Dealing with Ties in Rank Correlation

Priyanka Radja

Technische Universiteit Delft

TU Delft Delft University of Technology

Challenge the future

Dealing with Ties in Rank Correlation

by

Priyanka Radja

in partial fulfillment of the requirements for the degree of

Master of Science
in Computer Science

at the Delft University of Technology,
to be defended publicly on Monday August 27, 2018 at 10:15 AM.

Supervisor:	Dr. J. Urbano	
Thesis committee:	Prof. dr. A. Hanjalic,	TU Delft
	Dr. J. Urbano,	TU Delft
	Dr. J. Van Gemert,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft Delft University of Technology

ABSTRACT

In the field of Information Retrieval (IR), rankings of systems evaluated under different conditions are often compared to each other. This measure of correspondence between rankings, termed as rank correlation, must accurately capture the scenario for which the correlation is computed. Very often, these rankings may have tied systems, for which new correlation coefficients arise. It is important that these coefficients account for the new scenarios in the presence of ties. It is also important that these coefficients provide some flexibility to the person performing the correlation to introduce artificial ties when items are so close to each other that, for practical purposes, they may be considered as tied. Accounting for these different scenarios of tied items in rankings permits performing per topic comparisons of IR systems, which was earlier limited due to the presence of ties on individual topics. Comparing rankings at the topic level, the expected variability of the rankings can be studied to potentially improve the systems on the topics for which they perform poorer than average. We show the application of these new correlation coefficients with two typical IR experiments.

List of Figures

1.1 Tables showing the X and Y rankings for m systems against n topics. 2

3.1 Dimensions of a tie. 12

7.1 Copy of Figure. 1 from W. Webber et al [1]: Predictive power ϕ of nDCG and P@10 of themselves, and nDCG of P@10, with different topic subset sizes, on the TREC 8 runs. 35
7.2 Predictive power of nDCG and P@10 of themselves and each other as a measure of τ_a and τ_{ap} with different topic subset sizes 36
7.3 Predictive power of nDCG and P@10 of themselves and each other as a measure of τ_a, τ_b, τ_e, τ_{ap}, $\tau_{ap,b}$, $\tau_{ap,e}$ with different topic subset sizes 37
7.4 Predictive power of nDCG and P@10 of themselves and each other as a measure of $\tau_a^{0.01}$, $\tau_b^{0.01}$, $\tau_e^{0.01}$, $\tau_{ap,a}^{0.01}$, $\tau_{ap,b}^{0.01}$, $\tau_{ap,e}^{0.01}$ with different topic subset sizes 38
7.5 Predictive power of nDCG and P@10 of themselves and each other as a measure of $\tau_a^{0.05}$, $\tau_b^{0.05}$, $\tau_e^{0.05}$, $\tau_{ap,a}^{0.05}$, $\tau_{ap,b}^{0.05}$, $\tau_{ap,e}^{0.05}$ with different topic subset sizes 39
7.6 Predictive power of nDCG and P@10 of themselves and each other as a measure of $\tau_a^{0.10}$, $\tau_b^{0.10}$, $\tau_e^{0.10}$, $\tau_{ap,a}^{0.10}$, $\tau_{ap,b}^{0.10}$, $\tau_{ap,e}^{0.10}$ with different topic subset sizes 40
7.7 Figure 6 from Yilmaz et al [2] showing TREC-8 mean inferred AP as the judgement set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP. 41
7.8 TREC-8 inferred AP as judgement set is reduced to 30, 10, 5 percent vs. actual AP for topics 47, 22, 6 and 37 42
7.9 Change in a, e variants for inferred AP vs actual AP as the judgement sets are reduced. 43
7.10 Change in a, e variants with threshold $w = 0.01$ for inferred AP vs actual AP as the judgement sets are reduced. 44
7.11 Change in a, e variants with threshold $w = 0.02$ for inferred AP vs actual AP as the judgement sets are reduced. 45
7.12 Change in a, e variants with threshold $w = 0.05$ for inferred AP vs actual AP as the judgement sets are reduced. 46
7.13 Change in a, e variants with threshold $w = 0.10$ for inferred AP vs actual AP as the judgement sets are reduced. 47

List of Tables

3.1 Different Variants of τ and τ_{ap} in the presence of ties. 12

6.1 Different Variants of τ and τ_{ap} in the presence of threshold ties. 23

7.1 Copy of Table. 1 from W. Webber et al [1]: Predictive power ϕ of different metrics on the top 75% of TREC 8 AdHoc Track systems, calculated from 2,000 random repartitionings of the topic set. . 32

7.2 Predictive power of different metrics as a measure of τ, τ_a, τ_b, τ_e, τ_{ap}, $\tau_{ap,a}$, $\tau_{ap,b}$, $\tau_{ap,e}$ 33

7.3 Predictive power of different metrics as a measure of $\tau_a^{0.01}$, $\tau_b^{0.01}$, $\tau_e^{0.01}$, $\tau_{ap,a}^{0.01}$, $\tau_{ap,b}^{0.01}$, $\tau_{ap,e}^{0.01}$ 33

7.4 Predictive power of different metrics as a measure of $\tau_a^{0.05}$, $\tau_b^{0.05}$, $\tau_e^{0.05}$, $\tau_{ap,a}^{0.05}$, $\tau_{ap,b}^{0.05}$, $\tau_{ap,e}^{0.05}$ 34

7.5 Predictive power of different metrics as a measure of $\tau_a^{0.10}$, $\tau_b^{0.10}$, $\tau_e^{0.10}$, $\tau_{ap,a}^{0.10}$, $\tau_{ap,b}^{0.10}$, $\tau_{ap,e}^{0.10}$ 34

CONTENTS

List of Figures v

List of Tables vii

1 Introduction 1
- 1.1 Problem Description . . . 1
- 1.2 Contribution of the Thesis . . . 2
- 1.3 Report Structure . . . 3

2 Previous Work 5
- 2.1 Kendall Rank Correlation Coefficients - τ, τ_a, τ_b . . . 5
  - 2.1.1 Kendall τ . . . 5
  - 2.1.2 Degree of Accuracy - τ_a . . . 6
  - 2.1.3 Degree of Agreement - τ_b . . . 7
- 2.2 AP Correlation Coefficients - $\tau_{ap}, \tau_{ap,a}, \tau_{ap,b}$ . . . 7
  - 2.2.1 AP Correlation τ_{ap} . . . 7
  - 2.2.2 Degree of Accuracy - $\tau_{ap,a}$ . . . 8
  - 2.2.3 Degree of Agreement - $\tau_{ap,b}$ . . . 9

3 Different Scenarios to consider in Rank Correlation Coefficients 11
- 3.1 Where, What and When? . . . 11
- 3.2 Different Variants of τ and τ_{ap} . . . 12

4 Reformulation of AP Correlation Coefficients using sign convention 15
- 4.1 AP Correlation τ_{ap_sign} . . . 15
- 4.2 Degree of Accuracy - τ_{ap,a_sign} . . . 16
- 4.3 Degree of Agreement - τ_{ap,b_sign} . . . 17

5 Equal ties 19
- 5.1 Kendall τ_e . . . 19
- 5.2 AP Correlation $\tau_{ap,e}$ . . . 20

6 Tied within threshold 23
- 6.1 Indiscernible Threshold Ties . . . 24
  - 6.1.1 Degree of Accuracy - τ_a^w . . . 24
  - 6.1.2 Degree of Accuracy - $\tau_{ap,a}^w$ . . . 24
  - 6.1.3 Degree of Agreement - τ_b^w . . . 26
  - 6.1.4 Degree of Agreement - $\tau_{ap,b}^w$ . . . 27
- 6.2 Equal Threshold Ties . . . 28
  - 6.2.1 Kendall τ_e^w . . . 28
  - 6.2.2 AP Correlation $\tau_{ap,e}^w$ . . . 29

7 Practical Assessment 31
- 7.1 Experiment 1 - Correlation Analysis in IR . . . 31
  - 7.1.1 Data and Method . . . 31
  - 7.1.2 Correlation Analysis . . . 32
  - 7.1.3 P@10 Predicted by nDCG . . . 35
- 7.2 Experiment 2 - Topic Variability of IR Systems . . . 36
  - 7.2.1 Data and Method . . . 36
  - 7.2.2 Topic Variability of Systems . . . 41

8 Conclusion **49**

Bibliography **51**

1

INTRODUCTION

1.1. PROBLEM DESCRIPTION

Rankings or ranked lists are extensively used in all fields, not restrictive to computer science, ranging from psychology to economics. A ranking is defined as an arrangement of items in the order of some quality they possess to a varying degree [3]. Therefore, a ranking could be a list of your favourite fruits, ranked from your most favourite to your least favourite or a list of students in a class, arranged according to their score in mathematics.

In [4], Voorhees investigated the stability of Information Retrieval (IR) system rankings obtained from TREC[1] and showed that order of the systems in the rankings is more important than absolute values. Following which, a great deal of research has been carried out in the field of IR for rankings of different groups of items like that of web pages by PageRank algorithm [5], system runs using test collections[4], topic difficulty by predictive measures [6], term ranking for query expansion [7], etc. Moreover, IR system rankings for different evaluation measures [8], topic sets [9], user ratings [10], rankings by experts vs non-experts [11], may be compared to identify system performances and to assess how well the rankings correspond to each other. This comparison, termed as rank correlation, between rankings of the same items, indicates the degree of correspondence between them and will be the focus of this research. Rank correlation helps in assessing the two alternatives, represented by the two rankings, for varied purposes like simulation of implicit user feedback [12], resource selection in distributed IR systems [13] or simply to measure the retrieval effectiveness [14], [15] [16], [17] of different IR systems.

To determine correlation between two variables, the Pearson correlation coefficient r [18] was developed based on many properties described by Karl Pearson, and is measured as the distance from a best fit line i.e. as the strength of the linear association between the two variables. To determine the correlation between two rankings, rank correlation coefficients are used. Some popular ones being the Spearman correlation coefficient ρ, Kendall τ and AP correlation coefficient τ_{ap}. The Spearman correlation coefficient ρ, is measured as the strength and direction of the monotonic relationship between the rankings. Since both the Spearman ρ and Pearson r coefficients measure the distance between items, the measure of how far the items are from each other affects the coefficients. The Kendall τ measures rank correlation as the pairwise concordance of all item pairs. Similar to τ, the AP correlation coefficient τ_{ap}, given by Yilmaz et al, is measured as the pairwise concordance of only the items above each item in the rankings, thereby introducing a top heaviness by which discordant item pairs at the top of the rankings are penalized more.

In IR, rank correlation analysis is the process of assessing the mean system performances over all topics[2]. For example, as shown in Figure. 1.1, the rankings X and Y of systems $(S_1, S_2, ..., S_m)$, against different topics $(T_1, T_2, ..., T_n)$, may differ in the evaluation measures[3] used like Recall, P@10, nDCG etc., or in general, the evaluation conditions used, while making the rankings. The column-wise average of these rankings, which provides the Rankings of Systems (RoS) over all topics, are compared to determine their correspondence for

[1] Text REtrieval Conference consists of different tracks, each with the necessary infrastructure (test collections, evaluation methodology, etc. to carry out large-scale evaluation of text retrieval methodologies

[2] Topics are the queries for which the different IR system runs are automatically generated, given the expected output and a standard set of judged results

[3] Evaluation measures assess how well the retrieved results satisfy the query's intent

X

	S_1	S_2	...	S_m
T_1	X_{11}	X_{12}	...	X_{1m}
T_2	X_{21}	X_{22}	...	X_{2m}
...	X_{*1}	X_{*2}	...	X_{*m}
T_n	X_{n1}	X_{n2}	...	X_{nm}
RoS	$\overline{X}_1$	$\overline{X}_2$	...	$\overline{X}_m$

Y

	S_1	S_2	...	S_m
T_1	Y_{11}	Y_{12}	...	Y_{1m}
T_2	Y_{21}	Y_{22}	...	Y_{2m}
...	Y_{*1}	Y_{*2}	...	Y_{*m}
T_n	Y_{n1}	Y_{n2}	...	Y_{nm}
RoS	$\overline{Y}_1$	$\overline{Y}_2$	...	$\overline{Y}_m$

Figure 1.1: Tables showing the X and Y rankings for m systems against n topics.

rank correlation analysis. If X and Y were system rankings made using different evaluation measures, rank correlation analysis provides sufficient details to check if the two evaluation measures capture different aspects of the systems, reflect different user models and if an evaluation measure is well motivated or not [19]. This method of comparing the averaged system performances (RoS), has become the standard for rank correlation analysis [19]. However, this method masks the system performances over individual topics, thereby hiding the topic-to-topic level variation of the systems [4].

Ties in rankings refer to items holding the same value for the variable based on which the ranking is made. For instance, in a ranking of students in a class, arranged based on their score in mathematics, more than one student can have the same grade. These students are, hence, said to be tied. Such ties can also occur in system rankings for a topic in IR. This means that more than one system may posses the same evaluation measure value for a topic denoted by the rows in Figure. 1.1. This is especially true for an evaluation measure like P@10 where the systems can only take 11 possible values from 0, 0.1, 0.2 up to 1.0 as it is the precision at 10 documents retrieved. Therefore, a major reason for the standard use of averaged RoS for rank correlation could be that on a topic-to-topic level, the systems can be tied with each other. The generic form of the popular rank correlation coefficients - Kendall τ and AP correlation coefficient τ_{ap} cannot account for tied items in the rankings.

In previous research conducted by Woodbury [20] and Student [21], the treatment of ties in rankings was addressed for Spearman correlation coefficient ρ. Following these research, Kendall had developed two variants of his coefficient τ - τ_a and τ_b which were initially named τ_w and τ_s respectively, to denote the reference from Woodbury and Student. In [22], J. Urbano and M. Marrero developed two variants of the AP correlation coefficient τ_{ap} - $\tau_{ap,a}$ and $\tau_{ap,b}$, similar to τ, to handle ties in rankings.

The ties addressed, so far, only referred to items that were indiscernible to the ranker (hereafter, referred as indiscernible ties). This refers to the situation when the ranker, in a state of indecision between items, ties them. In [22], the authors proposed as future work, two scenarios of tied items - items tied due to their equivalence (equal ties) and items tied due to a small difference in their values that they lie within a customizable threshold (threshold ties). The proposal for items tied as they are equivalent, accounts for strictly tied items by the rankers. The proposal for a threshold value, to determine whether items are tied or not, provides flexibility to the person performing the rank correlation, to determine how close items need to be, in order to be tied. A similar choice of threshold was also suggested in [23] in automatic indexing for analysing documents and manipulating their descriptions in searching to generate index language used for these purposes and in [24] to check the probabilities for which the users, given a DCG score with a small difference in value, find a system satisfactory over another. This is helpful in practical applications, where it is appealing to allow some negligible differences between item values, while considering them as tied. These two cases of equal ties and threshold ties proposed in [22] will be the focus of this thesis.

1.2. CONTRIBUTION OF THE THESIS

The main contributions of the thesis are as follows:

- Reformulation of the AP correlation coefficient τ_{ap}, $\tau_{ap,a}$ and $\tau_{ap,b}$ based on the sign convention followed by Kendall τ.

  Due to the use of the sign convention, ties are allowed in both the rankings while computing τ_a. In $\tau_{ap,a}$, on the other hand, the correlation cannot be computed with ties in the reference ranking. To allow ties in reference ranking while computing $\tau_{ap,a}$ and to simplify the current formulations of τ_{ap}, $\tau_{ap,a}$, $\tau_{ap,b}$, it is necessary to reformulate the coefficients using the sign convention adopted by Kendall.

- Formulation of a new variant for Kendall τ and the AP correlation coefficient τ_{ap} to handle equal ties.

  For the case where ties refer to items that are truly equal to each other, a ranker must be awarded or penalized in the rank correlation coefficient for correctly identifying an equal tie and for incorrectly tying untied items with respect to the other ranking. Therefore, a new variant for the τ and τ_{ap} is necessary to be formulated to handle this special case of equal ties.

- Study the effect of a threshold, to determine whether items are tied or not, for the Kendall τ and AP τ_{ap} correlation coefficient.

  As suggested in [22], the existence of a customizable threshold in establishing items as tied in rankings has practical implications. This necessitates the reformulation of the coefficients τ and τ_{ap} to account for threshold ties.

- Practical assessment

  Evaluate the existing and the newly formulated variants of the τ and τ_{ap} coefficients for equal and threshold ties by performing experiments similar to [1].

  To understand the topic-to-topic level variability of IR systems and to justify the necessity in carrying out rank correlation of IR systems on per topic level, perform experiment similar to [2].

All the formulations and experiments carried out in this thesis are implemented in R and will be made available publicly as part of the ircor[4] package for rank correlation analysis in IR .

1.3. REPORT STRUCTURE

The remainder of this report is organized as follows, with the explanation of the different, existing variants of the two important correlation coefficients τ and τ_{ap} in Chapter 2. A decision tree highlighting the different possible scenarios that rank correlation coefficients should consider in Chapter 3. This is followed by Chapter 4, where the first goal of reformulating the τ_{ap} coefficients, in terms of the sign convention used in Kendall τ, will be detailed. Chapter 5 will address the second goal of handling equal ties in rankings while computing rank correlation. Chapter 6 will address the third goal of handling items, which fall very close to each other i.e. within a threshold in the evaluation measure, that they are tied while computing the rank correlation. Finally, the practical assessment, given as the final goal in section .1.2, will be detailed in Chapter 7, followed by conclusion and discussion in Chapter 8.

