The Treatment of Ties in AP Correlation

Urbano, Julián; Marrero, Mónica

DOI
10.1145/3121050.3121106

Publication date
2017

Document Version
Accepted author manuscript

Published in
Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval, ICTIR 2017

Citation (APA)
Urbano, J., & Marrero, M. (2017). The Treatment of Ties in AP Correlation. In *Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval, ICTIR 2017* (pp. 321-324). Association for Computing Machinery (ACM). https://doi.org/10.1145/3121050.3121106

The Treatment of Ties in AP Correlation

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

Mónica Marrero
Delft University of Technology
Delft, The Netherlands
m.marrerollinares@tudelft.nl

ABSTRACT

The Kendall tau and AP correlation coefficients are very commonly use to compare two rankings over the same set of items. Even though Kendall tau was originally defined assuming that there are no ties in the rankings, two alternative versions were soon developed to account for ties in two different scenarios: measure the accuracy of an observer with respect to a true and objective ranking, and measure the agreement between two observers in the absence of a true ranking. These two variants prove useful in cases where ties are possible in either ranking, and may indeed result in very different scores. AP correlation was devised to incorporate a top-heaviness component into Kendall tau, penalizing more heavily if differences occur between items at the top of the rankings, making it a very compelling coefficient in Information Retrieval settings. However, the treatment of ties in AP correlation remains an open problem. In this paper we fill this gap, providing closed analytical formulations of AP correlation under the two scenarios of ties contemplated in Kendall tau. In addition, we developed an R package that implements these coefficients.

KEYWORDS

Evaluation; Correlation; Kendall; Average Precision; Ties

1 INTRODUCTION

The Kendall τ [5] and Yilmaz τ_{ap} [17] rank correlation coefficients are frequently employed in Information Retrieval to compare two rankings X and Y given to a set of n items. For instance, Baeza-Yates et al. [1] compared the ranking of webpages produced by crawling algorithms with the ranking produced by PageRank, and White et al. [15] compared different rankings of terms in a study of implicit feedback. These correlation coefficients are particularly common in evaluation studies to compare the rankings of retrieval systems produced by different evaluation conditions, such as different evaluation measures [9], topic sets [3], assessors [14], experts vs. non-experts [2], or even to compare it to the ranking produced by user ratings [10] or the ranking over populations of topics [13].

Both τ and τ_{ap} were originally defined under the assumption that no ties are present in either ranking, so that every item is assigned one integer rank from 1 to n. However, in practical applications there are cases in which several items are considered equal and no preference is given to any of them. As Kendall [6] put it himself:

> *this effect may arise either because the objects really are indistinguishable, [...] or because the observer is unable to discern such differences as exist.*

According to Student [12], Pearson was first in contemplating the issue of ties in ranking problems, for which he suggested several ways to assign ranks to tied items [8]. Following Pearson, Student investigated the effect of ties in the calculation of the Spearman ρ correlation coefficient through its analogy to the product-moment correlation between the rankings. Woodbury [16] also studied the treatment of ties in Spearman ρ, but suggested a different alternative. Following a general definition of correlation by Daniels [4], Kendall [6] applied the principles of Student and Woodbury to his τ correlation coefficient, and identified the two versions as pertaining to two different scenarios:

a) The variant by Woodbury [16] assumes that one of the rankings, say X, is in fact a true and objective ranking in which no ties are present, and Y is the ranking given by an observer which may sometimes fail to distinguish some items and therefore assigns them the same rank. The correlation in this scenario is hence used as a measure of the *accuracy* of the observer. He coined this coefficient τ_a.
b) The variant by Student [12] assumes that both X and Y are rankings given by two observers, both of which may decide to tie some items. In this case, there is no objective ranking to compare with, so the correlation is used as a measure of *agreement* between the two observers. He coined this coefficient[1] τ_b.

As will be evident in the following sections, these two scenarios are fundamentally different and may lead to significantly different scores, so it is important to choose the most appropriate in each case. To the best of our knowledge though, the τ_{ap} correlation coefficient of Yilmaz et al. [17] has not been defined in the presence of ties as τ has. Smucker et al. [11] briefly confronted this problem in scenario a), but approached it numerically. In this paper we fill this gap and provide closed analytical formulations of τ_{ap} under both scenarios of ties. Of course, we coin them $\tau_{ap,a}$ and $\tau_{ap,b}$.

