Extending rank correlation coefficients to relevance profiles

Andrea Vezzuto[1]

Supervisor: Julián Urbano[1]

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

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

June 23, 2025

Name of the student: Andrea Vezzuto
Final project course: CSE3000 Research Project
Thesis committee: Julián Urbano, Lilika Markatou

Extending Rank Correlation Coefficients to Relevance Profiles

Andrea Vezzuto

Abstract

Frequently used in modern applications, rankings provide users with a list of the most relevant items. In information retrieval research, the τ, τ_{ap}, and τ_h correlation coefficients are commonly applied to assess the similarity of the underlying systems by comparing the rankings they produce. Traditionally, these comparisons focus solely on item ranking, but introducing relevance values has enabled systems to be analysed based on how element utility relates to retrieval order. In this work, τ, τ_{ap}, and τ_h are extended to incorporate relevance values, presenting several coefficients rooted in an axiomatic approach. These measures compare the utility of items or indices, enabling a granular relevance-based ranking comparison. Overall, the results demonstrate that including relevance judgments leads to significant deviations from traditional rank correlation metrics, highlighting the impact of relevance-aware measures in evaluating system performance and similarity.

CCS Concepts: • **Information systems** → **Similarity measures**; *Relevance assessment*.

Keywords: Rankings, Ranking profiles, Relevance judgments

1 Introduction

Ranked lists are essential in many modern applications to present users with the most relevant information, from search engines to recommender systems [9]. A ranking refers to an ordered list of items, sorted by a notion of perceived relevance or preference. Therefore, evaluating such systems depends on the placement of items with respect to a reference ordering. Traditionally, metrics such as Kendall's τ [8] have done so by quantifying the number of pairwise agreements between two lists. However, comparisons have evolved to include relevance judgments: a more nuanced, multi-level representation of item utility. These are often integer values that indicate the pertinence of an element to a topic, for example assigning a 4 to item A (highly relevant), and a 0 to item B (not relevant) on a scale between 0 and 4.

To this end, ranking profiles are defined as ordered lists of items arranged according to a specific ranking criterion. Additionally, each element possesses an inherent relevance judgment that reflects its utility to a user. Therefore, while the ranking criterion determines the ordering imposed by the ranker, the relevance judgments represent a system-independent, ground-truth attribute of the items. To compare two systems using relevance values, there exist several challenges: traditional Kendall-style measures ignore the interaction of relevance between the lists, while vector similarity coefficients, such as Pearson's correlation [12] and the cosine similarity, do not capture the relative ordering of items, treating each element independently. On the other hand, most information retrieval (IR) metrics, including the discounted cumulative gain (DCG) and the normalised DCG (nDCG) [7], are measured on singular rankings, offering no mechanism to assess the underlying ordering relationship between two profiles. Recent studies have partially bridged this gap by applying rank-based measures like Kendall's τ on transformed metrics such as nDCG [5, 13, 15]. However, this approach measures aggregated similarity over the entire ranking, rather than capturing per-item differences.

Furthermore, rank-based weighted variants such as τ_{ap} [19] and τ_h [17] have been proposed to give higher importance to items at the top of a ranking. This reflects the human behaviour of examining retrieved items top-down, only proceeding to the next document if what has been observed so far is considered insufficient [10]. Similarly to τ, they rely solely on discrete rank positions and do not consider the relevance of items.

This paper addresses the methodological gap by introducing a framework for extending Kendall's τ and its top-weighted variants to incorporate relevance judgments. To do so, several redefinitions of the concordance function, $c(i, j)$, are proposed, guided by a set of axioms established in this work. As such, correlation coefficients calculated on graded lists are introduced, enabling a relevance-based similarity assessment on the relative ordering between rankings.

The remainder of this paper is laid out as follows: Section 2 reviews existing measures, laying the groundwork for their adaptation to relevance profiles proposed in Section 3. Thereafter, Section 4 presents the empirical testing setup and highlights crucial results, with Section 5 discussing the implications of key findings. Finally, Section 6 summarises the work presented in this paper, highlighting some paths for future research, and Section 7 describes how the research process has been ensured to be reproducible.

2 Background

To assess the similarity between two rankings, x and y, Kendall's τ correlation coefficient [8] can be used, as defined by Vigna [17]:

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

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

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

where w is a weighting function, applied uniformly across all elements for Kendall's τ, such that $w = 1$. The term $c(i, j)$ represents the concordance function, which evaluates whether the relative order of items i and j is preserved between x and y. If both rankings agree on the order of a pair, the product of signs is positive (concordant); otherwise, it is negative (discordant). The denominator, on the other hand, ensures the coefficients remain bounded in $[-1, 1]$, using $\|x\|_w = \sqrt{\langle x, x \rangle_w}$ to represent the norm of ranking x, with an analogous definition for $\|y\|_w$. Overall, assuming both rankings are filtered to contain the same items, τ reflects the degree of agreement between the two orderings.

Additionally, a nice probabilistic interpretation for τ is based on the following experiment: consider a pair of items at random; return 1 if the pair is ranked in the same order in both lists, otherwise return 0. With p being the expected outcome, τ is defined as $\tau = 2p - 1$, and is therefore proportional to the probability of concordance between the rankings. If the rankings are independent, $p = \frac{1}{2}$, such that the expected value of the coefficient is 0.

While τ has been widely accepted as the standard for measuring the correlation between two rankings [6, 19], it is crucially not top-weighted. Generally, this does not accurately represent real-world human behaviour. Items at the top of the list are more likely to be considered by the user [11], so disagreements in these positions should carry greater importance in the coefficient. To this end, τ_{ap} [19] has been introduced as a measure that reflects user actions, maintaining the same structure as Equations 1 and 2, but with a weighing function defined as:

$$w = w(y_i, y_j) = \frac{1}{\max(y_i, y_j) - 1}. \tag{3}$$

Additionally, the random experiment is now as follows: pick one item at random from x and another one ranked above it; return 1 if they are in the same relative order in y, or 0 otherwise. Similarly to τ, the expected value of τ_{ap} is 0 if x and y are independent.