[4]The ircor package can be found at https://github.com/julian-urbano/ircor

2

PREVIOUS WORK

All previous work with regards to the correlation coefficients, τ and τ_{ap} will be discussed in this chapter.

2.1. KENDALL RANK CORRELATION COEFFICIENTS - τ, τ_a, τ_b

The rank correlation coefficient τ along with the two variants - τ_a and τ_b developed by Kendall, following the research by Woodbury [20] and Student [21], are discussed below in their respective subsections.

2.1.1. KENDALL τ

In 1938, M. Kendall proposed his rank correlation coefficient τ in the paper [25]. The τ computes the correlation between two rankings with no ties as the number of pairwise adjacent swaps required to convert one ranking into the other. This is given by:

$$\tau = \frac{\#concordant - \#discordant}{\#total} = \frac{S}{n(n-1)/2} = \frac{2}{n(n-1)/2}\sum_{i<j} c_{ij} - 1 \tag{2.1}$$

where,

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) = +1 (or) -1 \\ 0 & otherwise \end{cases}$$

Given two rankings X and Y of size n, for every possible item pairs, Kendall τ calculates the correlation by identifying the total number of concordant and discordant pairs. The item pairs that appear in the same order in the two rankings and in opposite order are termed concordant pairs and discordant pairs respectively. For rankings of size n, the total number of pairs possible is $\binom{n}{2}$ which equates to $n(n-1)/2$. As given in the formula, $\sum_{i<j} c_{ij}$ measures the total concordant pairs by awarding a 1 only for item pairs (i,j) that appear in the same order i.e. $sign(x_j - x_i) = sign(y_j - y_i)$ and a 0 otherwise. Hence, by general probability theory, the total correspondence (S) between rankings measured as total concordant pairs (P) minus total discordant pairs (Q) can be reformulated as

$$S = P - Q = P - (1 - P) = 2P - 1 = 2\sum_{i<j} c_{ij} - 1$$

Using the sign convention in the definition of c_{ij}, the τ can also be formulated as:

$$\tau = \sum_{i<j} \frac{sign(x_j - x_i) \cdot sign(y_j - y_i)}{n(n-1)/2} \tag{2.2}$$

where,

$$sign(x_j - x_i) \cdot sign(y_j - y_i) = \begin{cases} +1 & concordant\ pair \\ -1 & discordant\ pair \end{cases}$$

Due to the normalization by the total number of item pairs in the rankings, the Kendall τ can have a value on a scale of +1 to -1. The τ value of +1, -1 and 0 denote that the rankings are identical to each other (perfect concordance), exactly reversed in their orders of items (perfect discordance) and independent of each other respectively. Examples of these three cases are given below:

Perfectly Concordant Rankings	Perfectly Discordant Rankings	Independent Rankings
X Y	X Y	X Y
1A 1A	1A 1D	1A 1B
2B 2B	2B 2C	2B 2D
3C 3C	3C 3B	3C 3A
4D 4D	4D 4A	4D 4C
$\tau = \left(\frac{2}{6} \cdot 6\right) - 1 = +1$	$\tau = \left(\frac{2}{6} \cdot 0\right) - 1 = -1$	$\tau = \left(\frac{2}{6} \cdot 3\right) - 1 = 0$

2.1.2. DEGREE OF ACCURACY - τ_a

Following the research by Woodbury [20], the τ_a was framed by M. Kendall. Given two rankings of size n, a true ranking X and a ranking by an observer Y, τ_a acts as a measure of the degree of accuracy of observer Y in his/her ranking to the true ranking X. This coefficient was initially named as τ_w in the paper [26] and was later renamed to τ_a in Kendall's book "Rank Correlation Methods" [3].

The τ_a is given as:

$$\tau_a = \sum_{i<j} \frac{sign(x_j - x_i) \cdot sign(y_j - y_i)}{n(n-1)/2} \tag{2.3}$$

where,

$$sign(x_j - x_i) \cdot sign(y_j - y_i) = \begin{cases} +1 & concordant\ pair \\ -1 & discordant\ pair \\ 0 & tied\ pair \end{cases}$$

In the presence of ties in the rankings, where a tie refers to items that the ranker was unable to tell apart (indiscernible ties), the τ_a in equation (2.3) must be equal to the average of the τ's obtained, by replacing the tied item groups with integral ranks for all possible permutations. This is handled by awarding a 0 for tied item pairs in τ_a as they appear in concordant order in half the permutations and in discordant order in the other half.

As shown in table below, for a tied item group (B, C, D) in Y where the tied items are represented by ranks equal to the average of the integral position numbers they extend over, the item pair (B, C) appears in the concordant order of B followed by C only in half the permutations (Y_1, Y_2, Y_5). The τ_a using equation (2.3) is 0.7 and the average of the possible τ's ($\tau_{X,Y_1} = 1.0$, $\tau_{X,Y_2} = 0.8$, $\tau_{X,Y_3} = 0.8$, $\tau_{X,Y_4} = 0.6$, $\tau_{X,Y_5} = 0.6$, $\tau_{X,Y_6} = 0.4$) using equation (2.1) is also equal to 0.7.

X	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1A	1A	1 A	1 A	1 A	1 A	1 A	1 A
2B	3B	2 B	2 B	2 C	2 C	2 D	2 D
3C	3C	3 C	3 D	3 B	3 D	3 B	3 C
4D	3D	4 D	4 C	4 D	4 B	4 C	4 B
5E	5E	5 E	5 E	5 E	5 E	5 E	5 E

In a similar setting with ties in both rankings, the τ_a is again equivalent to the average of the possible τ's.

X	X_1	X_2	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1 A	1 A	1 A	1 A	1 A	1 A	1 A	1 A	1 A	1 A
2 B	2 B	2 B	3 B	2 B	2 B	2 C	2 C	2 D	2 D
3 C	3 C	3 C	3 C	3 C	3 D	3 B	3 D	3 B	3 C
4.5 D	4 D	4 E	3 D	4 D	4 C	4 D	4 B	4 C	4 B
4.5 E	5 E	5 D	5 E	5 E	5 E	5 E	5 E	5 E	5 E

The τ_a for the example above with ties in both rankings is 0.6 and the average of the τ's ($\tau_{X_1,Y_1} = 1.0$, $\tau_{X_1,Y_2} = 0.8$, $\tau_{X_1,Y_3} = 0.8$, $\tau_{X_1,Y_4} = 0.6$, $\tau_{X_1,Y_5} = 0.6$, $\tau_{X_1,Y_6} = 0.4$, $\tau_{X_2,Y_1} = 0.8$, $\tau_{X_2,Y_2} = 0.6$, $\tau_{X_2,Y_3} = 0.6$, $\tau_{X_2,Y_4} = 0.4$, $\tau_{X_2,Y_5} = 0.4$, $\tau_{X_2,Y_6} = 0.2$) is also equal to 0.6.

When neither of the rankings have ties, the Kendall τ_a is equal to τ. This satisfies the Generalization axiom given in [15], by which *"Notwithstanding the richness criteria, any proposed metric should collapse to a natural metric in cases where the richer criteria do not play a role"*. When all items are tied in either or both the rankings, a 0 τ_a will be derived, meaning that the two rankings are independent of each other. This

is because the tied items in the ranking, which in this case is the entire ranked list, are indiscernible to the ranker. Therefore, no conclusion can be arrived at regarding the ranker's judgement for all items, leading to a 0 correlation.

2.1.3. DEGREE OF AGREEMENT - τ_b

For two rankings by observers X and Y, the τ_b was framed by Kendall to measure the degree of agreement between them, following the research by Student [21]. The coefficient was initially named τ_s in the paper [26] and was later renamed to τ_b in Kendall's book "Rank Correlation Methods" [3].

In the presence of items, tied as they were indiscernible to the ranker, the τ_b awards a neutral 0 for the tied item pair similar to τ_a. In addition to this, the tied item pairs are not expected in the denominator. This is because while computing τ_b, an observer should not be expected to be indiscernible about the items, the other observer tied.

The τ_b given by M. Kendall is as follows:

$$\tau_b = \sum_{i<j} \frac{sign(x_j - x_i) \cdot sign(y_j - y_i)}{\sqrt{n(n-1)/2 - t_X}\sqrt{n(n-1)/2 - t_Y}} \tag{2.4}$$

where,

$$sign(x_j - x_i) \cdot sign(y_j - y_i) = \begin{cases} +1 & concordant\ pair \\ -1 & discordant\ pair \\ 0 & tied\ pair \end{cases}$$

t_X = #tied item pairs in the ranking by observer X
t_Y = #tied item pairs in the ranking by observer Y

For the example below, the items B and C are tied as indiscernible in X and items B, C and D are tied as indiscernible in Y. The tied item pair (B,C) in X ranking is not expected to be concordant in Y ($t_X = 1$) and the 3 tied item pairs {(B,C), (B,D), (C,D)} in Y ranking are not expected to be concordant in X ($t_Y = 3$). The τ_b calculated for this example using equation (2.4) is 0.8819.

X	Y
1 A	1 A
2.5 B	3 B
2.5 C	3 C
4 D	3 D
5 E	5 E

Therefore, to calculate the degree of agreement between observers, it is agreeable to eliminate the tied groups of items which the observers were unable to tell apart from the total number of item pairs, the two observers are expected to be concordant in. In the absence of ties, τ_b simplifies into τ. This satisfies the Generalization axiom given in [15]. When all the items are tied in either or both the rankings, the τ_b is undefined as no item pairs are expected to be concordant between the two rankings, leading to a zero in the denominator of τ_b.

2.2. AP CORRELATION COEFFICIENTS - τ_{ap}, $\tau_{ap,a}$, $\tau_{ap,b}$

The AP correlation coefficient - τ_{ap} given by Yilmaz et al [27] and the two variants- $\tau_{ap,a}$ and $\tau_{ap,b}$ developed by J. Urbano and M. Marrero [22] will be discussed in this section.

2.2.1. AP CORRELATION τ_{ap}

The Kendall τ, τ_a and τ_b treat the item pairs appearing in any part of the rankings equally, meaning that an incorrectly ordered item pair in the top, middle or bottom of rankings are penalized similarly. However, the top of a ranking is more important to be ranked in the correct sequence in many practical applications. E. Yilmaz et al [27] and B. Carterette [28] have criticized that this makes the Kendall τ, τ_a and τ_b a poor correlation coefficient.

In order to introduce a top heaviness, E. Yilmaz et al [27] formulated a new coefficient based on average precision called the AP (Average Precision) correlation coefficient, τ_{ap}. The τ_{ap} takes one ranking as reference and checks how well the other ranking corresponds to it, by measuring the concordance of only the items

above each item. This introduces top heaviness by ignoring the influence of the items below. The τ_{ap} does not account for ties and is calculated as:

$$\tau_{ap} = \frac{2}{n-1}\sum_{i=2}^{n}\frac{\#\text{concordant above i}}{i-1} - 1 = \frac{2}{n-1}\sum_{i=2}^{n}\sum_{j<i}\frac{c_{ij}}{i-1} - 1 \tag{2.5}$$

where,

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) = +1 (or) -1 \\ 0 & otherwise \end{cases}$$

Given two rankings X and Y of size n, for every item i acting as a pivot from position 2 to n, the τ_{ap} checks the concordance of only the i-1 items above it, for the n-1 possible pivots.

For the example below with no ties, the correlation between the two rankings with top heaviness was calculated as $\tau_{ap} = \frac{2}{4-1}\left(\frac{0}{1} + \frac{1}{2} + \frac{3}{3}\right) - 1 = 0$, using the equation (2.5).

X	Y
1 A	1 B
2 B	2 C
3 C	3 A
4 D	4 D

Since most of the discordant pairs {(A,B), (A,C)} are on the top of the rankings, the τ_{ap} penalizes these discordances more. In the next example, even though the rankings have the same number of discordant pairs as the previous example, they are penalized lesser with a $\tau_{ap} = \frac{2}{4-1}\left(\frac{1}{1} + \frac{1}{2} + \frac{2}{3}\right) - 1 = 0.4444$, as the discordant pairs {(B,C), (B,D)} appear lower in the rankings.

X	Y
1 A	1 A
2 B	2 C
3 C	3 D
4 D	4 B

2.2.2. DEGREE OF ACCURACY - $\tau_{ap,a}$

For a true ranking X acting as reference and a ranking provided by an observer Y, J. Urbano and M. Marrero in [22] had formulated $\tau_{ap,a}$, similar to τ_a, to measure the degree of accuracy of the observer's ranking Y to X. The $\tau_{ap,a}$ was formulated to deal with ties only in the observer's ranking Y, where ties referred to items the observer was unable to tell apart (indiscernible ties).

In the presence of indiscernible ties in Y, when the pivot i while computing $\tau_{ap,a}$ is a non-tied item, the formulation is similar to τ_{ap}. This is because even when there exits a tied group above a non-tied pivot, the entire tied group is either concordant with the pivot or discordant.

However, when the pivot is a tied item in Y, the contribution of all items above the tied group, of which the pivot is an element, will change according to the position of the pivot within the tied group. Moreover, within the tied group of the pivot, the other tied items are in concordance with the pivot only in certain permutations. Therefore, when the pivot is a tied item, it is necessary to consider the contribution of all items above the tied group, based on the position of the pivot within the tied group and the contribution of all items within the tied group. These two cases are handled separately by two terms in $\tau_{ap,a}$.

$$\text{Term 1: Contribution of items above the tied group} = \sum_{i=t_1+1}^{n}\sum_{j<p_i} c_{ij}\sum_{k=1}^{t_i}\frac{1}{t_i(p_i+k-2)} \tag{2.6}$$

where,

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) = +1 (or) -1 \\ 0 & otherwise \end{cases}$$

p_i = position of the 1^{st} element in the tied group of which pivot i is an element
t_i = size of the tied group of which pivot i is an element

The $\sum_{k=1}^{t_i} \frac{1}{(p_i+k-2)}$ measures the number of expected items to be concordant with the pivot, for all the positions the pivot i can take within its tied group, as the summation of all possible $(i-1)$. The $\sum_{j<p_i} c_{ij}$ in Term 1 measures all items above the pivot's tied group that are concordant with the pivot in the two rankings. Due to the summation $\sum_{i=t_1+1}^{n}$ in Term 1, which runs over all items in the rankings except the first item and its ties, calculated as t_1+1 (where t_1 is the size of first item's tied group), the summation of the different possible $(i-1)$'s is calculated t_i times for each of the items, in a tied group of size t_i, acting as pivot. This explains the division by t_i in Term 1.

$$\text{Term 2 : Contribution of items within the tied group} = \sum_{i=1}^{n} \frac{1}{2} \sum_{k=1}^{t_i-1} \frac{k}{t_i(p_i+k-1)} \tag{2.7}$$

The tied pivot and an item within its tied group in Y, will be concordant with X only in half of the possible permutations. Following the example below, the tied pairs {(B,C), (B,D), (C,D)} in Y, appear in concordant order with X only in half of the six permutations possible. Therefore, the Term 2 is divided by 2 to account for only half of the permutations in which a tied pair is concordant, with the correct $(i-1)$ for normalization calculated by $\sum_{k=1}^{t_i-1} \frac{1}{(p_i+k-1)}$ and the k in the numerator, restricting the permutations possible by only allowing the non-pivot item to occupy places above the pivot but within the tied group. The presence of t_i in the denominator of Term 2 follows the same explanation as for Term 1.

X	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1A	1A	1 A	1 A	1 A	1 A	1 A	1 A
2B	3B	2 B	2 B	2 C	2 C	2 D	2 D
3C	3C	3 C	3 D	3 B	3 D	3 B	3 C
4D	3D	4 D	4 C	4 D	4 B	4 C	4 B

Combining the two terms, the $\tau_{ap,a}$ is given as:

$$\tau_{ap,a} = \frac{2}{n-1}\left(\sum_{i=t_1+1}^{n}\sum_{j<p_i} c_{ij}\sum_{k=1}^{t_i}\frac{1}{t_i(p_i+k-2)} + \sum_{i=1}^{n}\frac{1}{2}\sum_{k=1}^{t_i-1}\frac{k}{t_i(p_i+k-1)}\right) - 1 \tag{2.8}$$

where,

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) = +1 (or) -1 \\ 0 & otherwise \end{cases}$$

p_i = position of the 1^{st} element in the tied group of which pivot i is an element

t_i = size of the tied group of which pivot i is an element

For the example above, the $\tau_{ap,a}$, calculated using equation (2.8) is 0.6111 and is equal to the average of the τ_{ap}'s possible, for each permutation of Y correlated with X ($\tau_{ap}(X,Y_1) = 1.0$, $\tau_{ap}(X,Y_2) = 0.7778$, $\tau_{ap}(X,Y_3) = 0.6667$, $\tau_{ap}(X,Y_4) = 0.4444$, $\tau_{ap}(X,Y_5) = 0.5556$, $\tau_{ap}(X,Y_6) = 0.2222$), similar to the Kendall τ_a coefficient.

