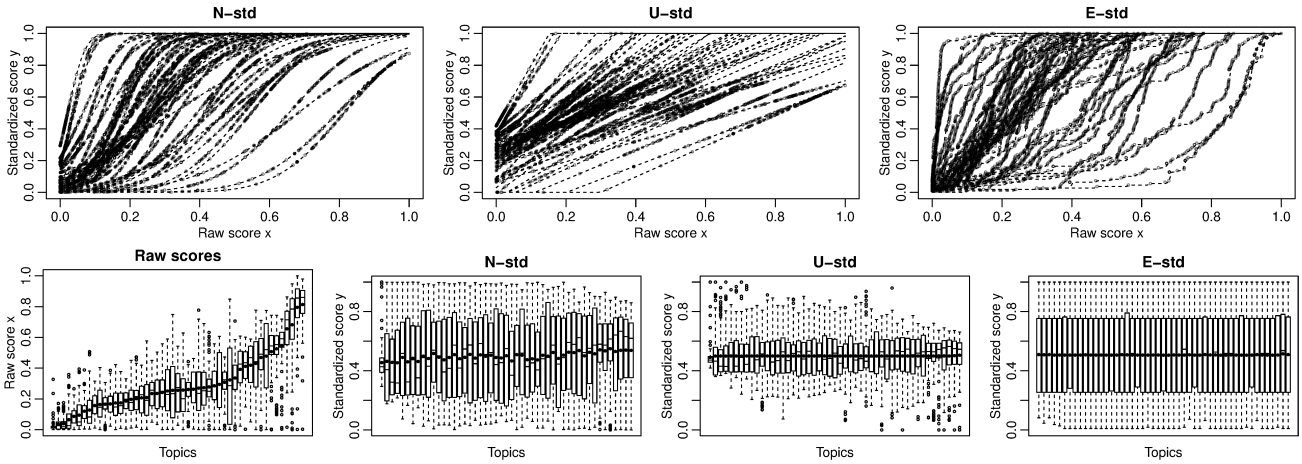

Figure 1: Effect of standardization on the AP scores of the TREC 2004 Robust systems and topics. Top: mapping from raw to standardized scores; points mark actual observations and lines mark the full mapping. Bottom: Distributions of per-topic scores without and with standardization; topics sorted by mean raw score; solid points mark the per-topic means.

scores. The maximum score for the topic, however low it might be for a difficult one, would turn into $y = 1$, and the minimum will similarly turn into $y = 0$, *regardless* of the distribution F.

The application of a distribution function is thus a simple and natural way to standardize. Ideally, one would use the true distribution of scores for the topic. The problem is, of course, that we do not know that distribution, but only the set of scores $X_1, \ldots, X_n \sim F$ *observed* with a set of n systems, that is, a mere sample. Still, from this sample we could make an estimate and use it to standardize:

$$y = \hat{F}(x). \tag{4}$$

Reconsidering the two-step standardization by Webber et al. [9] where the z-score is passed through the *cdf* of the standard normal, we note that it is equivalent to the one-step standardization in eq. (4) when using a normal distribution with mean μ and variance σ^2:

$$y = \Phi\left(\frac{x-\mu}{\sigma}\right) = F_{\mathcal{N}}\left(x\ ;\ \mu, \sigma^2\right). \tag{5}$$

In a sense, what eq. (5) does is assume $X \sim \mathcal{N}(\mu, \sigma^2)$ for the raw scores and standardize through the *cdf*.

In the case of Sakai [3], the transformation is linear with slope A and intercept B. A distribution whose *cdf* has this form is the Uniform. In particular, the *cdf* of a uniform variate between a and b is $F_{\mathcal{U}}(x\ ;\ a, b) = (x-a)/(b-a)$, that is, a linear function with slope $1/(b-a)$ and intercept $a/(a-b)$. Therefore, if we account for the initial step that computes the z-score, and solve for a and b, we can easily see that the two-step standardization by Sakai [3] is equivalent to the one-step distribution-based standardization in eq. (4) when using a uniform distribution with appropriate boundaries:

$$y = A\frac{x-\mu}{\sigma} + B = F_{\mathcal{U}}\left(x\ ;\ \mu - \sigma\frac{B}{A}, \mu + \sigma\frac{1-B}{A}\right). \tag{6}$$

Similarly, what eq. (6) does is assume $X \sim \mathcal{U}\left(\mu - \sigma\frac{B}{A}, \mu + \sigma\frac{1-B}{A}\right)$ for the raw scores and standardize through the *cdf*.

