Scales of Evaluation Measures: From Theory to Experimentation*

Marco Ferrante
ferrante@math.unipd.it
Dept. of Mathematics, University of Padua
Padua, Italy

Nicola Ferro
ferro@dei.unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Eleonora Losiouk
elosiouk@math.unipd.it
Dept. of Mathematics, University of Padua
Padua, Italy

Silvia Pontarollo
spontaro@math.unipd.it
Dept. of Mathematics, University of Padua
Padua, Italy

1 INTRODUCTION

Evaluation measures are the basis for quantifying the performance of IR systems and *measurement scales* play a central role since they determine the operations that can be performed with the measured values and, as a consequence, the statistical analyses that can be applied. Stevens [4] identifies four major types of scales with increasing properties: (i) the *nominal scale* consists of discrete unordered values, i.e. categories; (ii) the *ordinal scale* introduces a natural order among the values; (iii) the *interval scale* preserves the equality of intervals or differences; and (iv) the *ratio scale* preserves the equality of ratios. For example, mean and variance should be computed only when relying on interval scales.

We present our formal theory of IR evaluation measures [2], based on the *representational theory of measurement* [3, 4], to determine whether and when IR measures are interval scales.

We found that common set- based retrieval measures – namely Precision, Recall, and F-measure – always are interval scales in the case of binary relevance while this does not happen in the multi-graded relevance case. In the case of rank-based retrieval measures – namely AP, gRBP, DCG, and ERR – only gRBP is an interval scale when we choose a specific value of the parameter p and define a specific total order among systems while all the other IR measures are not interval scales. We also introduce some brand new set-based and rank-based IR evaluation measures which ensure to be interval scales.

Finally, we discuss the outcomes of an extensive evaluation [1], based on standard TREC collections, to study how our theoretical findings impact on the experimental ones. In particular, we report here a correlation analysis to study the relationship among the above-mentioned state-of-the-art evaluation measures and their scales.

2 SET-BASED MEASURES

Let us start by introducing an order relation $\preceq$ on the set of judged runs R(N). Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$, and let k be the biggest relevance degree at which the two runs differ for the first time, i.e. $k = \max\{j \leq c : \left|\{i : \hat{r}_i = a_j\}\right| \neq \left|\{i : \hat{s}_i = a_j\}\right|\}$. We strictly order any pair of distinct system runs as follows

$$\hat{r} \prec \hat{s} \iff \left|\{i : \hat{r}_i = a_k\}\right| < \left|\{i : \hat{s}_i = a_k\}\right| . \tag{1}$$

$R(N)$ is a totally ordered set with respect to the ordering $\preceq$ defined by (1). As for any totally order set, $R(N)$ is a poset consisting of only one maximal chain (the whole set); therefore it is *graded of rank* $|R(N)| - 1$, where $\left|R(N)\right| = \binom{N+c}{N}$ since it consists of all the N combinations of $c + 1 = |REL|$ objects with repetition. Since $R(N)$ is graded of rank $|R(N)| - 1$, there exists a unique *rank function* $\rho(\hat{r}) : R(N) \longrightarrow \mathbb{N}$ such that $\rho(\hat{0}) = 0$ and $\rho(\hat{s}) = \rho(\hat{r}) + 1$ if $\hat{s}$ covers $\hat{r}$:

$$\rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_a(\hat{r}_j) + N - j}{N - j + 1} , \tag{2}$$

where $\hat{r} = \{\hat{r}_1, \ldots, \hat{r}_N\} \in R(N)$ with $\hat{r}_i \preceq \hat{r}_{i+1}$ for any $i < N$.