For cases when there are ties in the true ranking X, the coefficient $\tau_{ap,a}$ becomes invalid and cannot be applied. During practical applications, even the true ranking X can have indiscernible ties. The Kendall τ_a accounted for ties in both rankings due to the use of the sign convention. Therefore, it is necessary to reformulate the $\tau_{ap,a}$ in terms of the sign convention, followed by Kendall.

It is important to note when there are no ties in X and Y, the $\tau_{ap,a}$ becomes equivalent to τ_{ap}. When all items in ranking Y are tied, the $\tau_{ap,a}$ becomes zero, indicating that no conclusion can be derived since the observer was unable to tell apart any of the items in the ranking.

2.2.3. DEGREE OF AGREEMENT - $\tau_{ap,b}$

For two rankings X and Y, both given by observers, the degree of agreement between them with top heaviness is calculated using $\tau_{ap,b}$, formulated by J. Urbano and M. Marrero [22]. In case of ties in the rankings, neither observer is expected to tie the items the other observer tied. This follows the same ideology adopted by M. Kendall in his coefficient τ_b . Therefore, tied groups are not expected to be concordant between the rankings while computing τ_b.

Since the τ_{ap} coefficient calculates correlation with one ranking as reference ranking, a symmetric version of $\tau_{ap,b}$ was formulated by J. Urbano and M. Marrero in [22], following the suggestion by Yilmaz et al [27]. This symmetric version of $\tau_{ap,b}$ requires averaging the two cases when each of the two rankings are held as

reference and the other ranking's correspondence to the reference is measured. The $\tau_{ap,b}$ is calculated as follows:

$$\tau_{ap,b} = \frac{\tau_{ap,ties}(X,Y) + \tau_{ap,ties}(Y,X)}{2} \tag{2.9}$$

where,

$$\tau_{ap,ties} = \frac{2}{n-t_1} \sum_{i=t_1+1}^{n} \sum_{j<p_i} \frac{c_{ij}}{p_i - 1} - 1 \tag{2.10}$$

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) = +1 (or) -1 \\ 0 & otherwise \end{cases}$$

p_i = position of the 1^{st} element in a tied group of which pivot i is an element
t_i = size of the tied group of which pivot i is an element

When the pivot is a tied item, only the items above the pivot's tied group are checked for concordance i.e. only the items above the first item p_i in the pivot's tied group, are checked for concordance. This explains the $p_i - 1$ in the denominator instead of the $i - 1$. This eliminates the expectation of tied pairs to be concordant in the rankings.

If the first item is tied, all the items tied with it can be ignored as pivots, while traversing through the rankings as there are no items above the first item's tied group that can form concordant pairs. This explains the $n - t_1$ in the denominator and the range of the summation $\sum_{i=t_1+1}^{n}$, where t_1 is the size of the first item's tied group.

X	Y
1 A	1 A
2.5 B	3 B
2.5 C	3 C
4 D	3 E
5 E	5 D

The $\tau_{ap,b}$, $\tau_{ap,ties}(X,Y)$ and $\tau_{ap,ties}(Y,X)$ for the example above are calculated to be 0.8333, 1.0 and 0.6667 respectively.

The discordant pair (D,E) is penalized in both $\tau_{ap,ties}(X,Y)$ and $\tau_{ap,ties}(Y,X)$. The tied pairs (B,E) and (C,E) are penalized only once as they are expected only when X is the reference ranking. This penalization of tied pairs removes the effect of its expectation in Y when X is the reference. Since (B,C) is tied in both rankings, this pair is not expected when either X or Y are held as reference.

When there are no ties in X and Y, $\tau_{ap,ties}(X,Y)$ becomes equal to $\tau_{ap}(X,Y)$, $\tau_{ap,ties}(Y,X)$ to $\tau_{ap}(Y,X)$ and $\tau_{ap,b}$ to the symmetric average of τ_{ap} i.e. $\frac{\tau_{ap}(X,Y)+\tau_{ap}(Y,X)}{2}$. When all items are tied in either or both of the rankings, the $\tau_{ap,ties}$ and therefore $\tau_{ap,b}$ is undefined as no item pairs are expected to be concordant between the two rankings. This is caused by the $n - t_1$ in the denominator which becomes 0 as all items are tied with the first item i.e $t_1 = n$.

3

Different Scenarios to consider in Rank Correlation Coefficients

The different scenarios that are possible while estimating the correlation between two rankings and the different implementations of the rank correlation coefficients for each of these scenarios are discussed in this chapter.

3.1. Where, What and When?

In this section, where a tie occurs, when items are said to be tied and what a tie means are some questions that will be investigated.

- Where does a tie occur?

  The two rankings, whose correlation is to be found, can either be true ranking vs. observer ranking or observer ranking vs. observer ranking. In the former case, the correlation score acts as a measure of the degree of accuracy of the observer in his/ her ranking to that of the true ranking. In the latter case, the correlation score is a measure of the degree of agreement between the two observers. These two cases were denoted as the a and b variant respectively, for τ and τ_{ap}. Therefore, there could be two scenarios where a tied group can occur. One scenario where one ranking is the true ranking and the other is a ranking by an observer. Other scenario where both rankings are by observers. This answers the question of where the items are tied.

- What does a tie mean?

  The next important question of what a tie means can have two answers: the items are equal or indiscernible to the ranker. The items which are equivalent to each other, that they are tied at the same rank are called equal ties. We will denote the rank correlation coefficients τ and τ_{ap} accounting for equal ties with “e" in subscript. The items, which the ranker is unable to tell apart, are called indiscernible ties.

- When are items said to be tied?

  The final important question of when items are tied can also have two answers: the tied items have the same value or the tied items have values, which are within a threshold set by the person performing the rank correlation. The value here denotes the quantity based on which the rankings were made. For many practical applications, a third person different from the rankers, who performs the rank correlation may, for his/ her requirements, need some flexibility in considering items very close to each other as ties. Therefore, this third person can set a customizable threshold value which will determine whether items are tied. For example, if this third person is accommodating of differences in item values of 0.5, the threshold shall be set to 0.5 and all items with a maximum difference of 0.5 will be tied. Such ties are termed as threshold ties. We will denote the rank correlation coefficients accounting for threshold ties with “w" in superscript. Moreover, when the threshold value is 0, the items are only when they have the same value. Therefore, the when? branch in Figure. 3.1 condenses into only the “same value" branch for threshold equal to 0.

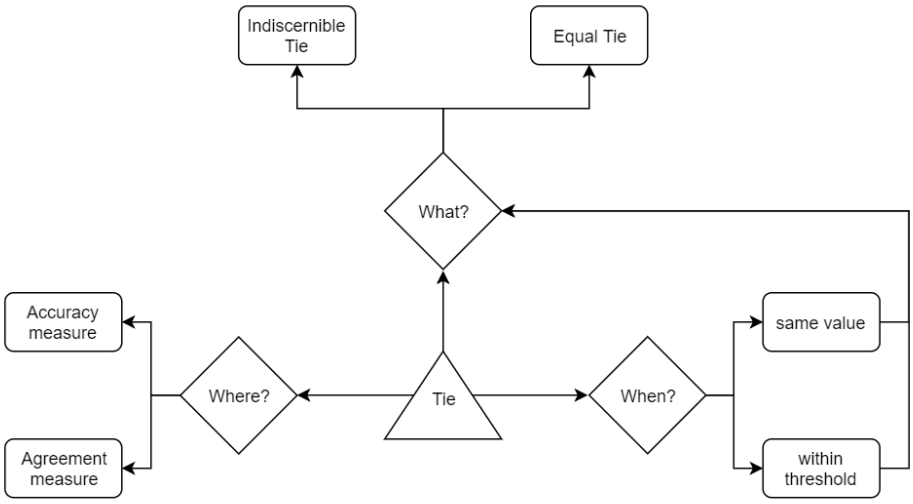

Figure 3.1: Dimensions of a tie.

Irrespective of whether a tie occurred because the tied items have exactly the same value or have values within the threshold set by the third party performing the rank correlation, the ties could have two meanings given by the previous question i.e. what a tie means - equal and indiscernible.

Based on these three questions, three dimensions of rank correlation coefficient arises. This is given in Figure. 3.1.

3.2. DIFFERENT VARIANTS OF τ AND τ_{ap}

The three important questions (where, what and when) determine the different cases that rank correlation coefficients should consider in the presence of ties.

Table 3.1: Different Variants of τ and τ_{ap} in the presence of ties.

No Ties	
Where?	
Truth vs. Observer	Observer vs. Observer
τ (Kendall, 1938) or τ_{ap} (Yilmaz et al, 2008)	

Ties			
When?	What?	Where?	
		True Ranking vs. Observer Ranking	Observer Ranking vs. Observer Ranking
Tied items have the same value	Indiscernible ties	τ_a (Kendall, 1945) or $\tau_{ap,a}$ (Urbano & Marrero, 2017)	τ_b (Kendall, 1945) or $\tau_{ap,b}$ (Urbano & Marrero, 2017)
	Equal ties	τ_e or $\tau_{ap,e}$	
Tied items have values within threshold	Indiscernible ties	τ_a^w or $\tau_{ap,a}^w$	τ_b^w or $\tau_{ap,b}^w$
	Equal ties	τ_e^w or $\tau_{ap,e}^w$	

Going through the branches of Figure. 3.1 gives rise to six different cases that a rank correlation coefficient should consider in the presence of ties, which gives rise to six variants for the Kendall and AP Correlation coefficients as given in table 1. Some of these cases have already been addressed in previous research, as referenced in the table. The unreferenced coefficients in Table. 3.1 will be formulated as a part of this thesis.

It is important to note that in the absence of ties, irrespective of the type of rankings (where), the τ and τ_{ap} have only one variant. This variant is the original form of the two coefficients given by M. Kendall [25] and Yilmaz et al [27]. Similarly, equal ties occurring when the items have same value or are within a threshold (when), irrespective of the type of rankings (where), both τ and τ_{ap} have only one variant (explained in detail in Chapter 5).

4

Reformulation of AP Correlation Coefficients using sign convention

The first goal of the thesis involving the reformulation of the AP Correlation coefficients τ_{ap}, $\tau_{ap,a}$ and $\tau_{ap,b}$, using the sign convention followed by M. Kendall, is addressed in this chapter. All the three variants of the AP correlation coefficient $\tau_{ap}, \tau_{ap,a}$ and $\tau_{ap,b}$ will be reformulated using the sign convention followed by Kendall. The successful reformulation of the AP correlation coefficients will simplify each of the coefficients and permit the presence of ties in the reference ranking X when computing $\tau_{ap,a}$. In [15], it is emphasized that simplicity is one of the several criteria that any metric should satisfy as it is this quality that makes a coefficient like Kendall τ prominent, despite its shortcomings. For this reason and for the ease of use and understanding, reformulation of the AP Correlation coefficient is essential.

4.1. AP Correlation τ_{ap_sign}

The reformulation of the Yilmaz τ_{ap} using the sign convention is detailed in this section. For two rankings X and Y, in order to introduce top heaviness, the correlation between them is found by holding X as reference and calculating the number of concordant items above each item in the Y.

The AP Correlation coefficient τ_{ap} given by Yilmaz et al in [27] is:

$$\tau_{ap} = \frac{2}{n-1}\sum_{i=2}^{n}\frac{\#\text{concordant above i}}{i-1} - 1 = \frac{2}{n-1}\sum_{i=2}^{n}\sum_{j<i}\frac{c_{ij}}{i-1} - 1 \tag{4.1}$$

where,

$$c_{ij} = \begin{cases} 1 & sign(x_j - x_i) = sign(y_j - y_i) = +1 (or) -1 \\ 0 & otherwise \end{cases}$$

Using the sign convention, we can reformulate τ_{ap} as:

$$\tau_{ap\_sign} = \frac{1}{n-1}\sum_{i=2}^{n}\sum_{j<i}\frac{sign(x_i - x_j)\cdot sign(y_i - y_j)}{i-1} \tag{4.2}$$

where,

$$sign(x_i - x_j)\cdot sign(y_i - y_j) = \begin{cases} +1 & concordant\ pair \\ -1 & discordant\ pair \end{cases}$$

By this new definition of τ_{ap_sign}, the top heaviness is still in effect by the unchanged denominator. The $2\sum_{j<i} c_{ij} - 1$ in the numerator of τ_{ap} measures the total concordance between the two rankings as the total number of concordant pairs minus the total number of discordant pairs, explained in section 2.2.1. This can be replaced by $sign(x_i - x_j)\cdot sign(y_i - y_j)$ as the $sign(x_i - x_j)$ and $sign(y_i - y_j)$ measure the order of the item pairs in X and Y with a +1 for item pairs in natural numbers order and -1 for reversed natural numbers order. Therefore, the multiplication of these sign parameters return a +1 for item pairs appearing in the same order (concordant) and a -1 for item pairs appearing in different orders (discordant) in the two rankings.

As shown by the examples below, the reformulated τ_{ap_sign} provides the same result as the τ_{ap} addressed in section 2.2.1, while maintaining the top heaviness.

X	Y
1 A	1 B
2 B	2 C
3 C	3 A
4 D	4 D

$$\tau_{ap} = \frac{2}{4-1}\left(\frac{0}{1}+\frac{1}{2}+\frac{3}{3}\right)-1 = 0$$

$$\tau_{ap\_sign} = \frac{1}{4-1}\left(\frac{-1}{1}+\frac{(-1+1)}{2}+\frac{(+1+1+1)}{3}\right) = 0$$

X	Y
1 A	1 A
2 B	2 C
3 C	3 D
4 D	4 B

$$\tau_{ap} = \frac{2}{4-1}\left(\frac{1}{1}+\frac{1}{2}+\frac{2}{3}\right)-1 = 0.4444$$

$$\tau_{ap\_sign} = \frac{1}{4-1}\left(\frac{+1}{1}+\frac{(+1-1)}{2}+\frac{(+1-1+1)}{3}\right) = 0.4444$$

4.2. DEGREE OF ACCURACY - τ_{ap,a_sign}

For a true ranking X and a ranking by an observer Y, the $\tau_{ap,a}$ was given by J. Urbano and M. Marrero to measure the degree of accuracy of the observer in Y to that of X. This variant, however, becomes invalid when the reference ranking X has ties due to the definition of c_{ij} and the framing of the two terms in $\tau_{ap,a}$.

However, in practical situations there may exist indiscernible ties even in the reference ranking X. Therefore, it is essential to reformulate the $\tau_{ap,a}$ using the sign convention followed by M. Kendall.

The $\tau_{ap,a}$ formulated by J Urbano and M Marrero [22] is:

$$\tau_{ap,a} = \frac{2}{n-1}\left(\sum_{i=t_1+1}^{n}\sum_{j<p_i} c_{ij}\sum_{k=1}^{t_i}\frac{1}{t_i(p_i+k-2)} + \sum_{i=1}^{n}\frac{1}{2}\sum_{k=1}^{t_i-1}\frac{k}{t_i(p_i+k-1)}\right)-1 \tag{4.3}$$

Using the sign convention, $\tau_{\text{ap,a}}$ can be reformulated as:

$$\tau_{ap,a\_sign} = \frac{1}{n-1}\left(\sum_{i=t_1+1}^{n}\sum_{j<p_i} sign(x_i-x_j)\cdot sign(y_i-y_j)\sum_{k=1}^{t_i}\frac{1}{t_i(p_i+k-2)}\right) \tag{4.4}$$

where,

$$sign(x_j-x_i)\cdot sign(y_j-y_i) = \begin{cases} +1 & concordant\ pair \\ -1 & discordant\ pair \\ 0 & tied\ pair \end{cases}$$

By this definition of τ_{ap,a_sign}, the $2\sum_{j<p_i} c_{ij} - 1$ in term 1, which calculates the contribution of items above the tied group based on the position of the pivot within the group, is replaced by $sign(x_i-x_j)\cdot sign(y_i-y_j)$, using the same ideology adopted in the previous section 4.1.1.

The term 2 of $\tau_{ap,a}$ using the sign convention becomes nil. This is because the contribution of an item pair within a tied group in Y, measured by term 2, is concordant to the reference ranking X, in half of the tied group's permutations and discordant in the other half. Therefore, half the +1 contributions cancel out the -1 from the other half, making it unnecessary to account for term 2, while calculating $\tau_{ap,a}$ using the sign convention.

X	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1A	1A	1 A	1 A	1 A	1 A	1 A	1 A
2B	3B	2 B	2 B	2 C	2 C	2 D	2 D
3C	3C	3 C	3 D	3 B	3 D	3 B	3 C
4D	3D	4 D	4 C	4 D	4 B	4 C	4 B

As shown above, the tied item pair (B, C) in Y appears in concordance with the reference ranking X, only in three permutations (Y_1, Y_2 and Y_5) and appears in discordance in the remaining three permutations leading to a total 0 contribution. This explains the elimination of term 2 while reformulating $\tau_{ap,a}$ using the sign convention.