However, the assumption of a normal or a uniform is a very strong one. In the absence of knowledge about F, which is virtually always the case, the empirical distribution function provides an estimate without the need for assumptions like in eq. (5–6):

$$y = ecdf\left(x\ ;\ x_1, \ldots, x_n\right) = \frac{1}{n}\sum I\left(x_i \leq x\right), \tag{7}$$

where I is the indicator function. In simple terms, *ecdf* computes the rank within the sample, normalized between 0 and 1. As noted by Webber et al. [9], this is precisely what we are after, because *"knowledge of a system's score is less useful than knowledge of its rank among the set of systems"*.

In summary, we see that the standardizations by Webber et al. [9] and Sakai [3] actually reduce to assuming a normal or uniform distribution on the per-topic scores, whose *cdf* is used to compute the fraction of scores that lie below. As an alternative, we propose to use the empirical distribution, which is free from such assumptions. Let us refer to these as *N-std*, *U-std* and *E-std*, respectively.

The top plots in Figure 1 illustrate the effect of standardization. In the case of *N-std* we can appreciate the typical *S*-shape of the normal *cdf*, which effectively eliminates the large effect of outliers. In the case of *U-std* we clearly see the linear transformation which, on the other hand, maintains the proportionality of scores but requires censoring. *E-std* naturally handles outliers because standardized scores increase in steps of $1/n$ units. A characteristic of *E-std* that is clear from the plots is that the standardized scores cover the full $[0, 1]$ range by design, while neither *N-std* nor *U-std* do so because of the model they impose on the data. It is also evident that *U-std* will not be sensitive to differences between systems with very low or very high scores, because they are all censored to be 0 or 1.

Additional advantages of *E-std* can be seen in the bottom plots in Figure 1, which show per-topic distributions. We first observe that, even though z-scores have zero mean and unit variance, the non-linear transformation of *N-std* no longer maintains common mean and variance ($\overline{y}$ ranges between 0.45 and 0.55, and s_y ranges between 0.14 and 0.31). Because *U-std* applies a linear transformation, we see that the mean scores remain $\overline{y} = B = 0.5$ and variability is much more stable. In fact, from eq. (3) we can see that $s_y = A = 0.15$,

with slight deviations if censoring is needed. In the case of *E-std*, we see nearly constant mean and variance. This is achieved by design, because, in general, if $X \sim F$, then $Y = F(X)$ follows a standard uniform. Therefore, *E-std* produces standardized scores that are uniformly distributed, ensuring $\overline{y} = 0.5$ and $s_y = \sqrt{1/12}$.

The point still remains that μ and σ are also unknown. The way around this limitation is to estimate them from a previous set of systems (called *standardization systems* by [3, 9]). Thus, given the scores $X_1, \ldots, X_n$ of these systems, the estimates are the per-topic sample mean $\hat{\mu} = \overline{X}$ and standard deviation $\hat{\sigma} = s_X$. For *E-std*, these are precisely the data used in eq. (7) to standardize. In principle, these standardization systems should represent the system population of interest, which ultimately determines the topic difficulty through the per-topic distributions. In our view, the most reasonable choice would be the state of the art systems, which in a TREC collection are arguably the set of systems participating in the track.

3 EXPERIMENTS

This section reports on two experiments to assess the effect of standardization. In the first one, we consider system comparisons using the same test collection (within-collection), while in the second one we consider comparisons between systems evaluated on different collections (between-collection). Comparisons will be made between results produced by the raw scores, *N-std*, *U-std* and *E-std*. For completeness, we also evaluate the standardization scheme that simply computes the z-score as in eq. (1) and therefore produces unbounded y scores. This scheme is called *z-std*.

The data used in our experiments are the TREC 2004 Robust (RB) and TREC 2006 Terabyte (TB) collections. The RB data contains 110 systems evaluated on the 99 TREC 2003–2004 Robust topics. The TB data contains 61 systems on the 149 TREC 2004–2006 Terabyte topics. In terms of measures, we use *AP* and *nDCG*.

3.1 Within-Collection Comparisons

In order to investigate the effect of standardization on within-collection comparisons, we proceed as follows. We randomly sample 50 topics from the full set and compute the raw scores and the standardized scores as per each of the standardization schemes. From these data we compute three statistics. First, we compute the correlation between the ranking of systems by raw scores and the ranking by standardized scores, using Kendall's τ and Yilmaz's τ_{ap} [10][1]. A high correlation indicates that the standardized scores are not much different from the raw scores, so in principle we look for lower coefficients. The third indicator evaluates the statistical power of the evaluation. In particular, we run a 2-tailed paired t-test between every pair of systems and, under the assumption that the null hypothesis is indeed false, look for schemes that maximize power. The process is repeated 10,000 times with both the RB and TB datasets, on both *AP* and *nDCG*.