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \preceq \hat{s}$, and we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$. $(R(N), \preceq_d)$ is a difference structure. Thus the rank function is an interval scale and we are able to define a new interval-scale measure that follows:

Definition 2.1. The *Set-Based Total Order (SBTO)* measure on $(R(N), \preceq_d)$ is:

$$\mathrm{SBTO}(\hat{r}) = \rho(\hat{r}) = \sum_{j=1}^{N} \binom{\delta_a(\hat{r}_j) + N - j}{N - j + 1} . \tag{3}$$

3 RANK-BASED MEASURES

Top-heaviness is a central property in *Information Retrieval (IR)*, stating that the higher a system ranks relevant documents the better it is. If we apply this property at each rank position and we take to extremes the importance of having a relevant document ranked higher, we can define a *strong top-heaviness* property which, in turn, will induce a total ordering among runs.

Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$, then there exists $k = \min\{j \leq N : \hat{r}[j] \neq \hat{s}[j]\} < \infty$, and we order system runs as follows

$$\hat{r} \prec \hat{s} \iff \hat{r}[k] < \hat{s}[k] . \tag{4}$$

This order prefers a single relevant document ranked higher to any number of relevant documents, with same relevance degree or higher, ranked just below it

$$(\hat{u}[1], \ldots, \hat{u}[m], a_0, a_c, \ldots, a_c), \prec (\hat{u}[1], \ldots, \hat{u}[m], a_j, a_0, \ldots, a_0) ,$$

*Extended abstract of [1, 2]

IIR 2019, September 16–18, 2019, Padova, Italy

Table 1: Correlation analysis among set-based measures.

Binary Relevance – T08		
Measure Pair	**Topic-by-Topic**	**Overall**
Precision vs SBTO	1.0000	0.9998
Recall vs SBTO	1.0000	0.8591
F-measure vs SBTO	1.0000	0.9670
Precision vs Recall	1.0000	0.8588
SBTO vs RBTO	0.4358	0.7410
Multi-graded Relevance – T26		
Measure Pair	**Topic-by-Topic**	**Overall**
Generalized Precision vs SBTO	0.7325	0.9175
Generalized Recall vs SBTO	0.7325	0.8453
Generalized Precision vs Generalized Recall	1.0000	0.9003
SBTO vs RBTO	0.3895	0.7352

Table 2: Correlation analysis among rank-based measures.

Binary Relevance – T08		
Measure Pair	**Topic-by-Topic**	**Overall**
RBP $p = 1/2$ vs RBTO	1.0000	1.0000
RBP $p = 0.2$ vs RBTO	0.9985	0.9225
RBP $p = 0.8$ vs RBTO	0.8553	0.9043
AP vs RBTO	0.6099	0.7439
Multi-graded Relevance – T26		
Measure Pair	**Topic-by-Topic**	**Overall**
gRBP $p = 1/3$ vs RBTO	1.0000	1.0000
gRBP $p = 1/3$, $W_3 = [0, 1, 3]$ vs RBTO	0.9867	0.9618
gRBP $p = 0.2$ vs RBTO	0.9996	0.9755
gRBP $p = 0.8$ vs RBTO	0.7420	0.9026
DCG vs RBTO	0.3774	0.6984
ERR vs RBTO	0.9468	0.9502
RBTO $W_1 = [0, 1, 2]$ vs RBTO $W_2 = [0, 2, 4]$	1.0000	1.0000
RBTO $W_1 = [0, 1, 2]$ vs RBTO $W_3 = [0, 1, 3]$	0.9866	0.9618

for any $1 \leq j \leq c$, for any length $N \in \mathbb{N}$ and any $m \in \{0, 1, \ldots, N-1\}$. This is why we call it *strong top-heaviness*.

$R(N)$ is totally ordered with respect to $\preceq$ and is *graded of rank* $(c+1)^N - 1$. Therefore, there is a unique rank function $\rho : R(N) \longrightarrow \{0, 1, \ldots, (c+1)^N - 1\}$ which is given by:

$$\rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathcal{a}}(\hat{r}[i])(c+1)^{N-i}, \tag{5}$$

where $\delta_{\mathcal{a}}$ is the indicator function.