In addition, and to promote its use, we provide implementations of these correlation coefficients in a fully-fledged R package called `ircor`, available from http://github.com/julian-urbano/ircor/.

2 CORRELATION WITHOUT TIES

Let $X = \langle x_1, \ldots, x_n \rangle$ be the true ranking of a set of n items, and let $Y = \langle y_1, \ldots, y_n \rangle$ be an alternative ranking given to the same items.

[1] Initially, Kendall [6] used the terms τ_W and τ_S after Woodbury and Student, but he recoined them as τ_a and τ_b when enumerated both scenarios a) and b) in his later book [7], similar to what we did here.

ICTIR'17, October 1–4, 2017, Amsterdam, The Netherlands.

978-1-4503-4490-6/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3121050.3121106

For illustration purposes, let us consider the following ranks given to a set of $n = 6$ items: $X = \langle 1, 2, 3, 4, 5, 6\rangle$ and $Y = \langle 2, 3, 1, 4, 6, 5\rangle$. If our items were identified by letters we would have rankings $X = \langle \mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}\rangle$ and $Y = \langle \mathsf{C}, \mathsf{A}, \mathsf{B}, \mathsf{D}, \mathsf{F}, \mathsf{E}\rangle$. If we consider all items in pairs, a distance between the two rankings can be computed by counting how many pairs are concordant or discordant between the two rankings: a pair is concordant if their relative order is the same in both rankings, and discordant otherwise. Kendall [5] followed this idea to define his τ correlation coefficient

$$\tau = \frac{\#concordants - \#discordants}{\#total} = \frac{4}{n(n-1)} \sum_{i<j} c_{ij} - 1, \quad (1)$$

where c_{ij} equals 1 if items i and j are concordant (recall that no ties are permitted yet):

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

In our example, the pair (B, C) is discordant because it has ranks $(2, 3)$ in X and ranks $(3, 1)$ in Y. Counting all pairs as in (1), the correlation is $\tau = 0.6$. The fraction of concordant pairs can be interpreted as the expected value of a random experiment: pick two arbitrary items and return 1 if they are concordant, or 0 if they are discordant. The Kendall τ coefficient can thus be interpreted in terms of the probability of concordance.

Yilmaz et al. [17] followed this idea to define a correlation coefficient with the same rationale as Average Precision, thus penalizing more heavily if swaps occur between items at the top of the ranking, much like AP penalizes more if the non-relevant documents appear at the top of the search results. The random experiment is now as follows: pick one item at random from Y and another one ranked above it, and return 1 if they are concordant, or 0 if they are discordant. Their AP correlation coefficient is similarly calculated by traversing the ranking Y from top to bottom[2]:

$$\tau_{ap} = \frac{2}{n-1} \sum_{i=2}^{n} \left(\frac{\#concordants\ above\ i}{i-1} \right) - 1 = \\ = \frac{2}{n-1} \sum_{i=2}^{n} \sum_{j<i} \frac{c_{ij}}{i-1} - 1. \quad (3)$$

In our example, we find $\tau_{ap} = 0.32$.

3 CORRELATION WITH TIES

Under the considerations of Woodbury [16] and Student [12], a tie reflects the *inability* of the observer to decide which of two items should be ranked first. Therefore, in the presence of a tie a pair of items can be considered neither concordant nor discordant; it is simply ignored. In the following subsections, we discuss how ties affect τ_a and τ_b, and provide the corresponding definitions for τ_{ap}.