Lastly, τ_h [17] extends τ to integrate a rank-based top-weighted factor. It can be computed symmetrically by:

$$\tau_h = \frac{\tau_w(x, y) + \tau_w(y, x)}{2}, \tag{4}$$

with the following weighing function:

$$w(\rho_{x,y}(i), \rho_{x,y}(j)) = \frac{1}{1 + \rho_{x,y}(i)} + \frac{1}{1 + \rho_{x,y}(j)}, \tag{5}$$

where $\rho_{x,y}$ is defined by ordering elements lexicographically with respect to x. Throughout this paper, it is assumed that both rankings do not contain ties, thus making the random experiment a scaled version of the one defined for τ, resulting in an expected value of τ_h that is also 0.

On the other hand, research on relevance judgments for system comparisons has primarily revolved around the notion of cumulative gain (CG) [7], which sums the total utility of elements up to a given rank. Its normalised form, normalised discounted CG (nDCG), accounts for the position of each document in the ranking by applying a logarithmic discount to lower-ranked items and normalises by an ideal ranking. Recent work in the field of relevance judgment generation with large language models makes use of the nDCG as a metric on which to create a ranking of systems, thereby computing Kendall's τ [5, 13, 15] to measure the overall agreement compared to a human-made reference ordering.

This work aims to streamline the comparison of systems using relevance values, allowing a direct and fine-grained similarity assessment between ranking profiles by extending τ, τ_{ap}, and τ_h. Given the additional information contained in the proposed coefficients, definitions must not only be logically sound but also present significant empirical deviations from the metrics on which they are based.

3 Definitions of τ, τ_{ap}, and τ_h for ranking profiles

To extend τ, τ_{ap}, and τ_h to ranking profiles, $c(i, j)$, defined in Equation 2, is modified to incorporate relevance judgements. This function lies at the heart of the similarity measure: it determines how item pairs are evaluated. In the original formulations, $c(i, j)$ performs element-wise comparison, adding a binary value if the items are concordant or discordant. In doing so, information regarding item relevance is discarded.

To address this limitation, the redefined concordance function must include the behaviour of the relevance values associated with the observed items. To this end, the following subsection introduces a set of axioms that logically follow from the extension of τ and its variants to ranking profiles. Based on these, several variants for $c(i, j)$ are proposed thereafter, each of which can be applied analogously to τ, τ_{ap}, and τ_h. It must be noted that while gain functions, highlighted in Section 2, are frequently used to model the evolution of utility in a ranking, they are not used in this paper. This is because Kendall's τ and its top-weighted variants inherently perform a pairwise aggregation of item scores based on their relative rankings. Introducing a separate gain function would duplicate this aggregation logic. Instead, the influence of each comparison makes direct use of the relevance values of the items involved, as detailed in the coefficients proposed below.

3.1 Desired properties of τ, τ_{ap}, and τ_h for ranking profiles

Extending τ and its variants to relevance profiles, it may be desirable to maintain some of their most essential properties. In particular, the behaviour of the correlation coefficient at the bounds of $[-1, 1]$ provides significant insight into how it can be interpreted. Accordingly, following the Kendall-based

measures, the extreme values of the revised measures may be obtained axiomatically as below.

Axiom 1. *If the compared rankings are equivalent, the coefficient is* 1.

Specifically, two rankings are equivalent if all items are placed in the same order. Consider, for instance, a ranking z:

$$z = \langle A, B \rangle,$$

with the relevance values $A = B = 2$. The existing correlation measures satisfy $\tau(z, z) = \tau_{ap}(z, z) = \tau_h(z, z) = 1$. It is therefore natural to extend this property to ranking profiles, ensuring that whenever a list is compared to itself, the coefficients yield a value of 1, **regardless** of the relevance values associated with each element. This logic can be further applied to reverse lists.

Axiom 2. *If the compared rankings are reversed, the coefficient is* -1.

Here, a reversed ranking lists items in the opposite order of the ranking to which it is compared. Considering once again z, its reverse, denoted by z', is:

$$z' = \langle B, A \rangle.$$

such that $\tau(z, z') = \tau_{ap}(z, z') = \tau_h(z, z') = -1$ holds. Once again, this property can be applied to ranking profiles such that reversed lists yield a value of -1, **regardless** of the relevance values.

However, a subtle point arises when comparing z to z'. Despite the underlying items being swapped, the ordering of the relevance values remains unchanged for both lists:

$$z_{rel} = z'_{rel} = \langle 2, 2 \rangle.$$

From this perspective, z_{rel} and z'_{rel} are equivalent. The shift from element rankings to ordering of relevance values motivates the following axiom.

Axiom 3. *If the ordering of relevance values in the compared rankings is equivalent, the coefficient is* 1.

Based on this property, computing a relevance-based coefficient between z_{rel} and z'_{rel} yields 1. This can be extended to reversed rankings.

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.

The property above introduces the notion of unique values in each list. This is because, without the underlying identity of the items at each rank, the differentiation between equivalence and a reversal is only possible when all relevance values are distinct. That is, elements with the same utility must either be considered concordant under Axiom 3 or discordant under Axiom 4. In a ranking with non-unique relevance values, the choice of relative priority between these is exclusive: one precludes the other from being satisfied. In this paper, a pair with the same relevance is considered concordant if their underlying identity is unknown. This is motivated by the prior example, in which z_{rel} and z'_{rel} are intuitively equivalent from the perspective of utility.

While Axioms 1 through 4 define behaviour at the extremes of concordance and discordance, it is also desirable to specify the expectation in the absence of any association between rankings.

Axiom 5. *If the compared rankings are independent, the expected value of the coefficient is* 0.

This property ensures that the measure is centred, assigning a neutral value when there is no systematic agreement or disagreement between rankings.

Generally, the first two axioms define concordance by item identity, while 3 and 4 determine it by relevance value ordering, with Axiom 5 always being applicable. Crucially, however, these groups are distinct and cannot always be satisfied with the same coefficient, as shown by z. A formal proof of this fact is as follows.

Lemma 1. *There does not exist a coefficient that satisfies Axioms 2 and 3 for all relevance profiles.*

Proof. Assume, for contradiction, that a coefficient α satisfies Axioms 2 and 3 for all relevance profiles. Consider the ranking $z = \langle A, B \rangle$ with relevance values $A = B = 2$, so that $z_{\text{rel}} = \langle 2, 2 \rangle$. Let $z' = \langle B, A \rangle$, its reverse, which also yields $z'_{\text{rel}} = \langle 2, 2 \rangle$, and thus $z_{\text{rel}} = z'_{\text{rel}}$.