Let us set $\boldsymbol{\delta_{\mathcal{a}}}(\hat{r}) = (\delta_{\mathcal{a}}(\hat{r}[1]), \ldots, \delta_{\mathcal{a}}(\hat{r}[N]))$. If we look at $\boldsymbol{\delta_{\mathcal{a}}}(\hat{r})$ as a string, the rank function is exactly the conversion in base 10 of the number in base $c+1$ identified by $\boldsymbol{\delta_{\mathcal{a}}}(\hat{r})$ and the ordering among runs $\preceq$ corresponds to the ordering $\leq$ among numbers in base $c+1$.

The *natural distance* is then given by $\ell(\hat{r}, \hat{s}) = \rho(\hat{s}) - \rho(\hat{r})$, for $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \preceq \hat{s}$, and we can define the difference as $\Delta_{\hat{r}\hat{s}} = \ell(\hat{r}, \hat{s})$ if $\hat{r} \preceq \hat{s}$, otherwise $\Delta_{\hat{r}\hat{s}} = -\ell(\hat{s}, \hat{r})$. $(R(N), \preceq_d)$ is a difference structure. As done before in the set-based case, an interval scale measure on $(R(N), \preceq_d)$ is given by the rank function itself.

Definition 3.1. The *Rank-Based Total Order (RBTO)* interval-scale measure on $(R(N), \preceq_d)$ is:

$$\mathrm{RBTO}(\hat{r}) = \rho(\hat{r}) = \sum_{i=1}^{N} \delta_{\mathcal{a}}(\hat{r}[i])(c+1)^{N-i} \tag{6}$$

4 EXPERIMENTS

We explore the following research question: "*How to scales determine the relationship among evaluation measures?*", i.e. what is the relationship between measures which are interval scales, ordinal scale or which are not on any scale? To this end, we will perform Kendall's τ correlation analysis on TREC 08 Ad-hoc (binary judgements) and TREC 26 Core (multi-graded judgments) collections.

Tables 1 and 2 report the correlation analysis in the case of set-based and rank-based evaluation measures for both binary and multi-graded relevance.

Note that two interval scale measures order systems in the same way on the same topic and their correlation must be 1.0. However, this may be no more true, if you first average performance across all the topics and then compute the correlation, which is the typical way of computing Kendall's τ correlation [5].

This can be, for example, observed in Table 1 where Precision, Recall, F-measure, and SBTO are all transformation of the same interval scale and thus their topic-by-topic correlation is 1; on the other hand, their overall correlation, i.e. the traditional one, is different from 1.0 because of the effect of the recall base when averaging across topics. This suggest that the difference between Precision and Recall ($\tau = 0.85$) is not due to them ranking systems differently but just to the fact that the recall base alters the scale properties from topic to topic.

Another interesting case is RBP. For $p = 1/2$ ($p = 1/3$ in the multigraded case) it is an interval-scale; for $p < 1/2$ it is an ordinal but not interval scale and its correlation starts departing from 1.0; the effect is much more pronounced for RBP with $p > 1/2$ which is neither an ordinal nor an interval scale, suggesting that simply acting on a parameter of a measure can completely alter its scale properties.

A final interesting case is RBTO with different weights for the relevance degrees: $W_1 = [0, 1, 2]$ vs $W_2 = [0, 2, 4]$ keep the RBTO on an interval-scale while $W_1 = [0, 1, 2]$ vs $W_3 = [0, 1, 3]$ show that it stops to be an interval-scale. Indeed, our theoretical findings [2] demonstrate that, in the multi-graded case, the interval-scale property is complied with only if the weights of the relevance degrees are on a ratio scale, which is not the case for $W_3 = [0, 1, 3]$.

REFERENCES

[1] M. Ferrante, N. Ferro, and E. Losiouk. 2019. How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal* (2019).

[2] M. Ferrante, N. Ferro, and S. Pontarollo. 2019. A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31, 3 (March 2019), 409–422.

[3] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement. Additive and Polynomial Representations.* Vol. 1. Academic Press, USA.

[4] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science, New Series* 103, 2684 (June 1946), 677–680.

[5] E. M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR 1998*, 315–323.