3.1 Correlation as Measure of Accuracy

For illustration purposes, let us consider the true objective ranking $X = \langle 1, 2, 3, 4, 5, 6\rangle$ and the ranking $Y = \langle 2, 4, 1, 4, 6, 4\rangle$ estimated by an observer, in which items B, D and F are tied. The observer was unable to distinguish these three items, but she should have because there really is an objective order. When counting the number of concordant pairs in (1), we can not really penalize or reward pairs (B, D), (B, F) and (D, F) because the observer did not really decide in either direction. However, these pairs are still counted in the denominator because she was *expected* to distinguish them. The correlation is thus defined as

$$\tau_a = \sum_{i<j} \frac{sign(x_j - x_i) \cdot sign(y_j - y_i)}{n(n-1)/2}, \quad (4)$$

Note that a concordant pair contributes +1 in the numerator, and a discordant pair contributes −1. A tied pair, on the other hand, contributes 0 in spite of it being expected in the denominator. In our example, the correlation is $\tau_a = 0.4$.

Woodbury [16] noted an interesting way of looking at the problem of ties in this scenario: what is the *average* correlation over all the possible permutations of the tied items in Y? As it turns out, if we replace any tied set t by integer ranks and average for all $t!$ possible orders we obtain the same formula as in (4). In our example there are six permutations: $\langle 2, 3, 1, 4, 6, 5\rangle$, $\langle 2, 3, 1, 5, 6, 4\rangle$, $\langle 2, 4, 1, 3, 6, 5\rangle$, $\langle 2, 4, 1, 5, 6, 3\rangle$, $\langle 2, 5, 1, 3, 6, 4\rangle$ and $\langle 2, 5, 1, 4, 6, 3\rangle$, with τ scores of $0.6, 0.467, 0.467, 0.333, 0.333$ and 0.2, respectively; their average is indeed $\tau_a = 0.4$. This result is precisely what we use next to define the corresponding version of τ_{ap} correlation: $\tau_{ap,a}$.

Back in (3), we can see that all untied items will contribute the same in all permutations when acting as the pivot item i. If there are groups of ties above it, each of their items will be concordant or discordant with respect to the pivot i, *regardless* of their position within the tie. For instance, when the pivot is E ($i = 6$), items B and D are both concordant and F is discordant, regardless of the position they have within their tied group. On the other hand, when the pivot i is a tied item we have to consider two terms separately:

I) The contribution of all items ranked above its tied group.
II) The contribution of the items within the group in all permutations in which they are ranked above the pivot.

Let t_i be the number of items tied with the pivot i, inclusive, and let p_i be the position of the first item in the group (the typical ranks used in sports). In our example, the estimated ranking is $Y = \langle C, A, (B, D, F), E\rangle$, so there are $t_3 = t_4 = t_5 = 3$ items in the tie, the first of which appears in position $p_3 = p_4 = p_5 = 3$. Note that if i is not tied, then $t_i = 1$ and $p_i = i$.

For the first term I, we can see that the items above the tied group which are concordant with the pivot remains the same in all permutations. In our example, A is always concordant with the pivot B regardless of the permutation, and C is always discordant. In general, the number of concordants above the pivot i is

$$\sum_{j<p_i} c_{ij}. \quad (5)$$

However, these items will have a different contribution depending on the specific position within the tie that the pivot has in each permutation. Across all $t_i!$ permutations, a tied item will have position p_i a total of $(t_i - 1)!$ times, position $p_i + 1$ another $(t_i - 1)!$ times, etc. Therefore, just like the factor $i-1$ normalizes the number of concordants in (3), these positions normalize the number of

[2] Throughout the paper, indexes refer to items sorted by the order given in Y (eg. in $Y = \langle 2, 3, 1\rangle$, the sorted items are $\langle \mathsf{C}, \mathsf{A}, \mathsf{B}\rangle$, so $i = 2$ refers to item A).

concordants in (5), so the average contribution of term I is:

$$\sum_{j<p_i} c_{ij} \cdot \sum_{k=1}^{t_i} \frac{1}{t_i\,(p_i+k-2)}. \tag{6}$$

For the second term II, we shall calculate the average contribution of all pairs within the tied group and over all permutations. There are $\binom{t_i}{2}$ such pairs to consider across $t_i!$ permutations. Note that two arbitrary items R and S will appear in order (R, S) in half the permutations and in the order (S, R) in the other half; on average, every pair will thus be concordant in half the cases. Without loss of generality, let us assume that the correct order is in fact (R, S).