For the example, the $\tau_{ap,a}$ calculated using equation (4.3) and the τ_{ap,a_sign} using (4.4) are both equal to 0.6111.

$$\tau_{ap,a}(X,Y) = \frac{2}{4-1}\left(1\cdot 3\cdot\left(\frac{1}{3(1)}+\frac{1}{3(2)}+\frac{1}{3(3)}\right)+\frac{1}{2}\cdot 3\cdot\left(\frac{1}{3(2)}+\frac{2}{3(3)}\right)\right)-1 = 0.6111$$

$$\tau_{ap,a\_sign}(X,Y) = \frac{1}{4-1}\left(1\cdot 3\cdot\left(\frac{1}{3(1)}+\frac{1}{3(2)}+\frac{1}{3(3)}\right)\right) = 0.6111$$

Moreover, the average of the different τ_{ap_sign}'s ($\tau_{ap\_sign}(X,Y_1) = 1.0$, $\tau_{ap\_sign}(X,Y_2) = 0.7778$, $\tau_{ap\_sign}(X,Y_3) = 0.6667$, $\tau_{ap\_sign}(X,Y_4) = 0.5556$, $\tau_{ap\_sign}(X,Y_5) = 0.4444$, $\tau_{ap\_sign}(X,Y_6) = 0.2222$) possible while computing the rank correlation of the permutations of ranking Y with ranking X is equal to $\tau_{ap,a\_sign} = 0.6111$, similar to the respective $c_{i,j}$ format.

Due to the reformulation of $\tau_{ap,a}$ using the sign convention, the correlation can be computed with ties even in reference ranking X. This is shown below:

X	X_1	X_2	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1.5 A	1 A	1 B	1 A	1 A	1 A	1 A	1 A	1 A	1 A
1.5 B	2 B	2 A	3 B	2 B	2 B	2 C	2 C	2 D	2 D
3 C	3 C	3 C	3 C	3 C	3 D	3 B	3 D	3 B	3 C
4 D	4 D	4 D	3 D	4 D	4 C	4 D	4 B	4 C	4 B

$$\tau_{ap,a\_sign}(X,Y) = \frac{1}{4-1}\left(0\cdot\left(\frac{1}{3(1)}+\frac{1}{3(2)}+\frac{1}{3(3)}\right)+2\cdot\left(\frac{1}{3(1)}+\frac{1}{3(2)}+\frac{1}{3(3)}\right)\right) = 0.4074$$

For the example with ties in X, the average of the different τ_{ap_sign}'s possible calculated using the equation (4.2) ($\tau_{ap\_sign}(X_1,Y_1) = 1.0$, $\tau_{ap\_sign}(X_1,Y_2) = 0.7778$, $\tau_{ap\_sign}(X_1,Y_3) = 0.6667$, $\tau_{ap\_sign}(X_1,Y_4) = 0.5556$, $\tau_{ap\_sign}(X_1,Y_5) = 0.4444$, $\tau_{ap\_sign}(X_1,Y_6) = 0.2222$, $\tau_{ap\_sign}(X_2,Y_1) = 0.3333$, $\tau_{ap\_sign}(X_2,Y_2) = 0.1111$, $\tau_{ap\_sign}(X_2,Y_3) = 0.3333$, $\tau_{ap\_sign}(X_2,Y_4) = 0.3333$, $\tau_{ap\_sign}(X_2,Y_5) = 0.1111$ and $\tau_{ap\_sign}X(_2,Y_6) = 0.0$) was also found to be equal to 0.4074. Therefore, using the sign convention permits ties even in the reference ranking X.

When there are no ties in either of the rankings, the τ_{ap,a_sign} becomes equivalent to τ_{ap_sign}. When all items in either or both of the rankings are tied, the τ_{ap,a_sign} becomes zero, indicating that no conclusion can be derived, since the either or both the rankers were unable to tell apart any of the items in the rankings.

4.3. DEGREE OF AGREEMENT - τ_{ap,b_sign}

For two rankings X and Y, both given by observers, the $\tau_{ap,b}$ variant of the AP Correlation coefficient that J Urbano and M Marrero [22] formulated to measure the degree of agreement between them was addressed in section 3.2.3. This variant calculates the total correlation as the average of the correlation scores obtained when holding one ranking as reference and then the other. This coefficient must neither expect the observers to identify the ties made by the other observer nor award tied pairs in both rankings as concordant and penalize tied pairs in only one ranking as discordant. This is because the ties here refer to items the observers were unable to tell apart (indiscernible ties). The $\tau_{ap,b}$ addressed in detail in section 3.2.3 is given below:

$$\tau_{ap,b} = \frac{\tau_{ap,ties}(X,Y)+\tau_{ap,ties}(Y,X)}{2} \tag{4.5}$$

where $\tau_{ap,ties}$ is given as

$$\tau_{ap,ties} = \frac{2}{n-t_1}\sum_{i=t_1+1}^{n}\sum_{j<p_i}\frac{c_{ij}}{p_i-1}-1 \tag{4.6}$$

Using the sign convention, the τ_{ap,b_sign} and $\tau_{ap,ties\_sign}$ can be reformulated as below:

$$\tau_{ap,b\_sign} = \frac{\tau_{ap,ties\_sign}(X,Y)+\tau_{ap,ties\_sign}(Y,X)}{2} \tag{4.7}$$

$$\tau_{ap,ties\_sign} = \frac{1}{n-t_1} \sum_{i=t_1+1}^{n} \sum_{j<p_i} \frac{sign(x_i - x_j) \cdot sign(y_i - y_j)}{p_i - 1} \tag{4.8}$$

where,

$$sign(x_j - x_i) \cdot sign(y_j - y_i) = \begin{cases} +1 & \textit{concordant pair} \\ -1 & \textit{discordant pair} \\ -1 & \textit{tied pair only in non-reference ranking} \end{cases}$$

The previous ideology of equating $2c_ij - 1$ to $sign(x_i - x_j) \cdot sign(y_i - y_j)$ adopted so far is used when converting $\tau_{ap,b}$ to sign convention as well. However, the $\tau_{ap,b}$ and τ_{ap,b_sign} compute correlation by taking one ranking as reference so a deliberate penalization of tied pairs that exist only in the reference ranking is required by $\tau_{ap,ties\_sign}$.

While holding one observer's ranking as reference and traversing through the other observer's ranking to compute correlation of items above p_i for each i, all the items in the reference raking, even if tied, are expected. This is repeated when the other observer's ranking is held as reference. To eliminate this expectation of tied item pairs in each reference ranking, a deliberate penalization of the tied pairs in the reference ranking is suggested.

This penalization of tied pairs in reference ranking is not equivalent to the penalization of discordant pairs. The discordant pairs are penalized twice when either of the two observers' rankings is held as reference in $\tau_{ap,ties\_sign}(X,Y)$ and $\tau_{ap,ties\_sign}(Y,X)$. This explains how the discordant pairs obtain a -1 total score in τ_{ap,b_sign}. The tied pairs in reference ranking due to their expectation are however penalized only once, simply to eliminate their expectation.

Using this reformulation with $sign(x_i - x_j) \cdot sign(y_i - y_j)$, the $\tau_{ap,ties\_sign}$ is equivalent to the $\tau_{ap,ties}$ formulated by J Urbano and M Marrero in [22]. Therefore, the τ_{ap,b_sign} and $\tau_{ap,b}$ are also equivalent as shown by the example below.

X	Y
1 A	1 A
2.5 B	3 B
2.5 C	3 C
4 D	3 E
5 E	5 D

The $\tau_{ap,b}$ using equation (4.5), computed as the average of $\tau_{ap,ties}(X,Y) = 1.0$ and $\tau_{ap,ties}(Y,X) = 0.6667$ from equation (4.6), is 0.8333. Using sign convention, τ_{ap,b_sign}, $\tau_{ap,ties\_sign}(X,Y)$, $\tau_{ap,ties\_sign}(X,Y)$ computed using equations (4.7) and (4.8) were found to be equivalent to their counterparts calculated using $c_{i,j}$. This is shown below.

$$\tau_{ap,b} = \frac{\tau_{ap,ties\_sign}(X,Y) + \tau_{ap,ties\_sign}(Y,X)}{2} = \frac{(1.0 + 0.6667)}{2} = 0.8333$$

$$\tau_{ap,ties\_sign}(X,Y) = \frac{1}{5-1}\left(\frac{1}{(2-1)} + \frac{1}{(2-1)} + \frac{1}{(2-1)} + \frac{(1+1+1+1)}{(5-1)}\right) = 1.0$$

$$\tau_{ap,ties\_sign}(Y,X) = \frac{1}{5-1}\left(\frac{1}{(2-1)} + \frac{1}{(2-1)} + \frac{(1-1-1)}{(4-1)} + \frac{(1+1+1+1)}{(5-1)}\right) = 0.6667$$

It is important to note that when there are no ties in either or both of the two rankings, $\tau_{ap,ties\_sign}(X,Y)$ becomes equal to $\tau_{ap\_sign}(X,Y)$, $\tau_{ap,ties\_sign}(Y,X)$ becomes equal to $\tau_{ap\_sign}(Y,X)$ and τ_{ap,b_sign} is equivalent to the symmetric average of τ_{ap_sign} which is given by the fraction $\frac{\tau_{ap\_sign}(X,Y)+\tau_{ap\_sign}(Y,X)}{2}$. In the extreme case when all items are tied in either or both of the rankings, the $\tau_{ap,ties\_sign}$ and therefore the τ_{ap,b_sign} is undefined as no item pairs are expected to be concordant between the two rankings. This is caused by the $n - t_1$ in the denominator which evaluates to a zero as all items are tied with the first item i.e $t_1 = n$.

5

EQUAL TIES

The second goal of the thesis to formulate a new variant for Kendall τ and the AP correlation coefficient τ_{ap} which handle equal ties will be addressed in this chapter. The previous chapters only dealt with the cases when there are no ties or indiscernible ties i.e. items the ranker is unable to tell apart, in the rankings. In practical applications, there may also exists truly tied items. This refers to items which are tied, as they are identical or equal to each other, and are hereby termed as equal ties. This section will deal with such ties in either or both of the rankings. As given by the decision tree in Chapter 2, for equal ties in a ranking, two scenarios (where?) are possible. One, when there exists a true ranking and a ranking by an observer and the value of the rank correlation coefficient acts as the degree of accuracy of the observer in his/ her ranking to that of the true ranking. The other, when both the rankings compared for rank correlation are by observers and the value of the rank correlation coefficient acts as a measure of the degree of agreement between the two observers in their rankings. In the following subsections, it will be shown how irrespective of the scenario (where the tie occurs i.e. the type of rankings involved), the case of equal ties has only one generic variant for the Kendall τ and AP correlation coefficient τ_{ap}.

5.1. KENDALL τ_e

In the case of rank correlation computed between a true ranking and an observer's ranking to measure the degree of accuracy of the observer, when there exists truly tied items or equal ties in either of the rankings, the ranker passes a judgement stating that the items are tied. The ranker is, in no means, in a state of indecision like in the case of indiscernible ties addressed earlier. Therefore, when there occurs an equal tie in either of the two rankings, in order for the other ranking to be concordant in this equally tied item pair, the other ranking is also expected to tie the item pair. This holds for the case when the two rankings are given by observers as well i.e. when the rank correlation coefficient acts as a degree of agreement between the two observers. When either of the two observers rank items as equally tied, the other observer's ranking is concordant in those item pairs only when he/ she also ties them as equally tied. In summary, an equal tied pair in only one of the two rankings under comparison is equivalent to a discordant item pair and an item pair tied as equal in both the rankings is equivalent to a concordant item pair.

As shown by table 1 in chapter 3, there exists only one variant for the rank correlation coefficients, irrespective of where the equal tie occurs (true vs. observer or observer vs. observer), when the tied items have the same value. The variant for Kendall rank correlation coefficient, τ_e which accounts for equal ties is given below:

$$\tau_e = \frac{\#concordant - \#discordant}{\#total} = \frac{2}{n(n-1)/2} \sum_{i<j} c_{ij} - 1 \tag{5.1}$$

where,

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

For all the $n(n-1)/2$ unique pairs possible, the $2\sum_{i<j} c_{ij} - 1$ computes the total concordant pairs including equally tied pairs in both the rankings, by the new definition of c_{ij} given above. Therefore, only item pairs occurring in the same order (either natural order of numbers or reversed natural numbers order) and equally

tied item pairs in both the rankings (irrespective of their order of appearance; with sign parameters equating to 0), are in concordance. This is illustrated by the example below:

Always Concordant Permutations of Y with X

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1 A	1 A	1 A	1 A	1 A	1 A	1 A
3 B	3 B	3 B	3 C	3 C	3 D	3 D
3 C	3 C	3 D	3 B	3 D	3 B	3 C
3 D	3 D	3 C	3 D	3 B	3 C	3 B
5 E	5 E	5 E	5 E	5 E	5 E	5 E

Always Discordant Permutations of Y with X

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1 A	1 A	1 A	1 A	1 A	1 A	1 A
2 B	3 B	3 B	3 C	3 C	3 D	3 D
3 C	3 C	3 D	3 B	3 D	3 B	3 C
4 D	3 D	3 C	3 D	3 B	3 C	3 B
5 E	5 E	5 E	5 E	5 E	5 E	5 E

When there is an equally tied group (B, C, D) in the ranking X and when the other ranker also ties this group as equal in his ranking Y, irrespective of the order of these items in Y, the items pairs are always in concordance, as demonstrated by the first example above. Similarly, when there is an equal tied group in only one of the two rankings, say Y (as given in the second example above), irrespective of the order of the equally tied items, the corresponding item pairs are always in discordance.

The τ_e calculated, for the two examples above - always concordant permutations of Y and always discordant permutations of Y with X, are given below:

(Always concordant) $\tau_e = \frac{2}{5(5-1)/2}((1+1+1+1)+(1+1+1)+(1+1)+(1))-1 = 1$

(Always discordant) $\tau_e = \frac{2}{5(5-1)/2}((1+1+1+1)+(1+0+0)+(1+0)+(1))-1 = 0.4$

Irrespective of the type of the two rankings X and Y (true or observer), the computation of τ_e is the same and leads to the same result given above. When all items in both the rankings are equally tied, the two rankings are in perfect concordance with correlation +1. When all items in only one ranking are equally tied, the rankings are in perfect discordance with correlation -1. When none of the items are equally tied in either or both the rankings, the τ_e becomes equal to τ.

5.2. AP CORRELATION $\tau_{ap,e}$

In the case of rank correlation computed with top heaviness between two rankings - true ranking vs. observer ranking or observer ranking vs. observer ranking, when there exists truly tied items or equal ties in the rankings, the ranker passes a judgement stating that the items are tied. The tied items are judged to be equal to each other and the ranker is not in a state of indecision while tying the items. The two rankers are in concordance with respect to the tied items only when the items are identified as equal ties by both the rankers. Otherwise, the two rankers are in discordance with respect to the equally tied item pair. The reformulation of AP Correlation coefficient to account for equal ties in both scenarios, regardless of the types of rankings involved, is given below:

$$\tau_{ap,e} = \frac{2}{n-1}\sum_{i=2}^{n}\sum_{j<i}\frac{c_{ij}}{i-1} - 1 \tag{5.2}$$

where,

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

As shown in chapter 3 and explained in the previous subsection 5.1.1, there is only one variants of τ_{ap}, when handling equal ties in the rankings, regardless of the type of rankings.

Irrespective of the order in which the items are tied as equal in the two rankings, the individual item pairs are either always concordant when the items pairs are equally tied in both rankings or the individual item pairs are always discordant when they are equally tied in only one ranking. This is shown by the examples below:

Always Concordant Permutations of Y with X

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1 A	1 A	1 A	1 A	1 A	1 A	1 A
3 B	3 B	3 B	3 C	3 C	3 D	3 D
3 C	3 C	3 D	3 B	3 D	3 B	3 C
3 D	3 D	3 C	3 D	3 B	3 C	3 B
5 E	5 E	5 E	5 E	5 E	5 E	5 E

Always Discordant Permutations of Y with X

X	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1 A	1 A	1 A	1 A	1 A	1 A	1 A
2 B	3 B	3 B	3 C	3 C	3 D	3 D
3 C	3 C	3 D	3 B	3 D	3 B	3 C
4 D	3 D	3 C	3 D	3 B	3 C	3 B
5 E	5 E	5 E	5 E	5 E	5 E	5 E

The $\tau_{ap,e}$ calculated for the two examples above - always concordant permutations of Y and always discordant permutations of Y with X, are given below:

(Always concordant) $\tau_{ap,e} = \frac{2}{5-1}\left(\frac{1}{2-1} + \frac{2}{3-1} + \frac{3}{4-1} + \frac{4}{5-1}\right) - 1 = 1$

(Always discordant) $\tau_{ap,e} = \frac{2}{5-1}\left(\frac{1}{2-1} + \frac{1}{3-1} + \frac{1}{4-1} + \frac{4}{5-1}\right) - 1 = 0.4167$

Irrespective of the type of the two rankings X and Y (true or observer), the computation of $\tau_{ap,e}$ is the same and leads to the same result given above. When all items in both the rankings are equally tied, the two rankings are in perfect concordance with correlation +1. When all items in only one ranking are equally tied, the rankings are in perfect discordance with correlation -1. When none of the items are equally tied in either or both the rankings, the $\tau_{ap,e}$ becomes equal to τ_{ap}.

6

TIED WITHIN THRESHOLD

The cases addressed earlier, where rank correlation was computed between two rankings with indiscernible or equal ties, only referred to ties that existed when the items had the same value in the rankings.