By Axiom 3, it follows that $\alpha(z, z') = 1$. However, since z' is the reverse of z, Axiom 2 implies $\alpha(z, z') = -1$, yielding the contradiction $1 = \alpha(z, z') = -1$. Therefore, no coefficient α can simultaneously satisfy Axioms 2 and 3 in all cases. □

Therefore, only a subset of all properties can hold simultaneously for the same coefficient. If the relevance of items is chosen to determine concordance, Axioms 3, 4, and 5 can be satisfied. On the other hand, if the identity of elements determines concordance, with relevance being used as a weighing factor, Axioms 1, 2, and 5 can hold. In both these cases, modifications for the concordance function are proposed in the following sections, allowing the reader to choose the most suitable option for their work.

3.2 Versions of the concordance function using relevance as a weighing factor

The following redefinition of $c(i, j)$ uses relevance as a weighing factor, with item identity determining the concordance of a pair. This revised measure is constructed to satisfy Axioms 1 and 2. Additionally, if the coefficient remains symmetric and appropriately bounded between $[-1, 1]$, it can directly replace the concordance function used in τ, τ_{ap}, and τ_h, satisfying Axiom 5.

3.2.1 Distance-weighted version. The distance-based similarity coefficient maintains the sign function using the

element ordering defined by Vigna [17], while weighing the relative relevance difference between the items as follows:

$$c_{dw}(i,j) = \text{sign}(x_j - x_i) \cdot \text{sign}(y_j - y_i) \cdot \frac{\left|rel_i - rel_j\right|}{\max(rel_i, rel_j)}, \quad (6)$$

where rel_i and rel_j denote the relevance of items i and j, respectively. Note that if rel_i and rel_j are both 0, $c_{dw}(i,j)$ is also set to 0 to avoid undefined behaviour. Generally, Equation 6 ensures that if both items being compared have the same relevance ($rel_i = rel_j$), the magnitude of the function is minimised, such that $c_{dw}(i,j) = 0$. If, on the other hand, both items have significantly different relevance relative to each other ($rel_i = 3$ and $rel_j = 0$, for example), the magnitude of $c_{dw}(i,j)$ is maximised, allowing for a total concordance (1) or discordance (−1). Overall, this behaviour follows the intuition that the ordering of items with significantly different relevance values conveys more meaningful information, and should therefore have a greater impact on the coefficient. Between its extremes, the Equation 6 linearly decays. Therefore, $c_{dw}(i,j)$ remains bounded in $[-1,1]$, ensuring that it can be used in the standard definitions of τ, τ_{ap}, and τ_h. Additionally, given that the concordance function remains symmetric around 0, all correlation coefficients retain their expected value of 0 when comparing two independent orderings, satisfying axiom 5.

It is important to note that if a ranking is compared to itself, the weighing factor for x and y in the denominator of Equation 1 simply cancels with the numerator. As such, the correlation coefficients yield a value of 1, regardless of the relevance values of the items. This extends to reverse rankings, which are given a correlation of −1. Such behaviour is crucial in maintaining the extension of τ and its derivatives as measures of similarity using element ordering as a weight, thereby satisfying Axioms 1 and 2. There is, however, an edge case that must be considered: if the all relevance values in the conjoint list of items are the same (such that $rel_i = rel_j$ for any feasible (i,j)) then $c_{dw}(i,j)$ would be 0, regardless of the ordering of the elements. In this instance, the relevance-based factor may be ignored (set to 1) for all comparisons, given that the relevance judgments provide no useful information. Therefore, Axioms 1, 2, and 5 still hold.

3.3 Versions of the concordance function using relevance to determine concordance

In contrast to the previous formulation, where item identity defines concordance, the following definitions use relevance values directly to determine the sign of $c(i,j)$ to satisfy Axioms 3 and 4. That is, concordance is assessed based on how relevance is ordered across the two rankings. To do so, a slight shift in notation is required: while the base versions of Kendall-style measures use x_i to indicate the rank of element i in x, the coefficients below use it to indicate the relevance at index i in list x. This enables index-based relevance comparisons. This change is necessary because comparing the relevance of the same item across both lists provides no meaningful information: if i and j denote items, then $rel_{x_i} = rel_{y_i}$ and $rel_{x_j} = rel_{y_j}$. Instead, to capture differences in relevance ordering between the rankings, the relevance values of the items *at* particular indices must be compared.

Additionally, as mentioned in Section 3.1, a reversed ranking of relevance judgments is only defined when the lists contain unique values. As such, this limits the length for perfectly inverse rankings to $max\_rel + 1$ or less. Such behaviour is similar to τ_a [18], which extends τ for rankings containing tied elements (two items are given the same rank) by assigning a concordance score of 0. Therefore, if the rankings include tied items, the lower bound of the coefficient, −1, cannot be obtained. In the case of the following coefficients, if two elements have the same relevance score, they will always be considered concordant according to Axiom 3, preventing a correlation of −1.

3.3.1 Sign-based version. This version introduces a simple sign-based criterion for concordance, $c_{sc}(i,j)$, relying only on whether the direction of change in relevance is consistent across the rankings. The general form is derived from the definition of Kendall's τ, with some modifications to more accurately reflect the interaction of ordered relevance values. As such, $c_{sc}(i,j)$ is defined as:

$$c_{sc}(i,j) = \begin{cases} 1 & \text{if } \text{sign}(rel_{x_j} - rel_{x_i}) = \text{sign}(rel_{y_j} - rel_{y_i}) \\ a & \text{if } \text{sign}(rel_{x_j} - rel_{x_i}) = 0 \\ a & \text{if } \text{sign}(rel_{y_j} - rel_{y_i}) = 0 \\ -1 & \text{otherwise,} \end{cases} \quad (7)$$

where rel_{x_j} refers to the relevance of the item at index j in list x. Therefore, the measure directly checks whether relevance ordering between rankings x and y is preserved. To satisfy Axiom 3, equivalent pairs in Equation 7 are assigned scores of 1. Furthermore, a is a variable representing a partial discordance in orderings, whose value must be set to ensure the expectation of the coefficient is 0 for independent rankings, according to Axiom 5. This property is carried over from τ and guarantees the measure is unbiased. To this end, the interaction between the two sign functions can be modelled as follows:

		List x		
		+1	0	−1
List y	+1	1	a	−1
	0	a	1	a
	−1	−1	a	1

Table 1. Values for $c_{sc}(i,j)$ according to the values of the sign functions for x and y, as presented in Equation 7.