Figure 2 shows the results for a selection of collection-measure combinations[2]. The two plots in the first row show the distributions of τ correlations. As expected, *U-std* and *z-std* perform very similarly because the former is simply a linear transformation of the latter; differences come from the necessity to censor outliers in *U-std*. Indeed, because they are both a linear transformation of the raw scores, they produce the most similar rankings. *N-std* results in slightly lower correlations, but *E-std* sets itself clearly apart from the others, yielding significantly lower τ scores. The plots in the second row show even clearer differences in terms of τ_{ap}. We see that *U-std* and *z-std* are almost identical, but more importantly we see that *N-std* and *E-std* are even further away, likely because they eliminate outliers that could affect the top ranked systems.

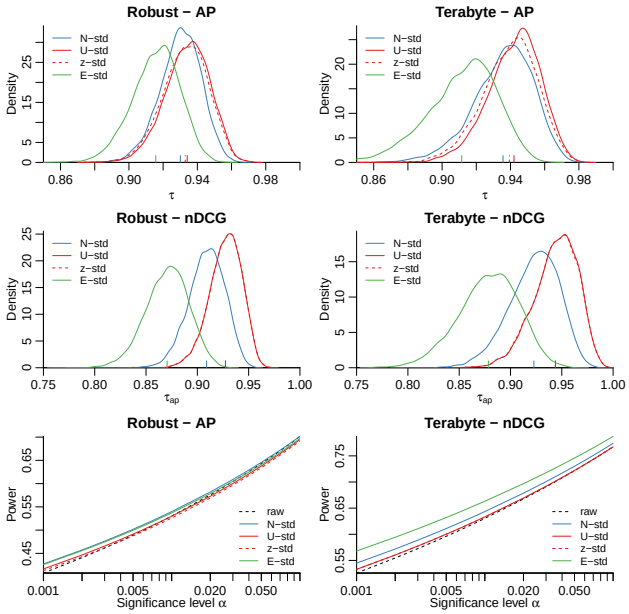

Figure 2: Within-collection comparisons. First row: τ correlation between rankings of systems with raw and standardized scores (lower is better); rugs mark the means. Second row: τ_{ap} correlation (lower is better). Third row: power of paired t-tests at various α levels (higher is better).

The two plots in the last row show statistical power for a range of significance levels. We can first observe that all standardization schemes achieve higher power than the raw scores, showing a clear advantage of the standardization principle. Once again, *U-std* and *z-std* perform nearly identically, and both are outperformed by *N-std* and, specially, *E-std*.

3.2 Between-Collection Comparisons

Here we study how standardization affects between-collection comparisons. In this case, we randomly sample two disjoint subsets of 50 topics each and compute raw and standardized scores on both subsets. Because topics are sampled from the full set, both results can be regarded as coming from two different collections having different topics from the same population. In this case we are not interested in how standardized scores compare to raw scores, but rather on how stable the results are between both sets of topics, so we compute the following four statistics. First, we compute the τ and τ_{ap} correlations between both rankings. We seek high correlations, indicating high score stability across topic sets. Third, for

[1] In particular, we compute τ_b and $\tau_{ap,b}$ to deal with tied systems. See [5] for details.
[2] All plots, along with data and code to reproduce results, are available from `https://github.com/julian-urbano/sigir2019-standardization`.

every system we run a 2-tailed unpaired t-test between both sets. By definition, the null hypothesis is true because we are comparing a system to itself simply on a different sample, so we expect as many Type I errors as the significance level α. Finally, we run another test between every system on one collection and every other system on the other collection, looking again to maximize statistical power under the assumption that all systems are different and thus null hypotheses are false. As before, this process is repeated 10,000 times with both the RB and TB datasets, on both *AP* and *nDCG*.

Figure 3 shows the results for a selection of collection-measure combinations. The plots in the first two rows show that standardization generally produces more stable results, as evidenced by raw scores yielding lower correlations. *U-std* and *z-std* perform very similarly once again, and *E-std* generally outperforms the others, producing slightly more stable comparisons between collections. An exception can be noticed for τ_{ap} on the TB dataset, which requires further investigation.

The third row of plots show the Type I error rates. We can see that all scoring schemes behave just as expected by the significance level α. This evidences on the one hand the robustness of the t-test [7] (recall the diversity of distributions from the boxplots in Figure 1), and on the other hand that standardization neither harms nor helps from the point of view of Type I errors (this is rather a characteristic of the test). Finally, the last two plots show the power achieved by the tests when comparing different systems. Here we first notice that all standardization schemes are substantially more powerful than the raw scores, achieving about twice as much power. While the results are very similar in the RB set, we see clear differences in the TB set, with *E-std* once again outperforming the other schemes.