Many a times, the rankings are compared to find the correlation between them by a third party, different from the rankers. For the benefit of this third party, it is favourable to allow threshold values set by him, which determine whether an item pair is tied or not, as suggested in the research article by J. Urbano and M. Marrero [22]. The idea being that items with values falling within the threshold can be perceived as tied - either equal or indiscernible. This provides flexibility by permitting customizable thresholds to be set by the person performing the rank correlation.

Consider the case where two rankings are composed of nDCG values of different information retrieval systems against different topics. The third party, different from the rankers, who performs the rank correlation, wishes to consider two systems with nDCG values with <0.05 difference as tied for practical purposes. Although such systems with nDCG values with <0.05 difference are not tied originally, artificial, threshold ties can now be induced by the third party, for the sake of practical applications. These threshold ties can have two meanings- indiscernible ties or equal ties, again depending on the choice of the third party. The two scenarios of where a threshold tie could occur (true vs. observer ranking or observer vs. observer ranking) and what the threshold tie means (indiscernible or equal tie), gives rise to 3 new variants for τ and τ_{ap}. Below, these variants are tabulated.

Table 6.1: Different Variants of τ and τ_{ap} in the presence of threshold ties.

Meaning of ties / Context of application	Indiscernible ties	Equal ties
Degree of Accuracy (true vs. observer)	τ_a^w or $\tau_{ap,a}^w$	τ_e^w or $\tau_{ap,e}^w$
Degree of Agreement (observer vs. observer)	τ_b^w or $\tau_{ap,b}^w$	τ_e^w or $\tau_{ap,e}^w$

As shown in the above table, the different thresholds that can be applied by the third party performing the rank correlation, are denoted by the letter w in the formulae. For example, for thresholds of 0.02, 0.15, 0.74, the τ_a^w computed for each of these thresholds are represented as $\tau_a^{0.02}$, $\tau_a^{0.15}$ and $\tau_a^{0.74}$.

When thresholds are used to determine whether items are tied or not, the problem of in-transitivity (i.e. the tied groups are no more transitive) arises. This is due to the presence of tied groups that overlap with each other. Say, A, B, C, D, E and C, D, E, F are two tied groups in a ranking when a threshold of w is applied. The general rule of transitivity where A being tied to C and C being tied to F, resulting in A and F being tied does not hold any more. This is because A and F are part of two different tied groups which share some common elements (C, D and E in this case). Addressing this problem while computing the rank correlation coefficients for the case of threshold ties is of high importance.

In this chapter, we will look at the 3 new variants of τ and τ_{ap} for threshold ties, while accounting for the in-transitivity in their respective sub-sections.

6.1. INDISCERNIBLE THRESHOLD TIES

When a threshold value is set to determine ties by the third party, different from the rankers, who performs the rank correlation, the third party is also given the choice of determining what the tie means (indiscernible or equal). The case when threshold ties refer to indiscernible items is addressed in this section. Based on where the indiscernible threshold ties appear (true vs. observer ranking or observer vs. observer ranking), two variants are possible for τ and τ_{ap}. Moreover, due to the potential difference in scale of the two rankings, it is is necessary to have two different thresholds for the two rankings under comparison.

6.1.1. DEGREE OF ACCURACY - τ_a^w

For a true ranking X and a ranking by an observer Y, the coefficients - τ and τ_{ap} to compute the degree of accuracy of the observer in his ranking Y to the true ranking X, are denoted as τ_a^w and $\tau_{ap,a}^w$ when threshold values determine whether an item pair is tied or not. It is feasible to compute correlation even when the true ranking X has ties, by both τ_a^w and $\tau_{ap,a}^w$, due to the reformulation of τ_{ap} using sign convention (addressed in chapter 4). Therefore, the same threshold or different thresholds (due to the possibility of differences in scale of two rankings) can be applied to the two rankings to induce ties and under these circumstances, the τ_a^w and $\tau_{ap,a}^w$ are calculated as follows.

The τ_a^w to address indiscernible threshold ties requires reformulating the current definition of the sign parameters in τ_a as follows:

$$\tau_a^w = \sum_{i<j} \frac{sign(x_j - x_i) \cdot sign(y_j - y_i)}{n(n-1)/2} \tag{6.1}$$

where,

$$sign(x_j - x_i) \cdot sign(y_j - y_i) = \begin{cases} +1 & \textit{concordant pair} \\ -1 & \textit{discordant pair} \\ 0 & \textit{threshold tied if} |x_j - x_i| \leq w_X \textit{ or } |y_j - y_i| \leq w_Y \end{cases}$$

Due to this reformulation of the sign parameters where w_X and w_Y denote thresholds used in rankings X and Y respectively, there is no need to explicitly account for in-transitivity of the threshold tied groups.

If the threshold for X and Y were set to $w_X = 0.5$ (inclusive) and $w_Y = 0.7$ (inclusive) respectively by the third party, in the following example, (A, B, C) will form an overlapping tied group with (B, C, D) in X and (A, B) will form an overlapping tied group with (B, C) in ranking Y.

X	Y
1 A	1 A
1.4 B	1.5 B
1.5 C	2 C
1.9 D	3 D
3 E	4 E

The τ_a^w is computed as follows:

$$\tau_a^w(X, Y, w_X = 0.5, w_Y = 0.7) = \frac{(0+0+1+1)+(0+0+1)+(0+1)+(1)}{5(5-1)/2} = 0.5$$

With threshold $w = 0$, the τ_a^w becomes equal to τ_a and only items with the same value are considered to be tied. Also, if threshold $w = 0$ and no items in the rankings are tied, $\tau_a^w = \tau_a = \tau$. When threshold w is very large that all items in either ranking are tied, τ_a^w is 0 as no conclusion can be derived about all the items that are indiscernible.

6.1.2. DEGREE OF ACCURACY - $\tau_{ap,a}^w$

As for the AP correlation coefficient, in order to compute the correct values for p_i and t_i used in the formula of $\tau_{ap,a}$ (refer section 2.2.2), for overlapping threshold tied groups to account for in-transitivity, a process that we call stepping is required. As per the process of stepping, the threshold tied groups are divided into sub-groups consisting of items that are exclusive to the overlapping tied groups and items that are common to the tied groups. For example, two overlapping threshold tied groups, as per stepping, will be divide into 3 subgroups - items exclusive to the first tied group, items common to both the groups and items exclusive to the second tied group. The non-tied items will form their own sub-group. For ranking X, the sub-groups are:

X	Sub-groups
1 A	1: (A)
1.4 B	2: (B, C)
1.5 C	3: (D)
1.9 D	4: (E)
3 E	

Since the threshold ties here refer to indiscernible items, this process of stepping is required to limit the total number of permutations to only the permutations where the in-transitivity of the overlapping tied groups is accounted for. For the same example used for τ_a^w, the only permutations permissible for X and Y without breaking the in-transitivity are:

X	X_1	X_2	Y	Y_1	Y_2
1 A	1 A	1 A	1 A	1 A	1 A
1.4 B	2 B	2 C	1.5 B	2 B	2 B
1.5 C	3 C	3 B	2 C	3 C	3 C
1.9 D	4 D	4 D	3 D	4 D	4 E
3 E	5 E	5 E	4 E	5 E	4 D

The numbers with respect to the permutations (X_1, X_2, Y_1 and Y_2) denote the position and the numbers with respect to X and Y denote the values, for the items A, B, C, D and E. Since in X, (A, B, C) forms an overlapping tied group with (B, C, D), only the common items B, C are allowed to interchange their positions without affecting the in-transitivity. Say for example, if B changes its position with A in X, The new group (A, C, D) with a threshold of $w_X = 0.5$ does not hold. Therefore, of the many permutations possible in X and Y, only the permutations tabulated above are permissible and the stepping process ensures this by dividing the overlapping tied groups into sub-groups.

After the tied groups are divided into sub- groups by stepping, the p_i and t_i which measure position of the 1^{st} item in the group of which i is also an item and the size of the group of which i is an item respectively, are computed within each sub-group for the calculation of $\tau_{ap,a}^w$. This is shown below for X and Y:

X			Y		
Sub-groups	p_i	t_i	Sub-groups	p_i	t_i
(A)	(1)	(1)	(A)	(1)	(1)
(B, C)	(2, 2)	(2, 2)	(B)	(2)	(1)
(D)	(4)	(1)	(C)	(3)	(1)
(E)	(5)	(1)	(D)	(4)	(1)
			(E)	(5)	(1)

As shown above, for the sub-group (B, C) in X, both the items take p_i as the position of the 1^{st} item in the group, which is the position of B, $p_B = 2$.

The $\tau_{ap,a}^w$ is then calculated, for the rankings X and Y after performing stepping, as follows with the same definition of the sign parameters like in τ_a^w.

$$\tau_{ap,a}^w = \frac{1}{n-1}\left(\sum_{i=t_1+1}^{n}\sum_{j<p_i} sign(x_i - x_j)\cdot sign(y_i - y_j)\sum_{k=1}^{t_i}\frac{1}{t_i(p_i+k-2)}\right) \tag{6.2}$$

where,

$$sign(x_j - x_i)\cdot sign(y_j - y_i) = \begin{cases} +1 & concordant\ pair \\ -1 & discordant\ pair \\ 0 & threshold\ tied\ if |x_j - x_i| \le w_X\ or\ |y_j - y_i| \le w_Y \end{cases}$$

The $\tau_{ap,a}^w(X,Y)$ for the example given in this section, after performing stepping on both X and Y and X being the reference ranking is given below:

$$\tau_{ap,a}^w(X,Y,w_X = 0.5, w_Y = 0.7) = \frac{1}{5-1}\left(0\cdot\frac{1}{1(1)} + 0\cdot\frac{1}{1(2)} + 1\cdot\frac{1}{1(3)} + 4\cdot\frac{1}{1(4)}\right) = 0.3333$$

Similarly, when Y is the reference ranking, the $\tau^{w}_{ap,a}(Y,X)$ calculated is given below, which happens to be equal to $\tau^{w}_{ap,a}(X,Y)$ for this example. However, this is not the case for all examples.

$$\tau^{w}_{ap,a}(Y,X,w_Y=0.7,w_X=0.5)=\frac{1}{5-1}\left(2\cdot 0\cdot(\frac{1}{2(1)}+\frac{1}{2(2)})+1\cdot\frac{1}{1(3)}+4\cdot\frac{1}{1(4)}\right)=0.3333$$

With threshold $w=0$, the $\tau^{w}_{ap,a}$ becomes equal to $\tau_{ap,a}$ and only items with the same value are considered to be tied. Also, if threshold $w=0$ and no items in the rankings are tied, $\tau^{w}_{ap,a}=\tau_{ap,a}=\tau_{ap}$. When threshold w is very large that all items in either ranking are tied, $\tau^{w}_{ap,a}$ is 0 as no conclusion can be derived about all the items that are indiscernible.

6.1.3. DEGREE OF AGREEMENT - τ^{w}_{b}

For two rankings X and Y, both given by observers, the coefficients τ_b and $\tau_{ap,b}$ to measure the degree of agreement between the observers in their rankings, are denoted by τ^{w}_{b} and $\tau^{w}_{ap,b}$ when threshold values, set by a third party performing the rank correlation, determine whether an item pair is tied or not. The ties in this section refer to indiscernible items with values falling within the threshold set by the third party, different from the rankers, who performs the rank correlation. Since the rankings X and Y can differ in their scale, it is required to allow different thresholds for the two rankings.

The τ^{w}_{b} to address indiscernible threshold ties is given below:

$$\tau^{w}_{b}=\sum_{i<j}\frac{sign(x_j-x_i)\cdot sign(y_j-y_i)}{\sqrt{n(n-1)/2-t_X}\sqrt{n(n-1)/2-t_Y}} \tag{6.3}$$

where,

$$sign(x_j-x_i)\cdot sign(y_j-y_i)=\begin{cases}+1 & \textit{concordant pair}\\ -1 & \textit{discordant pair}\\ 0 & \textit{threshold tied if } |x_j-x_i|\le w_X \textit{ or } |y_j-y_i|\le w_Y\end{cases}$$

t_X = #item pairs for which $x_j - x_i <= w_X$ in X
t_Y = #item pairs for which $y_j - y_i <= w_Y$ in Y

Due to this reformulation of sign parameters where w_X and w_Y denote the thresholds for ranking X and Y, the threshold ties are accounted for by checking whether two items fall within the respective threshold of the rankings. The t_X and t_Y are the count of the item pairs that fall within the respective threshold used in the two rankings X and Y.

For the example below, if the thresholds for X and Y were set to $w_X=0.5$ (inclusive) and $w_Y=0.7$ (inclusive) respectively by the third party, (A, B, C) will form an overlapping tied group with (B, C, D) in X and (A, B) will form an overlapping tied group with (B, C) in ranking Y. The number of tied pairs in X, $t_X=5$ as (A, B), (A, C), (B, C), (B, D) and (C, D) are tied in X, by the application of threshold w_X. The number of tied pairs in Y, $t_Y=2$ as (A, B) and (B, C) are tied in Y, by application of threshold w_Y.

X	Y
1 A	1 A
1.4 B	1.5 B
1.5 C	2 C
1.9 D	3 D
3 E	4 E

For this example, τ^{w}_{b} is computed as follows:

$$\tau^{w}_{b}(X,Y,w_X=0.5,w_Y=0.7)=\frac{(0+0+1+1)+(0+0+1)+(0+1)+(1)}{\sqrt{5(5-1)/2-5}\sqrt{5(5-1)/2-2}}=0.5$$

With threshold $w=0$, the τ^{w}_{b} becomes equal to τ_b and only items with the same value are considered to be tied. Also, if threshold $w=0$ and no items in the rankings are tied, $\tau^{w}_{b}=\tau_b=\tau$. When threshold w is very large that all items in either ranking are tied, τ^{w}_{b} is undefined as no item pairs are expected to be concordant.

6.1.4. DEGREE OF AGREEMENT - $\tau^{w}_{ap,b}$

For two rankings by observers X and Y, the AP correlation coefficient $tau^{w}_{ap,b}$, to calculate the degree of agreement between the observers while handling indiscernible threshold ties, requires redefining p_i and t_i and permitting different thresholds for X and Y. The $tau^{w}_{ap,b}$, however, does not require the stepping process used in $tau^{w}_{ap,a}$.

The p_i is simply the minimum position of all the items i is tied with. In cases where i is the common item between overlapping tied groups, the minimum item position among all these groups is p_i. The reason for this is that, in $tau^{w}_{ap,b}$, only the untied items need to be compared for concordance. By finding the minimum of all items from all tied groups of i, we find the first item above which i could form concordant or discordant item pairs.

The t_i is only used to find t_1 which is the number of items that are tied with and including the 1^{st} item. This eliminates considering first item and its groups as there are no other items above them that can form concordant or discordant pairs.

The $tau^{w}_{ap,b}$ can then be easily computed as

$$\tau^{w}_{ap,b} = \frac{\tau^{w}_{ap,ties}(X,Y) + \tau^{w}_{ap,ties}(Y,X)}{2} \tag{6.4}$$

where $\tau^{w}_{ap,ties}$ is given as

$$\tau^{w}_{ap,ties} = \frac{1}{n-t_1} \sum_{i=t_1+1}^{n} \sum_{j<p_i} \frac{sign(x_i - x_j) \cdot sign(y_i - y_j)}{p_i - 1} \tag{6.5}$$

where,

$$sign(x_j - x_i) \cdot sign(y_j - y_i) = \begin{cases} +1 & \textit{concordant pair} \\ -1 & \textit{discordant pair} \\ -1 & \textit{threshold tied only in non-reference ranking tied } |y_j - y_i| \leq w_Y \end{cases}$$

p_i = minimum position of all items tied with i
t_i = number of items tied with and including i

X		Y	
1	A	1	A
1.4	B	1.5	B
1.5	C	2	C
1.9	D	3	D
3	E	4	E

For the example above, if the thresholds for X and Y were set to $w_X = 0.5$ (inclusive) and $w_Y = 0.7$ (inclusive) respectively, the p_i and t_i when each ranking is the reference is calculated as follows:

X			Y		
Sub-groups	p_i	t_i	Sub-groups	p_i	t_i
1 A	1	3	1 A	1	2
1.4 B	1	4	1.5 B	1	3
1.5 C	1	3	2 C	2	2
1.9 D	2	3	3 D	4	1
3 E	5	1	4 E	5	1

For the X ranking, the items (A, B, C) form an overlapping tied group with (B, C, D). Therefore, for items A, B and C, the minimum item position that they are tied with is $p_i = 1$ referring to A's position. For item D, the minimum item position that it is tied with is B and hence $p_i = 2$. The $t_i = 4$ for item B, as combining the two tied groups, the total number of items that are tied with and including B is 4 (A, B, C, D). Similarly, p_i and t_i are calculated for ranking Y to compute $\tau^{w}_{ap,ties}$ when Y is the reference ranking.