Again, the individual contribution of this pair needs to be normalized according to the position of the pivot S. When it is in position p_i (ie. the first of the group), the pair (R, S) is not possible because R can never appear before S. When the pivot is in position $p_i + 1$ (ie. the second of the group), there are $(t_i - 2)!$ permutations in which R is arranged before it. When the pivot is in position $p_i + 2$, there are $2(t_i - 2)!$ permutations with R arranged before it: $\langle$R,∗,S, . . . $\rangle$ and $\langle$∗, R, S, . . . $\rangle$. In general, when the pivot S is in position $p_i + k$, there are $k \cdot (t_i - 2)!$ permutations where R appears before it.

As before, these positions are used to normalize the contribution of each pair, and the number of permutations is used to average these contributions across permutations. All in all, there are thus $\binom{t_i}{2}$ pairs of items in the tied group, each of which can appear in positions $k = 1, \ldots, t_i - 1$ within the group a total of $k \cdot (t_i - 2)!$ times, in each of which it contributes one concordant pair normalized by $p_i + k - 1$. Averaging over all $t_i!$ permutations, term II becomes

$$\frac{1}{2} \sum_{k=1}^{t_i-1} \frac{k}{p_i + k - 1}. \tag{7}$$

Putting together terms I and II, the $\tau_{ap,a}$ correlation is therefore

$$\tau_{ap,a} = \frac{2}{n-1}\left(\sum_{i=t_1+1}^{n} \sum_{j<p_i} c_{ij} \cdot \sum_{k=1}^{t_i} \frac{1}{t_i\,(p_i+k-2)} + \sum_{i=1}^{n} \frac{1}{2t_i} \sum_{k=1}^{t_i-1} \frac{k}{p_i+k-1} \right) - 1, \tag{8}$$

where the second term II is further divided by t_i because it will be added t_i times when traversing the tied elements in the outer summation. There are three final remarks worth mentioning. First, note that the summation is taken over all items in the estimated ranking Y, but the order in which the tied items are arranged does not alter the final score. Second, in the absence of ties the third summation equals $1/(i-1)$ and the last summation equals 0, so $\tau_{ap,a}$ reduces to τ_{ap}. Third, because tied groups are disjoint and the summation considers separately the contribution of items outside and within groups, this formulation generalizes to several tied groups. Also note that in the extreme case that the observer ties all elements, $\tau_{ap,a}$ equals 0, as does τ_a. In our example, the correlations with each of the permutations are 0.32, 0.22, 0.253, 0.153, 0.22 and 0.087, for an average of $\tau_{ap,a} = 0.209$.

3.2 Correlation as Measure of Agreement

When both X and Y are produced by observers, there is no notion of true and objective ranking. Similar to the scenario of τ_a, if either observer produced a tie for a pair of items, we can not really reward or penalize his indecision, regardless of how the other observer ranked them. However, in this scenario all these tied pairs are not really expected to be untied from one observer to another: whether the other observer is right or wrong with respect to some unknown truth, the fact remains that his agreement with the current observer can not be measured. In the extreme case of an observer tying all items, τ_a is 0 to reflect that he is no better than chance at ranking the items, but τ_b would be undefined because there is no pair of systems to calculate his agreement with the other observer. The correlation is thus defined as

$$\tau_b = \sum_{i<j} \frac{sign(x_i - x_j) \cdot sign(y_i - y_j)}{\sqrt{n(n-1)/2 - t_X}\sqrt{n(n-1)/2 - t_Y}}, \tag{9}$$

where t_X and t_Y are the number of tied pairs in X and Y, respectively. Note that (4) and (9) differ only in the denominator, reflecting out intuition as to how the former expects all possible pairs to be concordant, while the latter only expects this of the untied pairs. Indeed, the fact that the denominator includes one term for each ranking nicely shows that both observers may expect a different number of concordant pairs, according to their own ranking and regardless of the other. Note that in this case it does not make sense to average over all permutations, because not all pairs are expected.

To illustrate, let us include a tie between the third and fourth elements (C and D) of the first ranking in the previous example, so that one observer produced ranking $X = \langle 1, 2, 3.5, 3.5, 5, 6\rangle$ and the other one produced the same $Y = \langle 2, 4, 1, 4, 6, 4\rangle$ as before. In this case there are $t_X = 1$ and $t_Y = 3$ tied pairs, so the denominator is 12.961. In the numerator there are 8 concordant pairs and 3 discordants, so the final correlation is $\tau_b = 0.386$.

