Extending rank correlation coefficients for relevance profiles

Author: Andrea Vezzuto[1] **Supervisor & Responsible Professor**: Julián Urbano [1]

[1]EEMCS, Delft University of Technology, The Netherlands

Introduction

- Modern search and recommendation systems rely on ranked lists of items.
- Relevance profiles are lists of multi-level (graded) relevance judgments, allowing for a similarity evaluation based on the utility of items.
- We propose new correlation metrics to measure similarity using the relevance information contained in profiles.

Background

- A generalised correlation coefficient on lists x and y is:

$$\tau_w(x, y) = \frac{\langle x, y\rangle_w}{\|x\|_w \cdot \|y\|_w}. \tag{1}$$

The numerator, $\langle x, y\rangle_w$, denotes a weighted concordance score given by:

$$\langle x, y\rangle_w = \sum_{i<j} c(i, j) \cdot w = \sum_{i<j} \mathrm{sign}(x_j - x_i) \cdot \mathrm{sign}(y_j - y_i) \cdot w, \tag{2}$$

while $\|x\|_w = \sqrt{\langle x, x\rangle_w}$ is the norm of ranking x, and x_j denotes the ranking of element j in list x.

- Kendall's τ measures rank agreement via equally-weighted pairwise comparisons. Therefore, $w = 1$.
- τ_{ap} and τ_h are top-weighted extensions focusing on items higher in the ranking. Their weighing factors are:
  - τ_{ap}: $w(y_i, y_j) = \frac{1}{\max(y_i, y_j) - 1}$
  - τ_h: $w(\rho_{x,y}(i), \rho_{x,y}(j)) = \frac{1}{1+\rho_{x,y}(i)} + \frac{1}{1+\rho_{x,y}(j)}$, where $\rho_{x,y}$ is defined by ordering elements lexicographically with respect to x.
- Recent work in the field of relevance judgment generation using LLMs compares profiles to human-made references by:
  - Computing a normalised discounted cumulative gain function for each system, creating an overall ranking using the cumulative score for each.
  - Comparing the similarity of system rankings using Kendall's τ.

Axiomatic Properties of the Extended Coefficients

- Axiom 1: If the compared rankings are equivalent (all items are placed in the same order), the coefficient is 1.
- Axiom 2: If the compared rankings are reversed (one ranking lists items in the opposite order of the other), the coefficient is -1.

$$\begin{aligned} z &= \langle A, B\rangle \\ z' &= \langle B, A\rangle \\ z_{rel} &= z'_{rel} = \langle 2, 2\rangle \end{aligned}$$

In terms of utility, z_{rel} and z'_{rel} are equivalent. This motivates the following two axioms.

- Axiom 3: If the ordering of relevance values in the compared rankings is equivalent, the coefficient is 1.
- Axiom 4: If the ordering of relevance in the compared rankings is reversed, and all values are unique in each ranking, the coefficient is -1.
- Axiom 5: If the compared rankings are independent, the expected value of the coefficient is 0.
- *Note:* no coefficient can simultaneously satisfy axioms 2 and 3 for any ranking. Therefore:
  - If the identity of elements determines concordance, with relevance being used as a weighing factor, Axioms 1, 2, and 5 can hold.
  - If the relevance of items determines concordance, Axioms 3, 4, and 5 can be satisfied.

Redefining Concordance

1. Relevance as a weighing factor

The coefficient maintains item identity to determine the concordance of a pair. As such, it satisfies axioms 1, 2, and 5.

Distance-weighted

$$c_{dw}(i, j) = sign(x_j - x_i) \cdot sign(y_j - y_i) \cdot \frac{|rel_i - rel_j|}{max(rel_i, rel_j)}, \tag{3}$$

where rel_i denotes the relevance of element i, on a relevance scale in $[0, max\_rel]$.

General intuition: the relative ordering of items with significantly different relevance values conveys more meaningful information and should therefore have a greater impact on the coefficient.

2. Relevance to determine concordance

This coefficient uses relevance values to determine the concordance of a pair. As such, it satisfies axioms 3, 4, and 5.

Augmented additive

$$c_{ac}(i, j) = \begin{cases} 0 & \text{if } rel_{x_i} = rel_{y_i} = rel_{y_{N-i}} \\ & \quad \text{and } rel_{x_j} = rel_{y_j} = rel_{y_{N-j}} \\ & \quad \text{and } (rel_{x_{N-i}} \neq rel_{y_{N-i}} \text{ or } rel_{x_{N-j}} \neq rel_{y_{N-j}}) \\ c_{sc}(i, j, N-i, N-j) \cdot r(i, j, N-i, N-j) & \text{if } (rel_{x_i} \neq rel_{y_i} \text{ or } rel_{x_j} \neq rel_{y_j}) \\ & \quad \text{and } ((rel_{x_i} = rel_{y_{N-i}} \text{ and } rel_{x_j} = rel_{y_{N-j}}) \\ & \quad \text{or } (rel_{y_i} = rel_{x_{N-i}} \text{ and } rel_{y_j} = rel_{x_{N-j}})) \\ c_{sc}(i, j, i, j) \cdot r(i, j, i, j) & \text{otherwise} \end{cases} \tag{4}$$

where r_{ij} measures the strength of relevance-based concordance at the two indices:

$$r(i, j, k, l) = 1 - \frac{\left|\left(rel_{x_i} + rel_{x_j}\right) - (rel_{y_k} + rel_{y_l})\right|}{rel_{x_i} + rel_{x_j} + rel_{y_k} + rel_{y_l}}, \tag{5}$$

and rel_{x_i} is the relevance of the item at index i in list x. Furthermore, $c_{sc}(i, j, k, l)$ is a sign-based concordance measure:

$$c_{sc}(i, j, k, l) = \begin{cases} 1 & \text{if } sign(rel_{x_j} - rel_{x_i}) = sign(rel_{y_k} - rel_{y_l}) \\ a & \text{if } sign(rel_{x_j} - rel_{x_i}) = 0 \\ a & \text{if } sign(rel_{y_k} - rel_{y_l}) = 0 \\ -1 & \text{otherwise.} \end{cases} \tag{6}$$