For this example, the $\tau^{w}_{ap,b}$, $\tau^{w}_{ap,ties}(X,Y)$ with X as reference ranking and $\tau^{w}_{ap,ties}(X,Y)$ with Y as reference ranking are calculated as:

$$\tau^{w}_{ap,b} = \frac{\tau^{w}_{ap,ties}(X,Y) + \tau^{w}_{ap,ties}(Y,X)}{2} = \frac{(1-0.1111)}{2} = 0.4444$$

$$\tau^{w}_{ap,ties}(X,Y,w_X = 0.5, w_Y = 0.7) = \frac{1}{5-3}\left(\frac{1}{(2-1)} + \frac{4}{(5-1)}\right) = 1$$

$$\tau^{w}_{ap,ties}(Y,X,w_Y = 0.7, w_X = 0.5) = \frac{1}{5-2}\left(\frac{-1}{(2-1)} + \frac{1-1-1}{(4-1)} + \frac{4}{(5-1)}\right) = -0.1111$$

With threshold $w = 0$, the $\tau^{w}_{ap,b}$ becomes equal to $\tau_{ap,b}$ and only items with the same value are considered to be tied. Also, if threshold $w = 0$ and no items in the rankings are tied, $\tau^{w}_{ap,b} = \tau_{ap,b}$ = symmetric τ_{ap}. When threshold w is very large that all items in either ranking are tied, $\tau^{w}_{ap,b}$ is undefined as no item pairs are expected to be concordant.

6.2. EQUAL THRESHOLD TIES

When a threshold value is set by the third party, different from the rankers, who performs the rank correlation to determine whether items are tied and a tie refers to items that are equivalent to each other, the ties are termed equal threshold ties. Similar to τ_e and $\tau_{ap,e}$, when there are equal ties in rankings, the two variants of the Kendall and AP correlation coefficients a and b are the same. This is because the two rankings, irrespective of whether they are the true vs. observer rankings or observer vs. observer rankings, are concordant with respect to tied items if the items are tied as equal in both the rankings or discordant otherwise. Unlike in indiscernible ties where a neutral 0 is awarded, the mismatch of equal tied items between rankings is treated as discordant with a -1 penalization because the ranker tying the items passes a judgement about the equal ties and hence these ties are expected to be tied by the other ranker in order to be concordant.

To account for the difference in scale between the two rankings X and Y, it is required to have different thresholds for the rankings that determine whether items are equal tied or not. The Kendall and AP correlation coefficient to address equal threshold ties will be discussed in their respective subsections.

6.2.1. KENDALL τ^{w}_{e}

The Kendall τ to handle equal threshold ties when the two rankings are either true vs observer ranking or observer vs observer ranking is given as:

$$\tau^{w}_{e} = \frac{2}{n(n-1)/2}\sum_{i<j} c_{ij} - 1 \tag{6.6}$$

where,

$$c_{ij} = \begin{cases} 1 & concordant\ if\ sign(x_j - x_i) = sign(y_j - y_i) \\ 1 & threshold\ tied\ if\ |x_j - x_i| \le w_X\ and\ |y_j - y_i| \le w_Y \\ 0 & otherwise \end{cases}$$

As shown above, having threshold ties in both ranking is treated as concordant by awarding a 1 and all other cases, including the case where tied items are threshold tied in only one of the two rankings are treated as discordant. This is satisfies the definition for equal threshold ties given in 6.3.

X	Y
1 A	1 A
1.4 B	1.5 B
1.5 C	2 C
1.9 D	3 D
3 E	4 E

For the example above, if the thresholds for X and Y were set to $w_X = 0.5$ (inclusive) and $w_Y = 0.7$ (inclusive) respectively, the τ^{w}_{e} can be calculated as

$$\tau^{w}_{e} = \frac{2}{5(5-1)/2}\left((1+0+1+1)+(1+0+1)+(0+1)+(1)\right) - 1 = 0.4$$

The equally tied items (A, B) and (B, C), due to threshold in both the rankings are awarded a +1 and treated as concordant item pairs.

With threshold $w = 0$, the τ_e^w becomes equal to τ_e and only items with the same value are considered to be tied. Also, if threshold $w = 0$ and no items in the rankings are tied, $\tau_e^w = \tau_e = \tau$. When threshold w is very large that all items in either ranking are tied, τ_e^w is 1 as all item pairs are equally tied in both rankings and are in perfect concordance.

6.2.2. AP CORRELATION $\tau_{ap,e}^w$

The AP Correlation coefficient to handle equal threshold ties when the two rankings are either true vs observer ranking or observer vs observer ranking is given as:

$$\tau_{ap,e}^w = \frac{2}{n-1}\sum_{i=2}^{n}\sum_{j<i}\frac{c_{ij}}{i-1} - 1 \tag{6.7}$$

where,

$$c_{ij} = \begin{cases} 1 & concordant\ if\ sign(x_j - x_i) = sign(y_j - y_i) \\ 1 & threshold\ tied\ if\ |x_j - x_i| \le w_X\ and\ |y_j - y_i| \le w_Y \\ 0 & otherwise \end{cases}$$

As shown above, having threshold ties in both ranking is treated as concordant by awarding a 1 and all other cases, including the case where items are threshold tied in only one of the two rankings, are treated as discordant. This is satisfies the definition for equal threshold ties given in 6.3.

X	Y
1 A	1 A
1.4 B	1.5 B
1.5 C	2 C
1.9 D	3 D
3 E	4 E

For the example above, if the thresholds for X and Y were set to $w_X = 0.5$ (inclusive) and $w_Y = 0.7$ (inclusive) respectively, the $\tau_{ap,e}^w$ with top heaviness can be calculated as

$$\tau_{ap,e}^w = \frac{2}{5-1}\left(\frac{1}{2-1} + \frac{0+1}{3-1} + \frac{1+0+0}{4-1} + \frac{1+1+1+1}{5-1}\right) - 1 = 0.4167$$

The equally tied items (A, B) and (B, C), due to threshold in both the rankings are awarded a +1 and treated as concordant item pairs.

With threshold $w = 0$, the $\tau_{ap,e}^w$ becomes equal to $\tau_{ap,e}$ and only items with the same value are considered to be tied. Also, if threshold $w = 0$ and no items in the rankings are tied, $\tau_{ap,e}^w = \tau_{ap,e} = \tau_{ap}$. When threshold w is very large that all items in either ranking are tied, $\tau_{ap,e}^w$ is 1 as all item pairs are equally tied in both rankings and are in perfect concordance.

7

PRACTICAL ASSESSMENT

In this chapter, the different correlation coefficients proposed will be assessed by a series of 2 experiments related to IR evaluation research. The purpose of the first experiment is to evaluate whether the new coefficients proposed capture different ideas about the rankings and also to test and draw inferences from the results of these coefficients. The purpose of the second experiment is to emphasize the importance and benefit of comparing system rankings on topic level.

7.1. EXPERIMENT 1 - CORRELATION ANALYSIS IN IR

Here, the experiment in [1] by W. Webber et al is carried out for all correlation coefficients proposed in this thesis. The experiment required performing correlation analysis on system rankings from different topic sets similar to that demonstrated in Figure. 1.1 using the 5 evaluation measures - P@10 (Precision at 10 documents retrieved), RR (Reciprocal Rank), RBP.95 (Rank Biased Precision), AP (Average Precision) and nDCG (Normalized Discounted Cumulative Gain). This experiment was carried out in [1] to demonstrate whether, the complex metrics like nDCG were better than the simple metrics like P@10 and RR, in predicting the performance of the same simple metrics on new, untested topics, thus making it redundant to report the simple metrics.

7.1.1. DATA AND METHOD

The data used was the TREC[1] 8 Adhoc track[2] which is a news track consisting of 50 topics from 401–450 with binary relevant judgements and 100 documents pool-depth. A total of 129 system runs are submitted to this TREC track.

W. Webber et al, to predict the performance of the 5 evaluation measures for untested topics based on the system ranking from experimental topics, randomly partitioned the topic sets in the trec_eval results of the five metrics into two halves, after eliminating the bottom 25% systems. The Kendall τ was then used to measure the predictive power between the mean system rankings from the two partitioned topic sets. In simple terms, correlation analysis, like shown in Figure. 1.1, was carried out to determine how well the 5 measures predict each other. This random partitioning and calculation of τ correlation was repeated 2000 times to reduce random error. It was concluded in [1] that the simple evaluation metrics like P@10 and RR were poorer self-predictors and that the more complex metrics like nDCG, being better at predicting the simple metric like P@10 than itself, makes calculating P@10 to be redundant.

Since W. Webber et al intended to find how well the evaluation measures predict each other, this case is an example of correlation between true vs. observer ranking and the correlation score ('a'variant of the coefficients) acts as a measure of the degree of accuracy of the observer to the true ranking. Here, the true ranking is the system rankings on new topics based on an evaluation measure that is going to be predicted. However, all topic collections are only a sample from the true total collection, which is infinitely large, comprising of every possible topic in a given track. Such total true collection is unknown, which makes calling the system

[1] Text REtrieval Conference consists of different tracks, each with the necessary infrastructure (test collections, evaluation methodology, etc. to carry out large-scale evaluation of text retrieval methodologies

[2] The different IR areas of research in TREC

rankings, from topic collections — available at TREC — partitioned into half, as true ranking, incorrect. Moreover, there is no clear mention of which variant of the τ coefficient was used to measure the predictive power of the evaluation measures.

7.1.2. CORRELATION ANALYSIS

For the correlation analysis between rankings from the 5 evaluation measures, the same experiment similar to [1], by random partitioning of topic collections and computing correlation for 2000 trials as explained in previous section, was repeated with the elimination of bottom 25% systems and also some duplicate systems from the 3 pairs - (sys 36, sys 37), (sys 69, sys 70) and (sys 94, sys 95), which lacked any mention in the paper [1]. Since the evaluation metrics - RR, RBP.95, nDCG and AP had a number of tied systems in the rankings, for some trials out of the 2000, τ and τ_{ap} could not be performed for such trials. However, in the case of P@10, all the trials had tied items and hence the predictive power of P@10 with itself and any of the other metrics could not be calculated using τ and τ_{ap}.

It was not mentioned how the ties were handled in [1], and it is highly probable that the R implementation of τ was used. In which case, it is important to note that the R implementation of τ correlation coefficient, by default, resorts to the calculation of τ_b in the presence of ties in rankings. Therefore, the predictive power measured in [1] could be τ_b and not τ_a as they intended.

Table 7.1: Copy of Table. 1 from W. Webber et al [1]: Predictive power ϕ of different metrics on the top 75% of TREC 8 AdHoc Track systems, calculated from 2,000 random repartitionings of the topic set.

	P@10	RR	RBP.95	AP	nDCG
P@10	0.50	0.40	0.53	0.51	0.50
RR		0.39	0.41	0.36	0.37
RBP.95			0.58	0.57	0.55
AP				0.63	0.60
nDCG					0.61

The table from W. Webber et al which calculates the predictive power of each measure against one another is shown in Table 7.1. The tables generated, as a result of the correlation analysis, for the existing correlation coefficients - τ, τ_a, τ_b, τ_{ap}, $\tau_{ap,a}$ and $\tau_{ap,b}$ and for the new coefficients proposed in this thesis - τ_e, $\tau_{ap,e}$ and τ_a^w, τ_b^w, τ_e^w, $\tau_{ap,a}^w$, $\tau_{ap,b}^w$, $\tau_{ap,e}^w$ with thresholds of 0.01, 0.05 and 0.1, are tabulated in Table 7.2 through 7.5, where the rows refer to evaluation measure used in true ranking and columns to that in observer's ranking.

From these tables, similar overall conclusions to [1], that simpler metrics are poorer self predictors and that even the simple metric P@10, on average, is a better predictor of RR than itself, making RR a poor metric, are observed. By the use of the new variants of the τ and τ_{ap} correlation coefficients, further interesting observations can be made about the coefficients from close analysis of the tables. These are given below:

- The τ_b values are relatively higher than τ_a for P@10 comparisons due to the presence of higher number of tied items in P@10 rankings. This high value of τ_b is the result of the indiscernible tied items, not being expected to be concordant between the rankings and hence, not accounted for, in the denominator of τ_b.

- For evaluation metrics other than P@10, the τ_a and τ_b are mostly equal to each other but not equal to τ, indicating the presence of tied items in low numbers in the rankings. This conclusion was derived because, for the τ_a and τ_b to be equal, their denominators should be very similar. This only happens when the number of tied item pairs subtracted from τ_b's denominator is very low, leading to the inference that the rankings for the evaluation metrics other than P@10 have lesser tied items. The same can be verified by looking at the results from trec_eval.

- The τ_e is smaller than τ_a and τ_b for rankings with higher number of ties (refer row and column of P@10) because equal ties appearing in only one ranking are considered to be discordant and are penalized.

- For τ_a^w, as the value of the threshold w increases, the closer the correlation is to 0. This relationship between the threshold and correlation is caused by the increased number of indiscernible ties with high

Table 7.2: Predictive power of different metrics as a measure of τ, τ_a, τ_b, τ_e, τ_{ap}, $\tau_{ap,a}$, $\tau_{ap,b}$, $\tau_{ap,e}$

τ

	P@10	RR	RBP.95	AP	nDCG
P@10	-	-	-	-	-
RR	-	0.39	0.40	0.35	0.34
RBP	-	0.40	0.56	0.56	0.53
AP	-	0.35	0.56	0.61	0.59
nDCG	-	0.34	0.53	0.59	0.61

τ_{ap}

	P@10	RR	RBP.95	AP	nDCG
P@10	-	-	-	-	-
RR	-	0.34	0.40	0.36	0.33
RBP	-	0.38	0.56	0.53	0.47
AP	-	0.35	0.53	0.56	0.53
nDCG	-	0.29	0.43	0.49	0.52

τ_a

	P@10	RR	RBP.95	AP	nDCG
P@10	0.47	0.39	0.52	0.49	0.47
RR	0.39	0.38	0.40	0.36	0.35
RBP	0.52	0.40	0.56	0.56	0.53
AP	0.49	0.35	0.56	0.62	0.59
Ndcg	0.47	0.34	0.53	0.59	0.61

$\tau_{ap,a}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.47	0.36	0.49	0.44	0.41
RR	0.40	0.34	0.40	0.36	0.33
RBP	0.51	0.39	0.56	0.53	0.47
AP	0.47	0.35	0.53	0.56	0.53
nDCG	0.39	0.29	0.43	0.49	0.52

τ_b

	P@10	RR	RBP.95	AP	nDCG
P@10	0.48	0.40	0.52	0.50	0.48
RR	0.40	0.38	0.40	0.36	0.35
RBP	0.52	0.40	0.56	0.56	0.53
AP	0.50	0.35	0.56	0.62	0.59
nDCG	0.47	0.34	0.53	0.59	0.61

$\tau_{ap,b}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.48	0.38	0.51	0.46	0.40
RR	0.38	0.34	0.40	0.36	0.31
RBP	0.51	0.39	0.56	0.53	0.45
AP	0.47	0.36	0.53	0.56	0.51
nDCG	0.40	0.31	0.45	0.51	0.52

τ_e

	P@10	RR	RBP.95	AP	nDCG
P@10	0.44	0.37	0.50	0.48	0.45
RR	0.38	0.38	0.40	0.36	0.35
RBP	0.50	0.40	0.56	0.56	0.53
AP	0.48	0.35	0.56	0.62	0.59
nDCG	0.45	0.34	0.53	0.59	0.61

$\tau_{ap,e}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.44	0.36	0.49	0.44	0.41
RR	0.37	0.34	0.40	0.36	0.33
RBP	0.49	0.38	0.56	0.53	0.47
AP	0.45	0.35	0.53	0.56	0.53
nDCG	0.37	0.29	0.43	0.49	0.52

Table 7.3: Predictive power of different metrics as a measure of $\tau_a^{0.01}$, $\tau_b^{0.01}$, $\tau_e^{0.01}$, $\tau_{ap,a}^{0.01}$, $\tau_{ap,b}^{0.01}$, $\tau_{ap,e}^{0.01}$

$\tau_a^{0.01}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.47	0.39	0.51	0.48	0.46
RR	0.39	0.38	0.40	0.35	0.34
RBP	0.51	0.39	0.55	0.54	0.51
AP	0.48	0.35	0.54	0.59	0.57
nDCG	0.46	0.34	0.51	0.57	0.59

$\tau_{ap,a}^{0.01}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.47	0.37	0.50	0.45	0.42
RR	0.39	0.34	0.40	0.36	0.32
RBP	0.50	0.38	0.55	0.51	0.46
AP	0.46	0.34	0.51	0.53	0.50
nDCG	0.38	0.29	0.42	0.47	0.50

$\tau_b^{0.01}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.52	0.42	0.57	0.54	0.52
RR	0.42	0.41	0.44	0.39	0.38
RBP	0.57	0.43	0.63	0.62	0.58
AP	0.54	0.39	0.62	0.68	0.65
nDCG	0.51	0.37	0.58	0.65	0.66

$\tau_{ap,b}^{0.01}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.54	0.43	0.58	0.53	0.46
RR	0.43	0.37	0.45	0.41	0.35
RBP	0.58	0.44	0.64	0.61	0.51
AP	0.54	0.41	0.61	0.64	0.57
nDCG	0.46	0.35	0.51	0.57	0.59

$\tau_e^{0.01}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.32	0.26	0.35	0.31	0.30
RR	0.26	0.27	0.24	0.19	0.19
RBP	0.34	0.24	0.38	0.36	0.34
AP	0.31	0.19	0.36	0.42	0.40
nDCG	0.30	0.19	0.34	0.40	0.43