It is not immediate how to adapt AP correlation in this scenario, because (3) is computed by traversing the estimated ranking from top to bottom, computing concordants with an objective truth. However, here there is no notion of true and estimated ranking, so in principle we can not decide which of the two rankings we traverse. In situations like this, Yilmaz et al. [17] suggested to compute a symmetrized version of τ_{ap} by computing the mean of the correlation of X with respect to Y and the correlation of Y with respect to X, that is, assuming that one ranking is the truth and the other one the estimate, and vice-versa.

Approaching $\tau_{ap,b}$ as a symmetrized version still requires two changes in the original formulation of (3) in order to mimic the behavior of τ_b. First, if the pivot i is part of a tied group we need only count its concordants among all items ranked above the group (term I in $\tau_{ap,a}$), because the items within the group will not contribute anything (term II in $\tau_{ap,a}$). The number of concordants is thus as in (5), but normalized by $p_i - 1$ rather than by $i - 1$. In our example, if the pivot is D ($i = 4$) in Y, only items C and A are ranked above its group, of which only A is concordant in X. By ignoring B and F from Y we mimic the second term in the denominator of τ_b as well as the second term in the numerator (ie. do not expect those pairs), and by ignoring C from X we mimic the first term in the numerator (ie. neither reward nor penalize a pair that is expected).

Second, the outer normalization by $n - 1$ that averages across all pivots now needs to normalize only across the number of pivots that are not tied with the top item. To clarify, consider the toy example in which the top m items are tied: by (5) there are no pairs

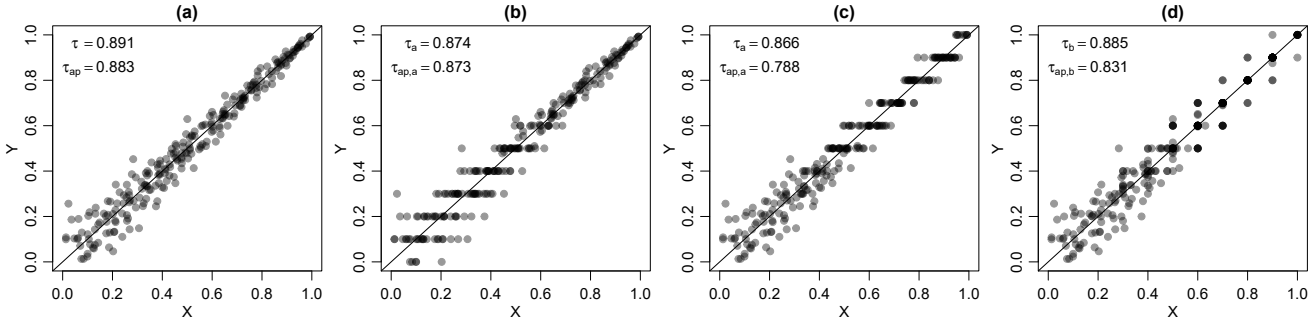

Figure 1: (a) No ties. (b) Small effect of ties at the bottom. (c) Large effect of ties at the top. (d) Ties in both rankings.

to consider concordance above those pivots because $p_i = 1$, so there is no point in counting them for the outer summation and normalization. This (still asymmetric) $\tau_{ap,ties}$ would therefore be

$$\tau_{ap,ties} = \frac{2}{n - t_1} \sum_{i=t_1+1}^{n} \sum_{j<p_i} \frac{c_{ij}}{p_i - 1} - 1. \tag{10}$$

In our example, we find that $\tau_{ap,ties}$ of Y with respect to X is 0.12, while it is 0.16 when correlating X with respect to Y. The final $\tau_{ap,b}$ is defined as the average

$$\tau_{ap,b} = \frac{\tau_{ap,ties}(X,Y) + \tau_{ap,ties}(Y,X)}{2}, \tag{11}$$

which in our case is $\tau_{ap,b} = 0.14$. There are three final remarks worth mentioning. First, note that the summation in (10) is taken over the items in the "estimated" ranking, but the order in which tied elements are arranged does not affect the final score because it only considers pairs above the group. Second, in the absence of ties $\tau_{ap,ties}$ in reduces to the original τ_{ap} because $t_i = 1$ and thus $p_i = i$. Third, because tied groups are disjoint, this formulation generalizes to several tied groups. Also note that in the extreme case that one observer ties all elements, $\tau_{ap,b}$ is not defined because there are no pairs to compare with, as happens with τ_b.