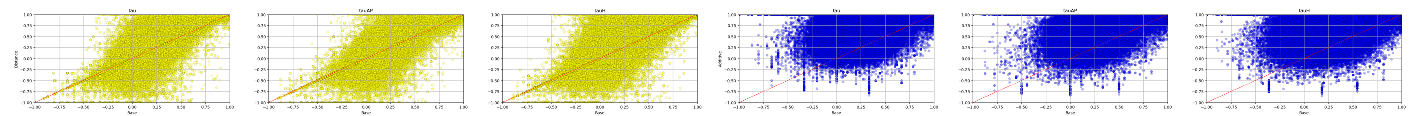
Note that a is empirically set to -0.7 for $n = 4$ and -0.58 for $n = 5$. These values are chosen to ensure that the average correlation value across all τ variants for the simulated dataset is approximately equal to that of $c_{sc}(i, j)$ with $a = -\frac{1}{2\cdot(n-1)}$, which is unbiased by construction.

General intuition: the worst possible ordering is the reversal of two elements, while greater distances are assigned lower magnitudes of the coefficient. In terms of similarity, the former is a complete disagreement, while the latter suggests weaker comparability.

Experimental Setup

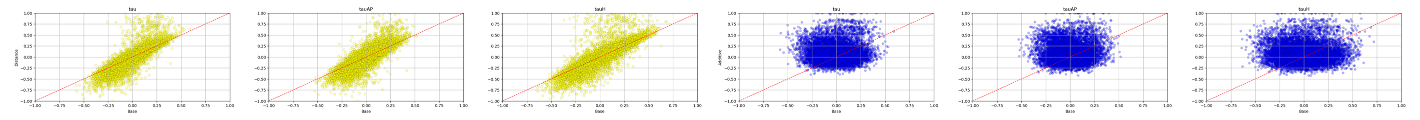
- Real-world data using the ad hoc task from the 2010 - 2014 TREC Web Track. In total, this consists of 150 systems containing the top 1000 items for 50 topics.
- For further testing, simulated profiles using the NSGA-II evolutionary algorithm can be used to generate numerous different systems via tunable parameters:
  - Each profile: list of 50 documents with a 4-point relevance scale (0–3).
  - Fitness: match target anDCG scores in $[0.1, 0.9]$.
  - Techniques: crossover adds or multiplies elements from two parent arrays, and mutation randomly swaps two elements or adds a random value to an item.

Results



(a) dw variant (Equation 3) using item-based concordance with relevance as a weighing factor, on 2010 - 2014 TREC data.

(b) ac variant (Equation 4) using relevance-based concordance, on 2010 - 2014 TREC data.



(c) dw variant (Equation 3) using item-based concordance with relevance as a weighing factor, on simulated data.

(d) ac variant (Equation 4) using relevance-based concordance, on simulated data.

Figure 1. Comparison of τ, τ_{ap}, and τ_h (left to right in each figure) to the proposed coefficients for the 2010 - 2014 TREC (**top row**) and simulated (**bottom row**) data.

Discussion

- Relevance-based concordance shows greater deviation than item-based dw, indicating a significant impact from redefining concordance.
- ac never reaches -1 due to limited relevance scores and overlapping document scores, preventing full discordance.
- dw yields many ± 1 outcomes, as ties in relevance are excluded—such ties don't affect ordering.
- Simulation data shows similar trends but with a narrower range due to independently generated systems, given that the expected coefficient is 0).
- In simulations, ac skews more positive since the relevance distribution is fixed, increasing the likelihood of concordant pairs.

Conclusion and Future Work

- By introducing new definitions of concordance to capture positional agreement and item utility, relevance-aware metrics demonstrate a significant divergence from traditional measures. This highlights the importance of relevance-sensitive evaluation in information retrieval.
- It is important to note that the correlation coefficients proposed in this work may not be suitable for all use cases. Rather, it is hoped that the measures can serve as a framework for other relevance-based extensions of correlation measures.
- Future work may include:
  - A mathematically derived value of a to ensure c_{ac} is unbiased for any n.
  - A stability analysis on the metrics introduced, as in the work by Buckley and Voorhees.
  - Integration of the proposed coefficients into evaluation IR libraries, such as pyircor.

Related Literature

1. Chris Buckley and Ellen M. Voorhees. 2017. Evaluating Evaluation Measure Stability. SIGIR Forum 51, 2 (2017), 235–242.
2. Maurice G. Kendall. 1938. A new measure of rank correlation. Biometrika 30, 1/2 (1938), 81–93.
3. Kevin Roitero, Andrea Brunello, Julián Urbano, and Stefano Mizzaro. 2019. Towards stochastic simulations of relevance profiles. International Conference on Information and Knowledge Management, Proceedings (2019).
4. Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. WWW 2015 - Proceedings of the 24th International Conference on World Wide Web (2015), 1166–1176.
5. Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings (2008), 587–594.