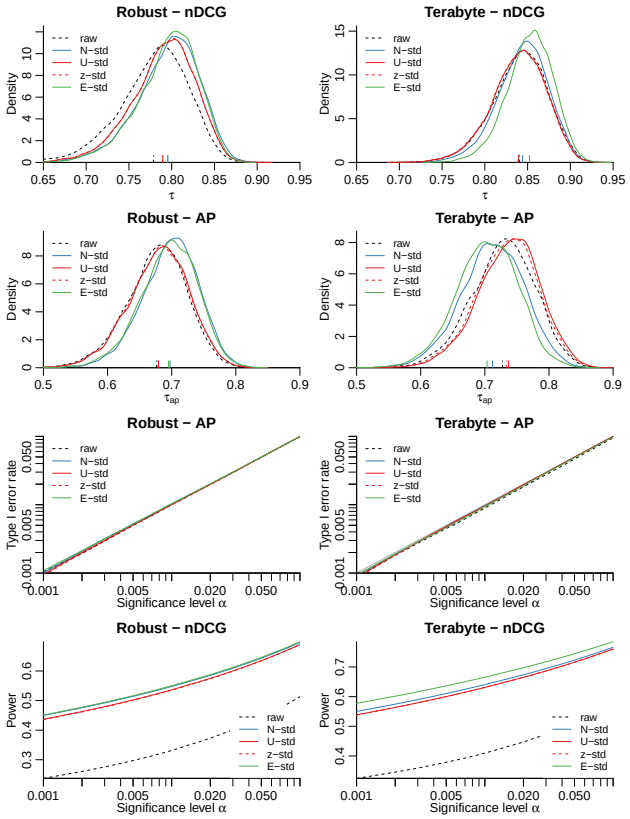

Figure 3: Between-collection comparisons. First row: τ correlation between the rankings of systems produced by the two collections (higher is better); rugs mark the means. Second row: τ_{ap} correlation (higher is better). Third row: Type I error rate of unpaired t-tests at various α levels (diagonal is better). Fourth row: statistical power (higher is better).

4 CONCLUSIONS

In this paper we revisit the problem of score standardization to make IR evaluation robust to variations in topic difficulty. We introduced a new scheme for standardization based on the distributions of per-topic scores, and showed that previous methods by Webber et al. [9] and Sakai [3] are special cases of this scheme. From this point of view we propose the empirical distribution as an alternative, and discuss a number of points that highlight its superiority.

In experiments with TREC data, we showed that, even though the raw and standardized rankings are the same topic by topic, the rankings by mean scores may differ considerably. In addition, standardization achieves higher statistical power. Thus, standardization offers an alternative and quite different view on system comparisons. However, it is important to note that these comparisons are made on a different scale altogether, so one may not just use standardized scores to make statements about raw scores. Nonetheless, standardization with the empirical distribution is arguably more faithful to our notion of relative system effectiveness.

Future work will follow three main lines. First, we will study additional datasets and measures for generality. However, because TREC collections are usually limited to 50 topics, we also plan on using recent simulation methods so that we can analyze more data [7]. Finally, we will study the stability of *E-std* for varying numbers of systems. This is interesting because, even though the empirical function converges to the true distribution, it is unclear how large the set of systems needs to be for the results to be stable.

ACKNOWLEDGMENTS

Work carried out on the Dutch national e-infrastructure (SURF Cooperative) and funded by European Union's H2020 programme (770376-2 TROMPA). `Eva.say("Hello World!");`

REFERENCES

[1] D. Bodoff. 2008. Test Theory for Evaluating Reliability of IR Test Collections. *Information Processing and Management* 44, 3 (2008), 1117–1145.
[2] J. Guiver, S. Mizzaro, and S. Robertson. 2009. A Few Good Topics: Experiments in Topic Set Reduction for Retrieval Evaluation. *ACM TOIS* 27, 4 (2009), 1–26.
[3] T. Sakai. 2016. A Simple and Effective Approach to Score Standardization. In *ACM ICTIR*. 95–104.
[4] M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247–375.
[5] J. Urbano and M. Marrero. 2017. The Treatment of Ties in AP Correlation. In *SIGIR ICTIR*. 321–324.
[6] J. Urbano, M. Marrero, and D. Martín. 2013. On the Measurement of Test Collection Reliability. In *ACM SIGIR*. 393–402.
[7] J. Urbano and T. Nagler. 2018. Stochastic Simulation of Test Collections: Evaluation Scores. In *ACM SIGIR*.
[8] E. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track. In *TREC*.
[9] W. Webber, A. Moffat, and J. Zobel. 2008. Score Standardization for Inter-collection Comparison of Retrieval Systems. In *AMC SIGIR*. 51–58.
[10] E. Yilmaz, J.A. Aslam, and S. Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *AMC SIGIR*. 587–594.