4 CONCLUSIONS AND FUTURE WORK

In this paper we tackled the problem of ties in the calculation of the τ_{ap} correlation coefficient. Following the principles by which Kendall [6] adapted his τ correlation to cope with ties under two different scenarios, we provided closed analytical formulations of τ_{ap} to accept ties in either ranking. Thanks to the accompanying software implementation, researchers can easily substitute τ_{ap} for τ to incorporate its top-heaviness component in problems where ties are possible.

For future work we will consider a third scenario that Kendall [6] mentioned implicitly but did not consider explicitly (see the quote in the introduction). In both τ_a and τ_b he assumed that a tie was given when the observer was unable to discern a difference, but it may be the case that the tied elements are in fact equal in the true ranking. In principle, this is an scenario for the measurement of the accuracy of an observer, so in τ_a this would mean that a tie in the true ranking is what we expect the observer to tell. If the observer orders the pair of items in either way, it should be discordant because he should have tied it. Similarly, if a pair is not tied in the true ranking but the observer did tie it, it should be considered discordant as well.

Yet another scenario to consider is that in which a tie means that the two items are very close together, within some threshold. For instance, two systems may be tied if their *nDCG* difference is smaller than 0.05. While this scenario surely is appealing because it allows us to compute correlations under customizable thresholds, it appears to be problematic because the ties are no longer transitive.

ACKNOWLEDGMENTS

Zizou entrenador. Casemiro titular.

REFERENCES

[1] Ricardo Baeza-Yates, Carlos Castillo, Mauricio Marin, and Andrea Rodriguez. 2005. Crawling a country: better strategies than breadth-first for web page ordering. In *ACM WWW*. 864–872.
[2] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does it Matter?. In *ACM SIGIR*. 667–674.
[3] Ben Carterette, Virgil Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million Query Track 2009 Overview. In *TREC*.
[4] H. E. Daniels. 1944. The Relation between Measures of Correlation in the Universe of Sample Permutations. *Biometrika* 33, 2 (1944), 129–135.
[5] Maurice G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1 (1938), 81–93.
[6] Maurice G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251.
[7] Maurice G. Kendall. 1948. *Rank Correlation Methods* (4th ed.). Charles Griffin & Company Limited.
[8] Karl Pearson. 1907. *On Further Methods of Determining Correlation.* Technical Report.
[9] Tetsuya Sakai. 2007. On the Reliability of Information Retrieval Metrics Based on Graded Relevance. *Information Processing and Management* 43, 2 (2007), 531–548.
[10] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do User Preferences and Evaluation Measures Line Up?. In *ACM SIGIR*. 555–562.
[11] Mark D. Smucker, Gabriella Kazai, and Matthew Lease. 2013. Overview of the TREC 2013 Crowdsourcing Track. In *TREC*.
[12] Student. 1921. An Experimental Determination of the Probable Error of Dr. Spearman's Correlation Coefficients. *Biometrika* 13, 2/3 (1921), 263–282.
[13] Julián Urbano and Mónica Marrero. 2016. Toward Estimating the Rank Correlation between the Test Collection Results and the True System Performance. In *ACM SIGIR*. 1033–1036.
[14] Ellen M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *ACM SIGIR*. 315–323.
[15] Ryen W White, Ian Ruthven, Joemon M Jose, and CJ Van Rijsbergen. 2005. Evaluating implicit feedback models using searcher simulations. *ACM TOIS* 23, 3 (2005), 325–361.
[16] Max A. Woodbury. 1940. Rank Correlation When There are Equal Variates. *Annals of Mathematical Statistics* 11, 3 (1940), 358–362.
[17] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *ACM SIGIR*. 587–594.