Since both rankings are defined on the same relevance scale with $n = \max\_rel + 1$, their sign value probabilities are symmetric: each has probability $\binom{n}{2} = \frac{n(n-1)}{2}$ for ±1, and n

for 0. Furthermore, x and y are independent, such that the joint probability of each cell in Table 1 is:

$$\mathbb{P}_{ij} = \mathbb{P}[\text{sign}(x) = j] \cdot \mathbb{P}[\text{sign}(y) = i]. \tag{8}$$

Therefore, the expected value of $c_{sc}(i, j)$ is:

$$\mathbb{E}[c_{sc}(i, j)] = \sum_{i,j \in \{+1,0,-1\}} \mathbb{P}_{ij} \cdot c_{sc}(i, j). \tag{9}$$

Plugging in the values from Equation 7 and simplifying, a can be expressed as a function of n by setting $\mathbb{E}[c_{sc}(i, j)]$ to 0:

$$a = -\frac{1}{2 \cdot (n-1)}. \tag{10}$$

The complete working between Equations 9 and 10 can be found in Appendix A.1.

Therefore, for a relevance scale between 0 and 3, the formulation above yields $a = -\frac{1}{6}$. Intuitively, this aligns with the piecewise definition in Equation 7: when one of the sign functions is zero and the other is non-zero, the pair is neither in full agreement nor full disagreement. In these cases, the contribution is treated as a scaled discordance, reflecting a partial misalignment.

Lastly, note that Axiom 4 is not satisfied by this concordance coefficient. Given the total length of the rankings, denoted by N, reversals in the global ordering must not only consider the indices i and j, but also $N - i$ and $N - j$. This can be illustrated by the following example:

$$x_{rel} = \langle 4, 0, 3, 1 \rangle,$$

$$y_{rel} = \langle 3, 1, 4, 0 \rangle,$$

where y_{rel} is the reverse of x_{rel}. By only observing the first two indices in both lists, it is impossible to establish that the rankings are reversed. In fact, according to Equation 7, their sign is concordant, thereby assigning it a 1. As such, $c_{sc}(ij)$ functions as a sign-based pairwise comparator that captures local ordering agreement but lacks sensitivity to full reversals. Nevertheless, this coefficient is valuable as a foundational component: it provides an intuitive translation of the concordance function used in Kendall's τ, making it a useful starting point for constructing more refined coefficients. In particular, it can be incorporated into measures with broader structural information, as shown in the following subsection.

3.3.2 Augmented additive version.

Extending the sign-based approach, the additive concordance coefficient, denoted by $c_{ac}(i, j)$, directly handles reversed rankings, as well as introducing a linearly-decaying structure to differentiate between partial and complete agreements.

$$c_{ac}(i,j) = \begin{cases} 0 & \text{if } rel_{x_i} = rel_{y_i} = rel_{y_{N-i}} \\ & \quad \text{and } rel_{x_j} = rel_{y_j} = rel_{y_{N-j}} \\ & \quad \text{and } (rel_{x_{N-i}} \neq rel_{y_{N-i}} \text{ or } rel_{x_{N-j}} \neq rel_{y_{N-j}}) \\ c_{sc}(i,j,N-i,N-j) \cdot r(i,j,N-i,N-j) & \text{if } (rel_{x_i} \neq rel_{y_i} \text{ or } rel_{x_j} \neq rel_{y_j}) \\ & \quad \text{and } ((rel_{x_i} = rel_{y_{N-i}} \text{ and } rel_{x_j} = rel_{y_{N-j}}) \\ & \quad \text{or } (rel_{y_i} = rel_{x_{N-i}} \text{ and } rel_{y_j} = rel_{x_{N-j}})) \\ c_{sc}(i,j,i,j) \cdot r(i,j,i,j) & \text{otherwise} \end{cases} \tag{11}$$

As mentioned in Section 3.3.1, the reversal of a pair requires the comparison of indices i and j to $N - i$ and $N - j$. To this end, $c_{sc}(i, j)$ can be redefined as a function of four variables:

$$c_{sc}(i,j,k,l) = \begin{cases} 1 & \text{if } \text{sign}(rel_{x_j} - rel_{x_i}) = \text{sign}(rel_{y_k} - rel_{y_l}) \\ a & \text{if } \text{sign}(rel_{x_j} - rel_{x_i}) = 0 \\ a & \text{if } \text{sign}(rel_{y_k} - rel_{y_l}) = 0 \\ -1 & \text{otherwise.} \end{cases} \tag{12}$$

This coefficient is modulated in the second and third cases by a relevance-based agreement factor, differentiating exact matching and partial agreement:

$$r(i,j,k,l) = 1 - \frac{\left|\left(rel_{x_i} + rel_{x_j}\right) - \left(rel_{y_k} + rel_{y_l}\right)\right|}{rel_{x_i} + rel_{x_j} + rel_{y_k} + rel_{y_l}}, \tag{13}$$

where $r(i, j, k, l) \in [0, 1]$. As such, $r(i, j, k, l)$ measures the strength of relevance-based concordance at the given indices, with values of 1 indicating total agreement. Therefore, the multiplication of c_{sc} by r considers the worst possible ordering to be 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.

Taking note of the conditions in Equation 11, each case serves to classify a pair (i, j) as contributing to an equivalent ranking, a reversed ranking, or neither, thereby satisfying Axioms 3 and 4.

The third case of the piecewise function is the default: it applies when none of the more specific conditions are met. In this case, if $rel_{x_i} = rel_{y_i}$ and $rel_{x_j} = rel_{y_j}$, then the items at positions i and j in list x agree with those at the same positions in list y, and thus $c_{ac}(i, j) = 1$, representing full concordance.