$\tau_{ap,e}^{0.01}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.32	0.22	0.33	0.26	0.22
RR	0.24	0.20	0.22	0.16	0.13
RBP	0.34	0.21	0.39	0.31	0.26
AP	0.29	0.17	0.33	0.34	0.31
nDCG	0.21	0.12	0.23	0.28	0.32

Table 7.4: Predictive power of different metrics as a measure of $\tau_a^{0.05}$, $\tau_b^{0.05}$, $\tau_e^{0.05}$, $\tau_{ap,a}^{0.05}$, $\tau_{ap,b}^{0.05}$, $\tau_{ap,e}^{0.05}$

$\tau_a^{0.05}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.35	0.32	0.34	0.29	0.30
RR	0.32	0.33	0.30	0.24	0.25
RBP	0.34	0.30	0.32	0.28	0.28
AP	0.29	0.24	0.28	0.28	0.28
nDCG	0.30	0.25	0.28	0.28	0.32

$\tau_{ap,a}^{0.05}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.36	0.30	0.33	0.26	0.24
RR	0.32	0.29	0.29	0.22	0.20
RBP	0.34	0.28	0.32	0.25	0.22
AP	0.28	0.23	0.26	0.23	0.21
nDCG	0.24	0.20	0.21	0.20	0.23

$\tau_b^{0.05}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.62	0.51	0.65	0.59	0.56
RR	0.51	0.49	0.53	0.45	0.43
RBP	0.65	0.53	0.69	0.64	0.59
AP	0.59	0.45	0.64	0.66	0.62
nDCG	0.56	0.43	0.59	0.62	0.67

$\tau_{ap,b}^{0.05}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.68	0.56	0.71	0.64	0.56
RR	0.56	0.48	0.59	0.54	0.46
RBP	0.71	0.59	0.74	0.68	0.56
AP	0.64	0.54	0.68	0.68	0.57
nDCG	0.56	0.46	0.56	0.57	0.61

$\tau_e^{0.05}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.24	0.09	0.31	0.23	0.18
RR	0.09	0.08	0.09	-0.01	-0.01
RBP	0.31	0.09	0.42	0.37	0.28
AP	0.23	-0.01	0.37	0.43	0.34
nDCG	0.18	-0.01	0.28	0.34	0.36

$\tau_{ap,e}^{0.05}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.30	0.11	0.33	0.20	0.16
RR	0.13	0.06	0.12	0.01	-0.03
RBP	0.37	0.14	0.44	0.35	0.28
AP	0.27	0.07	0.37	0.42	0.35
nDCG	0.17	0.00	0.25	0.30	0.35

Table 7.5: Predictive power of different metrics as a measure of $\tau_a^{0.10}$, $\tau_b^{0.10}$, $\tau_e^{0.10}$, $\tau_{ap,a}^{0.10}$, $\tau_{ap,b}^{0.10}$, $\tau_{ap,e}^{0.10}$

$\tau_a^{0.10}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.22	0.20	0.17	0.11	0.11
RR	0.20	0.21	0.15	0.09	0.10
RBP	0.18	0.15	0.15	0.10	0.09
AP	0.11	0.09	0.10	0.07	0.06
nDCG	0.11	0.10	0.09	0.06	0.09

$\tau_{ap,a}^{0.10}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.24	0.18	0.16	0.09	0.07
RR	0.20	0.17	0.14	0.07	0.06
RBP	0.18	0.14	0.13	0.08	0.06
AP	0.10	0.08	0.08	0.05	0.04
nDCG	0.08	0.07	0.06	0.04	0.05

$\tau_b^{0.10}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.71	0.54	0.68	0.53	0.47
RR	0.54	0.50	0.51	0.39	0.37
RBP	0.69	0.51	0.71	0.59	0.47
AP	0.53	0.40	0.59	0.55	0.44
nDCG	0.47	0.37	0.47	0.44	0.52

$\tau_{ap,b}^{0.10}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.79	0.64	0.75	0.64	0.54
RR	0.64	0.55	0.62	0.55	0.45
RBP	0.75	0.62	0.75	0.65	0.47
AP	0.64	0.55	0.65	0.58	0.42
nDCG	0.54	0.45	0.47	0.42	0.48

$\tau_e^{0.10}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.62	0.33	0.65	0.53	0.46
RR	0.33	0.21	0.34	0.26	0.22
RBP	0.65	0.34	0.75	0.70	0.59
AP	0.54	0.27	0.71	0.76	0.65
nDCG	0.46	0.22	0.59	0.65	0.66

$\tau_{ap,e}^{0.10}$

	P@10	RR	RBP.95	AP	nDCG
P@10	0.66	0.37	0.59	0.46	0.40
RR	0.40	0.26	0.37	0.28	0.24
RBP	0.65	0.42	0.72	0.66	0.60
AP	0.54	0.39	0.71	0.76	0.72
nDCG	0.43	0.33	0.59	0.68	0.74

values of threshold, which makes the rankings more independent of each other, leading to the 0 correlation. For previously concordant pairs, the increased threshold which artificially ties them as indiscernible, reduces the overall correlation towards 0. For previously discordant pairs, the high threshold increases the correlation towards 0. This complies with τ_a^w becoming 0 when all items in rankings are

tied.

- For τ_b^w, a definite pattern cannot be observed always because with the increase in threshold and the subsequent increase in the number of tied items, the correlation of the entire ranked lists solely depends on the untied item pairs which could either be concordant or discordant. This explains the lack of a definite pattern between $\tau_b^{0.01}$, $\tau_b^{0.05}$ and $\tau_b^{0.10}$.

- Another very interesting observation is for the τ_e^w coefficient, which initially decreases in correlation when threshold is increased from 0.01 to 0.05 and later increases even higher than the original value τ_e for most of the metrics. This is explained by the presence of tied groups in the two rankings that do not span over the same items, which initially cause the correlation to drop as the mismatch of equal ties in rankings is treated as discordant. With the increase in threshold to 0.10, the equal tied groups in the two rankings have increased in size to span over the same items, leading to a higher concordance and hence, a higher correlation. Therefore, the threshold which causes the correlation to stop decreasing can act as a measure of how far apart the equal tied groups in the two rankings are.

- The τ_{ap} variants ($\tau_{ap,a}$, $\tau_{ap,b}$, $\tau_{ap,e}$, $\tau_{ap,a}^w$, $\tau_{ap,b}^w$ and $\tau_{ap,e}^w$) all follow their τ counterparts.

- However, with top heaviness in the AP correlation coefficient, any pattern discussed above may not be obvious, i.e. the top heaviness may cloud any expected pattern between the correlation scores. For example, in computing $\tau_{ap,a}^w$ in a ranking with bottom items closer in value and top items very far apart, the increase in threshold value will reduce a previously concordant correlation, closer to 0 as expected, but not as much to make the pattern easily observable. This is because only the bottom items will be artificially tied, with low increase in thresholds, whose weight is not as high as the top items. Therefore, the pattern of correlation approaching 0 with increased threshold will not be very obvious, until the threshold is large enough that even the top items, far apart from each other, are tied.

7.1.3. P@10 PREDICTED BY NDCG

During the random partitioning into two halves for the entire topic collection, the predictive power, calculated as τ in [1], was also computed for reduced total topic set sizes of (10, 20, 30, 40) as well. This was intended to show whether the complex meric, nDCG was better at predicting P@10 than itself, for varied topic set sizes, to emphasize the argument that it is redundant to report a simple metric like P@10. The original figure from W. Webber et al paper is shown below:

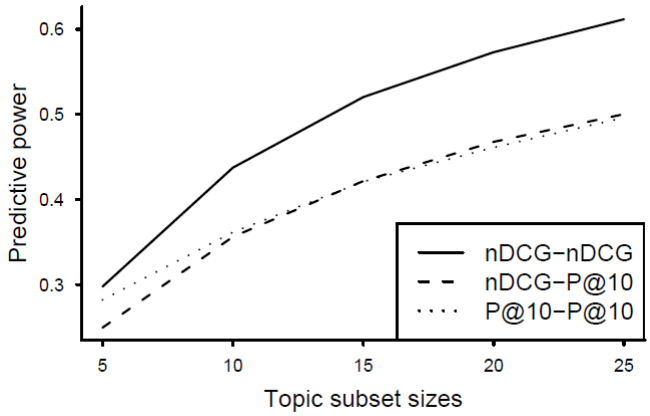

Figure 7.1: Copy of Figure. 1 from W. Webber et al [1]: Predictive power ϕ of nDCG and P@10 of themselves, and nDCG of P@10, with different topic subset sizes, on the TREC 8 runs.

Figures, similar to Figure. 7.1, were made for the different correlation coefficients addressed in this thesis. The legends in the graphs denote the first measure to be predicted by the other. Therefore, P@10-nDCG refer to P@10 being predicted by nDCG. For τ and τ_{ap}, all comparisons involving P@10 had ties and the coefficients

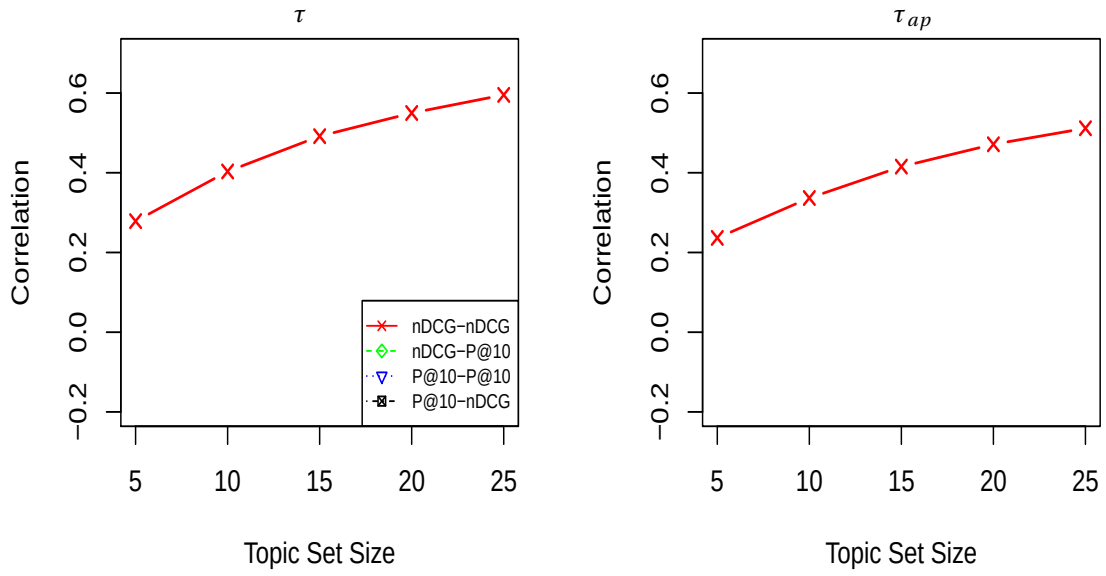

Figure 7.2: Predictive power of nDCG and P@10 of themselves and each other as a measure of τ_a and τ_{ap} with different topic subset sizes

could not be calculated. Therefore, this leads us to believe that either the 'a'or 'b'variant of τ were used to find the correlation values for P@10 but we are not sure as there is no mention about this in W. Webber et al

As reported by W. Webber et al in [1], that nDCG with increased topic set sizes is better at predicting P@10, which makes reporting P@10 redundant is observed only when τ_e is used to measure the predictive power. For the rest of the coefficients, as shown by the charts above, P@10 is at least as good as nDCG (if not better) in predicting itself. This can be noticed from observing the nDCG-P@10 and P@10-P@10 lines in the Figure. 7.2 through 7.6. Therefore, concluding P@10 to be redundant is not valid because under different scenarios, determined by the three dimensions of ties in rankings (where, when, what), the corresponding correlation coefficients calculating the predictive power show different results.

Moreover, for the chart corresponding to $\tau^{0.05}_{ap,a}$ coefficient, the simple metric P@10 is observed to be as good as nDCG in predicting nDCG, requiring further investigated as this was observed only for the $\tau^{0.05}_{ap,a}$ coefficient. This contradicts the claim that only complex metrics in [1] are able to predict the simple metrics as good as, if not better than the simple metrics themselves.

7.2. EXPERIMENT 2 - TOPIC VARIABILITY OF IR SYSTEMS

Here, the experiment similar to that in [2] for the TREC 8 Adhoc track, will be carried out to assess the topic-wise performance of the systems based on infAP (inferred AP) with incomplete judgements against actual AP with complete judgements. For this, the experiment involves computing AP with a full pool of judgments and infAP with different sizes of judgements generated randomly as proposed in [2]. The incomplete judgements were generated by random sampling of only p% of the judgements for each topic as judged and the remaining (100-p)% of the total judgements are marked as not judged. This experiment was framed to check how well the measure infAP for mean system rankings with incomplete judgements correlate to the actual AP for depth 100 complete judgements to find the tradeoff between p and the estimation accuracy.

7.2.1. DATA AND METHOD

The data used is the same TREC 8 Adhoc track from the previous experiment. To test whether the infAP proposed in [2] performed well with incomplete relevance judgements, random sampling of the judgements was performed to create incomplete judgements of different pool sizes. This was explained in the introduction of this section 7.2. It is important to note that during pooling at least one of the retrieved judgements per topic needs to be relevant (i.e. at least one judgement with +1 relevance). The remaining (100-p)% judgements are marked as not judged by deliberately assigning -1 relevance. The relevance scores -1, 0 and +1 for the judge-

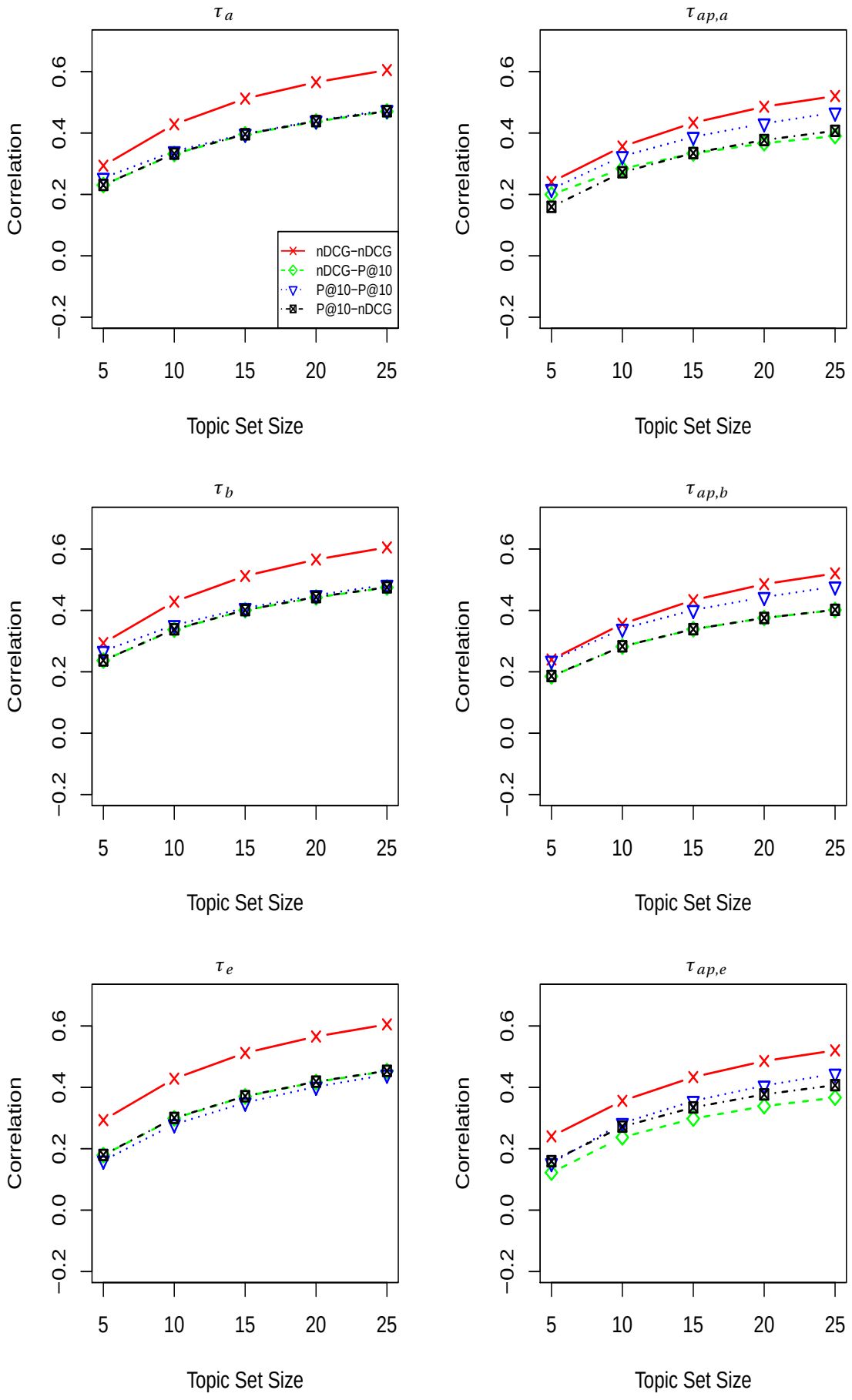

Figure 7.3: Predictive power of nDCG and P@10 of themselves and each other as a measure of τ_a, τ_b, τ_e, τ_{ap}, $\tau_{ap,b}$, $\tau_{ap,e}$ with different topic subset sizes