The second case handles reversed rankings. This occurs when $rel_{x_i} = rel_{y_{N-i}}$ and $rel_{x_j} = rel_{y_{N-j}}$, indicating that the items at i and j in x match those at $N - i$ and $N - j$ in y, suggesting a reversed ordering. Additionally, to avoid considering elements in ranking y twice, the constraint ($rel_{y_i} = rel_{x_{N-i}}$ and $rel_{y_j} = rel_{x_{N-j}}$) checks whether indices i and j are part of a reversed pair. If this is satisfied, $c_{sc} \cdot r$ is once again computed using $(i, j, N - i, N - j)$. This case also includes a negation: it applies only if at least one of the items differs between the two lists at the same index: $rel_{x_i} \neq rel_{y_i}$ or $rel_{x_j} \neq rel_{y_j}$. This, in conjunction with the condition described in the following paragraph, prevents overlap with the third case when the lists are potentially equivalent.

The first case further resolves ambiguity when both agreement and reversal conditions appear to be satisfied. Specifically, if $rel_{x_i} = rel_{y_i} = rel_{y_{N-i}}$ and $rel_{x_j} = rel_{y_j} = rel_{y_{N-j}}$, the pair may satisfy both concordance and reversal. A tertiary check is then used to determine if the lists are truly equivalent: the values at $N - i$ and $N - j$ must also match in both lists, such that $rel_{x_{N-i}} = rel_{y_{N-i}}$ and $rel_{x_{N-j}} = rel_{y_{N-j}}$.

If this is not the case, the pair is considered ambiguous and is assigned a neutral score of 0. This is illustrated by the following example:

$$x_{rel} = \langle 3, 2, 0, 3, 2 \rangle, \quad y_{rel} = \langle 3, 2, 0, 2, 3 \rangle.$$

Considering indices $i = 0$ and $j = 1$ $rel_{x_0} = 3 = rel_{y_0} = rel_{y_4}$, and $rel_{x_1} = 2 = rel_{y_1} = rel_{y_3}$, suggesting both agreement and reversal. However, the tertiary check reveals:

$$rel_{x_4} = 2 \neq rel_{y_4} = 3, \quad rel_{x_3} = 3 \neq rel_{y_3} = 2.$$

Since the corresponding elements at $N - i = 4$ and $N - j = 3$ do not match, the rankings are neither truly equivalent nor fully reversed. As a result, the first case applies, and the pair is assigned a neutral score of 0.

Lastly, it is important to note that, due to the conditions imposed on the piecewise function, the choice of $a = -\frac{1}{2 \cdot (n-1)}$ does not ensure that $c_{sc}(i, j, k, l)$ is unbiased for independent rankings. Since the function depends on conditionals for both x and y, the probabilities of each cell in Table 1 can no longer be calculated separately for each list. This is further exacerbated in $c_{ac}(i, j)$ by the agreement factor, r, whose scaling also depends on the values of both rankings. Therefore, mathematically deriving the value of a such that Equation 11 maintains an expectation of 0 for independent rankings is outside the scope of this research project, and is left as future work. However, using the independently simulated rankings described in Section 4.2, 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 this dataset is approximately equal to that of $c_{sc}(i, j)$, which is unbiased by construction. By extension, therefore, $c_{ac}(i, j)$ is approximately unbiased for relevance scales of lengths 4 and 5.

4 Experimental Setup and Results

4.1 Real-world data

The TREC (Text REtrieval Conference) Web Track data from 2010 to 2014 [3] consists of benchmark datasets and evaluation resources designed to support research in information retrieval. Using collections of crawled web pages, the ad hoc retrieval task involves finding documents relevant to a query without knowledge of their ground-truth utility. Participants submit retrieval runs, which are evaluated against relevance judgments provided by the National Institute of Standards and Technology (NIST). Over the 5-year span, approximately 150 systems containing rankings of the top 1000 items for 50 topics have been made available. Therefore, the rankings and item-relevance value mapping are used to validate the extended correlation coefficients against a real-world benchmark. Note that max_rel is 3 for the 2010 and 2011 datasets, whereas it is 4 for the remaining 2012 to 2014 period. Furthermore, any ties in a ranking, corresponding to items with the same score, are randomly broken using a seed set to 42, ensuring reproducibility. Lastly, τ and its derivatives require that the lists are conjoint. By preprocessing every pair of rankings to satisfy this constraint, each list often contains significantly fewer than 1000 items.

4.2 Stochastic simulation of ranking profiles

To further test the behaviour of the proposed coefficients over a range of diverse systems, stochastically simulated rankings can be generated using an adaptation of the NSGA-II evolutionary algorithm [4] as presented by Roitero et. al [14], using the jMetalPy framework [1]. In contrast with real-world data, each parameter in the simulation can be tuned, allowing a controlled experimentation environment. For the results obtained in the following section, each relevance profile is represented by an integer array of length 50, with elements bound to the interval $[0, max\_rel]$. As is common practice, $max\_rel = 3$ for the stochastic data.

The initial population consists of a set of profiles characterised by an array $\mathbf{R} = [R_0, R_1, \ldots, R_{max\_rel}]$, where each R_i denotes the maximum number of documents with relevance value i. This constraint is further enforced throughout the evolutionary process. Initially, each rank is given a value based on the following probability:

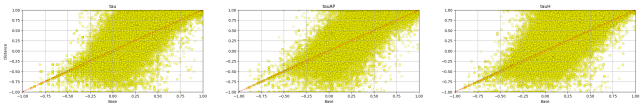
$$\mathbb{P}(rel, rank) = e^{-\frac{(rel+1)(rank - 0.1 \cdot N)}{0.2 \cdot N}}, \tag{14}$$

defined for each possible relevance judgment, rel, where $rank$ is the index of the current item, and N is the length of the list. This formulation, along with the population representation and the anDCG fitness function, are original contributions of this work. On the other hand, the genetic mutations of rankings are adapted from the work of Roitero et al., allowing different profiles to be generated. Specifically, crossovers add or multiply elements from two parent arrays, and mutations randomly swap two elements or add a random value to an item. Lastly, different systems are simulated by randomly generating a target average nDCG (anDCG) value [7] in the range $[0.1, 0.9]$, which is defined as the mean of the nDCG scores over all possible depths in the ranking. Specifically, the nDCG at rank p is given by $nDCG_p = \frac{DCG_p}{IDCG_p}$, where $IDCG_p$ is the maximum possible DCG (ideal DCG) at p, computed by sorting the relevance scores in decreasing order. Additionally, the DCG at rank p is $DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$, where rel_i is the relevance score of the item at rank i.

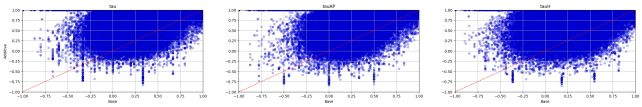
Using the simulated profiles obtained as described above, a semi-random ranking is achieved via a correspondence between random hashes and relevance values. As such, each item in a profile is given a unique hash selected from a pool associated with its relevance level, preserving the relevance distributions. This mapping allows lists to be compared and evaluated as rankings of distinct, relevance-labelled items.

4.3 Results

Although the reasoning behind the coefficients proposed in Section 3 is described in detail, it is crucial to understand

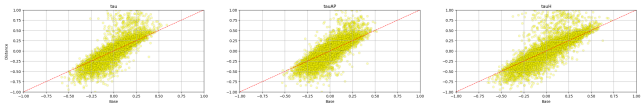


(a) dw variant (Equation 6) using item-based concordance, with relevance as a weighing factor.

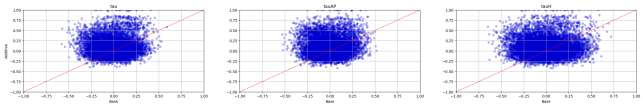


(b) ac variant (Equation 11) using relevance-based concordance.

Figure 1. Comparison of τ, τ_{ap}, and τ_h (left to right) to the proposed coefficients for the 2010 - 2014 TREC data.



(a) dw variant (Equation 6) using item-based concordance, with relevance as a weighing factor.



(b) ac variant (Equation 11) using relevance-based concordance.

Figure 2. Comparison of τ, τ_{ap}, and τ_h (left to right) to the proposed coefficients for simulated data.

their behaviour using the collected ranking data. To this end, the concordance functions form two new similarity measures for each variant of τ. Namely, dw denotes the distance-based version of $c(i, j)$ that uses relevance as a weighing function (Equation 6), while ac refers to the augmented additive version of $c(i, j)$ that uses relevance to determine concordance (Equation 11). As such, a comparison to τ, τ_{ap}, and τ_h is made to grasp the impact of relevance-based measures on the similarity evaluation. Specifically, a significant difference between the base and relevance-aware coefficients is expected, given that relevance values capture information regarding utility, rather than a strict item ordering.

Figure 1 illustrates the transparency scatter plots comparing the proposed coefficients to the Kendall-type variants, aggregated over all system comparisons for the 2010 to 2014 TREC data. Specifically, Figures 1a and 1b demonstrate the scatter plots comparing the dw and ac variants, respectively, to τ, τ_{ap}, and τ_h, from left to right. As such, the points in a given plot represent the paired values of a base metric and its relevance-aware counterpart for two systems over a particular topic, with lower transparency indicating higher frequency. Additionally, the diagonal red line represents the identity line, such that points deviating from it reflect the extent to which incorporating relevance affects rank correlation.

Comparing the behaviour of the dw and ac variants in Figure 1, the former generally shows higher agreement with the baseline metrics. This indicates that maintaining the item-based concordance and introducing a distance-based weighting causes a relatively modest shift in correlation. Notably, however, the dw variant produces a significant quantity of ± 1 outcomes. This can be attributed to the fact that pairs with the same relevance value are not considered: given that their utility is the same, their ordering does not matter in evaluating relevance similarity between lists. In contrast, the ac variant, which uses relevance to determine the sign of the concordance coefficient, differs substantially from the Kendall-style measures. Given that item-identity, the core of τ and its variants, is discarded, $c_{ac}(i, j)$ is expected to produce significantly different interpretations of similarity. These trends can be confirmed numerically by comparing the absolute differences between the base and relevance-aware variants, as shown in Table 2 of Appendix A.2. For instance, considering Kendall's τ, the average difference increases from 0.14 for $|\tau - \tau_{dw}|$ to 0.39 for $|\tau - \tau_{ac}|$. This pattern is further observed for all variants of τ, as well as in the proportion of large differences (in $(0.1, 2]$).

Furthermore, as discussed in Section 3.3, never ac attains a score of -1. This is because the maximum relevance score in the TREC datasets is 3 or 4, depending on the year, while the length of common items is greater. Thus, many documents share the same score, preventing complete discordance. On the other hand, a significant number of rankings in Figure 1b return 1, while the base coefficients do not. These indicate that many rankings are fundamentally the same, regardless of the underlying ordering of the elements. This can be partially attributed to the fact that, on average, more than half of the available documents in the TREC dataset have a relevance of 0. As a result, ranking pairs with few non-zero relevance scores and many zeros are likely to contain a large number of concordant items based solely on relevance ordering.

In addition, Figure 2 illustrates the transparency scatter plots for the simulated data. Generally, similar trends are observed as with the TREC data, with the dw variant exhibiting the least variation. In contrast, the relevance-based concordance variant shows larger deviations, differing by more than 0.1 in over 65% of comparisons to all Kendall-style measures (the full numerical results can be found in Table 3 of Appendix A.3). Notably, however, the ac variant tends to produce more positive correlation values. While dw averages approximately 0.0 across all τ types for this dataset, ac has a higher mean of 0.09. This difference can be attributed to the fact that, although the underlying items are randomised for each ranking, the distribution used to generate each relevance value remains unchanged. As a result, the relevance-based concordance function is more likely to produce concordant pairs on average. Furthermore, given that

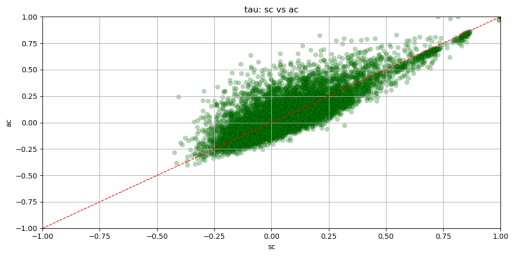


Figure 3. Comparison of the *sc* and *ac* variants using simulated data.

the simulations generate independent rankings, the range of τ similarity is limited to approximately $[-0.6, 0.6]$, centred around 0.

Lastly, an interesting comparison can be made between the sign-based concordance (*sc*, Equation 7) and augmented additive concordance (*ac*) coefficients, given that they both solely rely on relevance values. To this end, Figure 3 compares these variants, using the simulated data to compute the relevance-aware measures of τ. Generally, the plot demonstrates a high agreement for correlation values close to 1, with gradually increasing spread for rankings with more discordant pairs. This effect can be explained by the inclusion of the agreement factor in the *ac* coefficient, r, which generates a range of values in $[-1, 1]$ rather being limited to $-1, 1,$ or a. Furthermore, the points in Figure 3 seem to be slightly skewed upwards compared to the identity line. This is supported by the average coefficient values: the concordance measure $c_{sc}(i, j)$ yields an average of 0.08, while $c_{ac}(i, j)$ gives a slightly higher mean of 0.09. This highlights the limitation of empirically setting the value of a for the additive variant, as the measure still demonstrates a small degree of bias for $n = 4$.

5 Discussion

The empirical analysis conducted using real-world and simulated data shows that integrating relevance values into rank correlation measures yields significantly different coefficients compared to traditional Kendall-style measures. From a practical standpoint, this result underscores a limitation of traditional rank correlation metrics when applied in information retrieval settings: they fail to capture how systems rank items based on their utility. In contrast, the proposed relevance-aware coefficients provide a more nuanced measure, distinguishing orderings that agree not only in rank positions but also in how similarly they identify element relevance.

Furthermore, the choice of concordance function has a measurable impact on sensitivity. Notably, the additive variant (*ac*) produces the greatest deviation from traditional metrics by fully decoupling rank similarity from document identity. In contrast, the distance-weighted variant (*dw*) maintains a partial connection to item order. As such, correlation measures can be selected based on the specific research objective, using either item identity or utility to determine the sign of the concordance function.

Additionally, an adaptation of the stochastic simulation algorithm to relevance profiles has been presented to complement real-world data. Using tunable parameters, realistic systems can be generated with specific rank-based distributions. However, a significant limitation of the implementation remains: each system is simulated independently, limiting the range of correlation values to around 0. To partially compensate for this, shorter list lengths were used to obtain more distinct rankings, although, as previously mentioned, the full $[-1, 1]$ range could not be achieved.

It is important to note that the correlation coefficients proposed in this work may not be suitable for all use cases. In particular, the relevance-based concordance measures have been designed to satisfy Axioms 3 and 4. While these provide a principled approach, they may not align with scenarios where greater emphasis should be placed on highly relevant documents, for instance. Moreover, the reversal property described in Axiom 4 introduces significant complexity: it requires the coefficient to account for positional changes across different indices in the rankings, resulting in a piecewise formulation with several conditions as shown in Equation 11. This not only reduces interpretability but also introduces interdependencies between rankings, making it more difficult to define an unbiased measure. Despite these limitations, it is hoped that the proposed coefficients can still serve as a framework for other relevance-based extensions of correlation.

6 Conclusion and Future Work

In conclusion, this study has extended Kendall's τ and its top-weighted variants to account for graded relevance judgments in ranking profiles. By introducing alternative definitions of the concordance function, the proposed coefficients incorporate the ordering of item utility following two distinct approaches.

Generally, the results obtained across TREC and simulated datasets confirm that relevance-aware measures diverge meaningfully from their traditional counterparts, particularly when relevance is used to determine concordance. These findings emphasise the need for relevance-sensitive evaluation in IR research, enabling a more complete and fine-grained similarity analysis based on the item utility.

Future work can proceed along several directions. First, a mathematically derived unbiased variation of the $c_{ac}(i, j)$ concordance coefficient may be investigated. Furthermore, an analysis of the stability of the proposed measures, as in the work by Buckley and Voorhees [2], may indicate how well relevance scores can discriminate between different systems. Lastly, exploring the integration of these coefficients into IR evaluation libraries, such as ircor [16], may facilitate their practical adoption.

7 Responsible Research

Given that this work solely involved redefining and testing correlation coefficients, there were few ethical concerns throughout the research process. Of note, however, is the reproducibility of all aspects of the paper, such that a fellow researcher can independently recreate what is described. To facilitate this, the stochastic simulations and computation of the results in Section 4 were performed using code that has been made publicly available on GitHub[1]. A ReadMe file detailing the structure of the codebase used to compute the coefficients has been added, ensuring that all findings can be reproduced with relatively little effort. Furthermore, the simulation algorithm is an adaptation of the work partially presented by Julián Urbano, a supervisor of this paper. Therefore, given their expertise, there was a bias towards choosing that particular algorithm, and other alternatives were not thoroughly considered. In addition, as the stochastic ranking generation is also a contribution of this paper, the codebase contains significant documentation, allowing the reader to gain a full understanding of the simulation's inner workings.

References

[1] Antonio Benítez-Hidalgo, Antonio J. Nebro, José García-Nieto, Izaskun Oregi, and Javier Del Ser. 2019. jMetalPy: A Python framework for multi-objective optimization with metaheuristics. *Swarm and Evolutionary Computation* (2019), 100598.

[2] Chris Buckley and Ellen M. Voorhees. 2017. Evaluating Evaluation Measure Stability. *SIGIR Forum* 51, 2 (2017), 235–242.

[3] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. 2010. Overview of the TREC 2010 Web Track. In *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*, Vol. 500-294. National Institute of Standards and Technology (NIST).

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.

[5] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. *ICTIR 2023 - Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval* (2023), 39–50.

[6] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing top k lists. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 28–36.

[7] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[8] Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[9] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[10] Massimo Melucci. 2007. On rank correlation in information retrieval evaluation. *SIGIR Forum* 41, 1 (2007), 18–33.

[11] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: what observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, 659–668.

[12] Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London* 58 (1895), 240–242.

[13] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2025. Judging the Judges: A Collection of LLM-Generated Relevance Judgements.

[14] 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), 2217–2220.

[15] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *arXiv:2406.06519* (2024).

[16] Julián Urbano and Mónica Marrero. 2017. The Treatment of Ties in AP Correlation. In *ACM SIGIR International Conference on the Theory of Information Retrieval*. 321–324.

[17] 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.

[18] Max A. Woodbury. 1940. Rank Correlation when There are Equal Variates. *The Annals of Mathematical Statistics* 11, 3 (1940), 358–362.

[19] 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.

A Appendix A

A.1 Working between Equations 9 and 10

Starting from equation 9:

$$\mathbb{E}[c_{ij,sc}] = \sum_{i,j \in \{+1,0,-1\}} \mathbb{P}_{ij} \cdot c_{ij,sc}.$$

Using the values from Table 1, we can organize the terms as follows:

$$\begin{aligned}\mathbb{E}[c_{ij,sc}] = &\underbrace{p_{y+}p_{x+} \cdot 1}_{(+1,+1)} + \underbrace{p_{y+}p_{x0} \cdot a}_{(+1,0)} + \underbrace{p_{y+}p_{x-} \cdot (-1)}_{(+1,-1)} \\ &+ \underbrace{p_{y0}p_{x+} \cdot a}_{(0,+1)} + \underbrace{p_{y0}p_{x0} \cdot 1}_{(0,0)} + \underbrace{p_{y0}p_{x-} \cdot a}_{(0,-1)} \\ &+ \underbrace{p_{y-}p_{x+} \cdot (-1)}_{(-1,+1)} + \underbrace{p_{y-}p_{x0} \cdot a}_{(-1,0)} + \underbrace{p_{y-}p_{x-} \cdot 1}_{(-1,-1)}.\end{aligned}$$

Grouping the coefficients of 1, a, and -1 terms:

$$\begin{aligned}\mathbb{E}[c_{ij,sc}] = &(p_{y+}p_{x+} + p_{y0}p_{x0} + p_{y-}p_{x-}) \cdot 1 \\ &+ (p_{y+}p_{x0} + p_{y0}p_{x+} + p_{y0}p_{x-} + p_{y-}p_{x0}) \cdot a \\ &+ (p_{y+}p_{x-} + p_{y-}p_{x+}) \cdot (-1).\end{aligned}$$

[1] Code available at https://github.com/avezzuto1

Substituting the expressions for the probabilities:

$$p_{x+} = p_{x-} = p_{y+} = p_{y-} = \frac{n(n-1)}{2}$$
$$p_{x0} = p_{y0} = n$$

Computing each group of terms:

- Terms with coefficient 1:

$$p_{y+}p_{x+} + p_{y0}p_{x0} + p_{y-}p_{x-} = \left(\frac{n(n-1)}{2}\right)^2 + n^2 + \left(\frac{n(n-1)}{2}\right)^2 = n^2 + 2 \cdot \left(\frac{n(n-1)}{2}\right)^2$$

- Terms with coefficient a:

$$p_{y+}p_{x0} + p_{y0}p_{x+} + p_{y0}p_{x-} + p_{y-}p_{x0} = 2 \cdot \left(\frac{n(n-1)}{2} \cdot n\right) + 2 \cdot \left(n \cdot \frac{n(n-1)}{2}\right) = 4 \cdot \frac{n^2(n-1)}{2}$$

- Terms with coefficient -1:

$$p_{y+}p_{x-} + p_{y-}p_{x+} = 2 \cdot \left(\frac{n(n-1)}{2} \cdot \frac{n(n-1)}{2}\right) = 2 \cdot \left(\frac{n(n-1)}{2}\right)^2$$

Summing these:

$$\mathbb{E}[c_{ij,sc}] = \left(n^2 + 2 \cdot \left(\frac{n(n-1)}{2}\right)^2\right) \cdot 1 + \left(4 \cdot \frac{n^2(n-1)}{2}\right) \cdot a - 2 \cdot \left(\frac{n(n-1)}{2}\right)^2.$$

Cancelling out the positive and negative contributions from the squared terms:

$$2 \cdot \left(\frac{n(n-1)}{2}\right)^2 - 2 \cdot \left(\frac{n(n-1)}{2}\right)^2 = 0$$

We are left with:

$$\mathbb{E}[c_{ij,sc}] = n^2 + 4 \cdot \frac{n^2(n-1)}{2} \cdot a$$

Simplifying to:

$$\mathbb{E}[c_{ij,sc}] = n^2 + 2n^2(n-1) \cdot a.$$

To ensure the coefficient is unbiased for independent rankings, the expected value must be zero:

$$n^2 + 2n^2(n-1) \cdot a = 0$$

Lastly, solving for a:

$$2n^2(n-1) \cdot a = -n^2$$
$$a = -\frac{1}{2(n-1)}$$

Which is the final result in Equation 10.

A.2 Absolute difference metrics between the base and relevance-aware coefficients for 2010 - 2014 TREC data

Target	Variant	Avg.	Max.	M	L
τ	$\lvert\tau - \tau_{dw}\rvert$	0.14	1.58	27.90%	49.11%
	$\lvert\tau - \tau_{ac}\rvert$	0.39	2	11.34%	77.71%
τ_{ap}	$\lvert\tau_{ap} - \tau_{ap,dw}\rvert$	0.12	1.59	34.09%	42.73%
	$\lvert\tau_{ap} - \tau_{ap,ac}\rvert$	0.37	2	13.76%	75.35%
τ_h	$\lvert\tau_h - \tau_{h,dw}\rvert$	0.18	1.62	22.33%	55.35%
	$\lvert\tau_h - \tau_{h,ac}\rvert$	0.40	2	11.82%	77.57%

Table 2. Summary of differences for τ, τ_{ap}, and τ_h for 2010 - 2014 TREC data. M represents medium absolute differences in $(0.01, 0.1]$, and L represents large differences in $(0.1, 2]$.

A.3 Absolute difference metrics between the base and relevance-aware coefficients for simulated data

Target	Variant	Avg.	Max.	M	L
τ	$\lvert\tau - \tau_{dw}\rvert$	0.09	1.08	62.41%	26.05%
	$\lvert\tau - \tau_{ac}\rvert$	0.21	1.23	30.25%	66.27%
τ_{ap}	$\lvert\tau_{ap} - \tau_{ap,dw}\rvert$	0.10	0.94	59.02%	31.61%
	$\lvert\tau_{ap} - \tau_{ap,ac}\rvert$	0.21	1.25	30.56%	65.89%
τ_h	$\lvert\tau_h - \tau_{h,dw}\rvert$	0.10	1.21	60.85%	28.69%
	$\lvert\tau_h - \tau_{h,ac}\rvert$	0.24	1.28	24.38%	72.66%

Table 3. Summary of differences for τ, τ_{ap}, and τ_h for simulated data. M represents medium absolute differences in $(0.01, 0.1]$, and L represents large differences in $(0.1, 2]$.