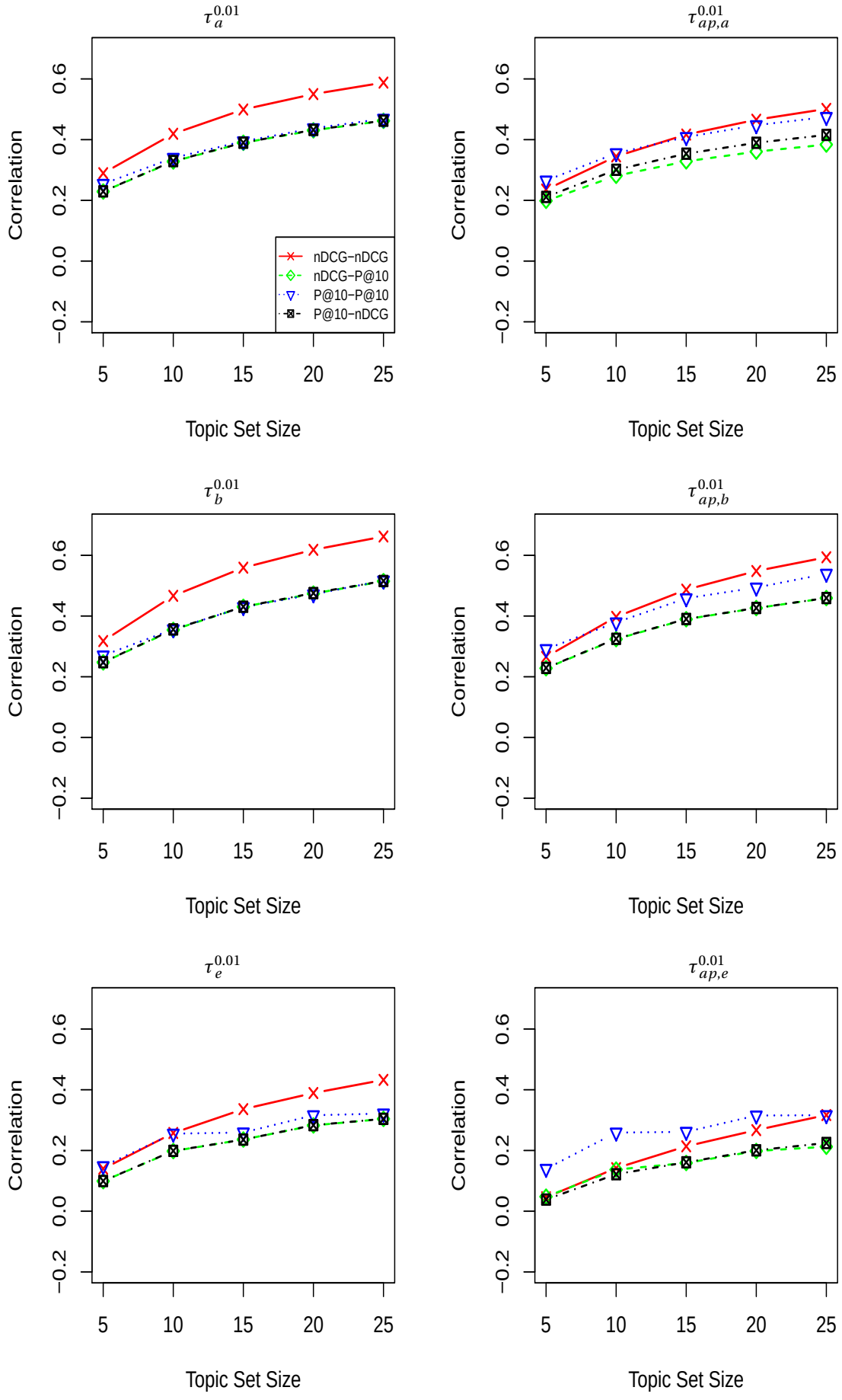

Figure 7.4: Predictive power of nDCG and P@10 of themselves and each other as a measure of $\tau_a^{0.01}$, $\tau_b^{0.01}$, $\tau_e^{0.01}$, $\tau_{ap,a}^{0.01}$, $\tau_{ap,b}^{0.01}$, $\tau_{ap,e}^{0.01}$ with different topic subset sizes

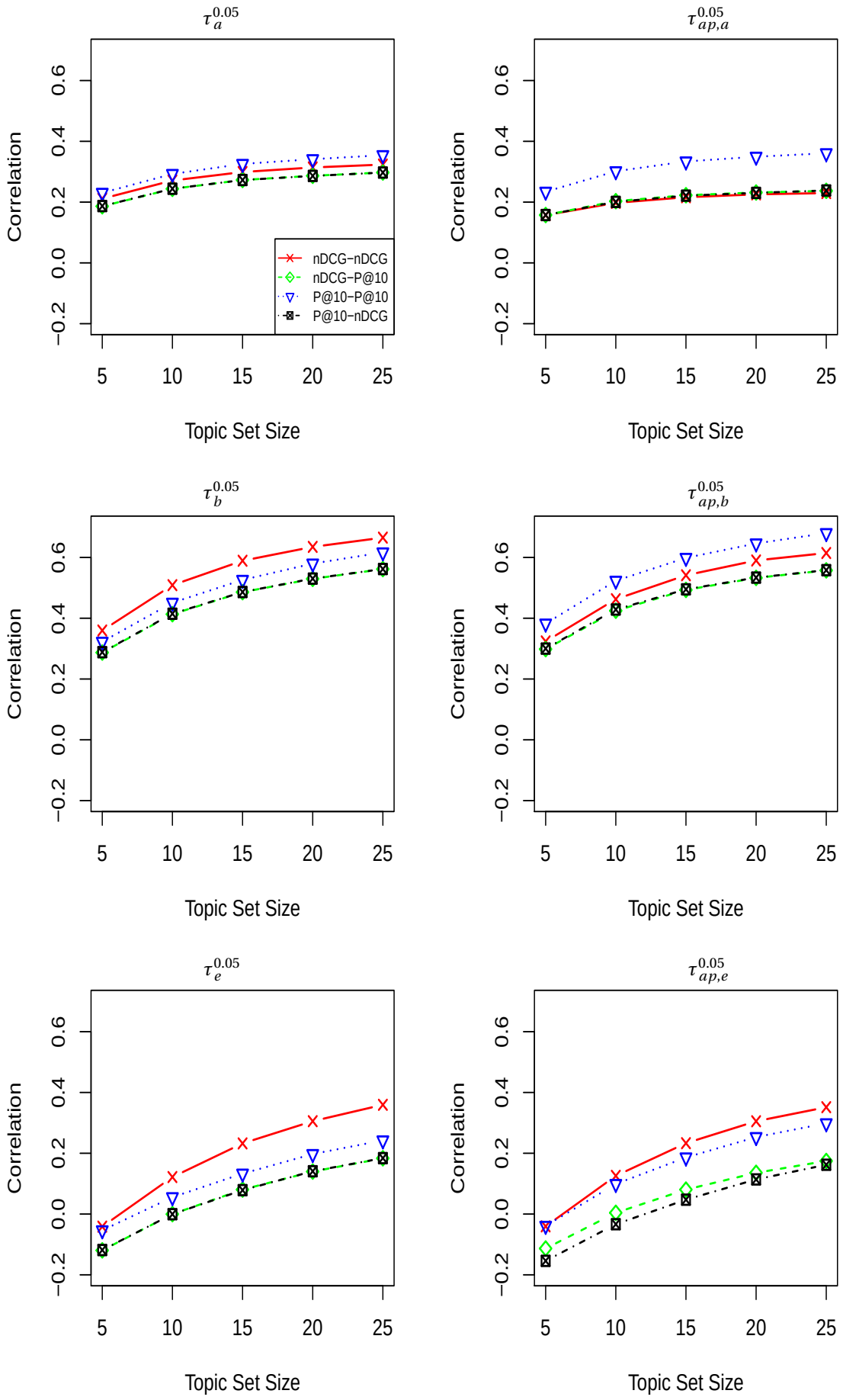

Figure 7.5: Predictive power of nDCG and P@10 of themselves and each other as a measure of $\tau_a^{0.05}$, $\tau_b^{0.05}$, $\tau_e^{0.05}$, $\tau_{ap,a}^{0.05}$, $\tau_{ap,b}^{0.05}$, $\tau_{ap,e}^{0.05}$ with different topic subset sizes

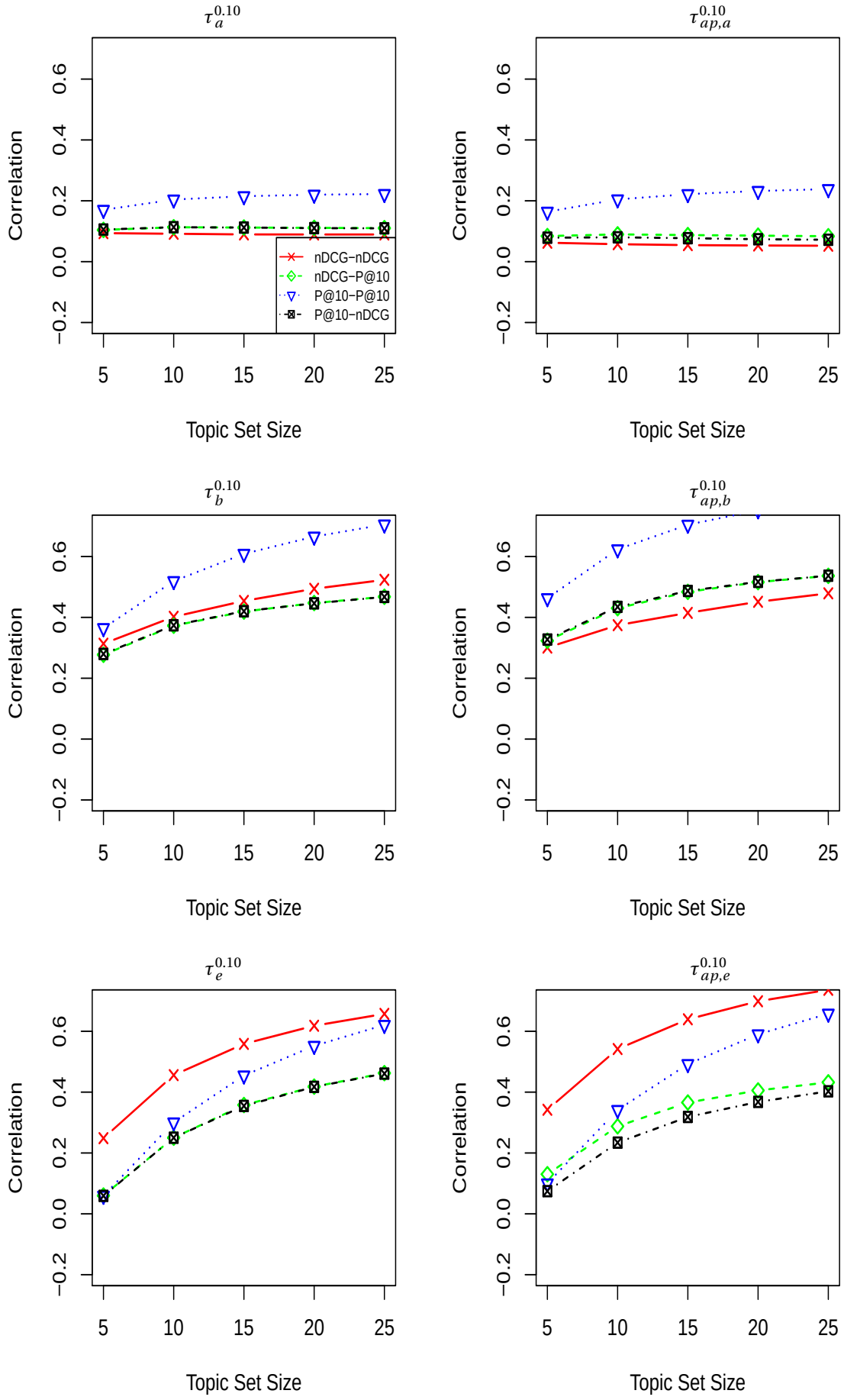

Figure 7.6: Predictive power of nDCG and P@10 of themselves and each other as a measure of $\tau_a^{0.10}$, $\tau_b^{0.10}$, $\tau_e^{0.10}$, $\tau_{ap,a}^{0.10}$, $\tau_{ap,b}^{0.10}$, $\tau_{ap,e}^{0.10}$ with different topic subset sizes

ments refer to not judged, judged as irrelevant and judged as relevant respectively. The experiment in [2] was carried out for p ∈ {1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100}. The mean system rankings over all topics with infAP were then assessed with the mean AP system rankings for the different pool sizes, using the Kendall's τ, linear correlation coefficient ρ, and root mean squared (RMS) error to measure how well infAP correlated with AP as shown in Figure. 7.7.

One should note that since we measure how well infAP corresponds to AP, this is the scenario of true vs. observer rankings. Therefore, in this thesis, for the system rankings for certain interesting topics, given in Figure. 7.8, the infAP with different pool sizes will be compared with the AP to depict the topic-variability, for all the different 'a'and 'e'variants proposed in this thesis.

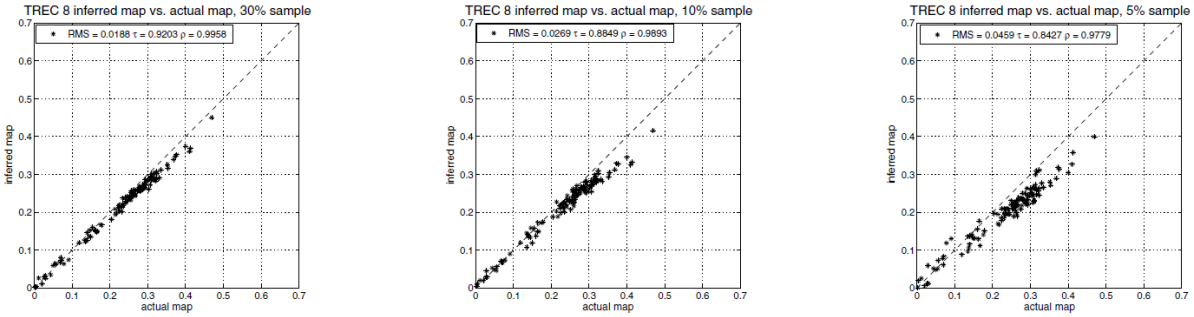

Figure 7.7: Figure 6 from Yilmaz et al [2] showing TREC-8 mean inferred AP as the judgement set is reduced to (from left to right) 30, 10, and 5 percent versus the mean actual AP.

7.2.2. TOPIC VARIABILITY OF SYSTEMS

For the sake of topic-wise comparison of systems, the correlations between infAP and AP system rankings for each topic and the mean of all these correlations are computed, in addition to the correlation between the mean infAP and AP. As shown by the charts in Figure. 7.8, the system rankings vary for the different topics. With increased pool sizes, infAP does not always correlate better with AP for the topics in comparison to the mean values from Figure. 7.7. Therefore, on topic level the systems may be performing poorly (overestimating or underestimating). So, it is necessary to compare the system rankings on topic level to know information, which was not revealed before when performing correlation on the mean system rankings (RoS) as given in Figure. 1.1. An interesting information that could be revealed by the topic comparisons is on which topics, a certain system performs poorer than its average performance over all topics etc. This will help in error analysis revealing on which topics, a certain system needs to improve. Moreover, a number of ties can be seen in Figure. 7.8 for Topic 37 but there was no mention of the ties in [2] despite the use of τ which is invalid in the presence of ties. It is concluded that the default R implementation of τ which resorts to τ_b was used, similar to the previous experiment given in subsection 7.1.2.

The charts given in Figure. 7.9 through 7.13 similar to the Figure.7 from [2], plot the correlation between mean infAP and mean AP system rankings over all topics (correlation(means)), mean of the correlations of system rankings for each topic (mean(correlations)), maximum (max(correlations)) and minimum (min(correlations)) topic correlation of the system rankings. Plotting the maximum and minimum correlations on topic level shows the extent of variability in the correlation scores of the system rankings on topic level, emphasizing the importance of topic level comparison of the system rankings.

As depicted in the Figure. 7.9 through 7.13, the system rankings for certain topics (shown as maximum correlation on topic level) are more concordant than the mean system rankings (correlation(means)), meaning that for the different pool sizes there are certain topics which are more interesting than the others as the system rankings are most correlated for these topics. It can be noted from the figures that the correlation of the mean system rankings over all topics, given as correlation(means), performed in the research by Yilmaz et al [2] is always overestimated in comparison to the mean of the per topic correlation of the system rankings i.e. mean(correlation). It is probable that this is the case in many other IR research articles where the standard averaging technique for Rank Correlation Analysis, which compares the RoS as given in Figure. 1.1, is adopted. In summary, when the system performances on all topics is given by correlation(means), estimating the system performances on individual topics, followed by its mean to account for all topics i.e. mean(correlations)

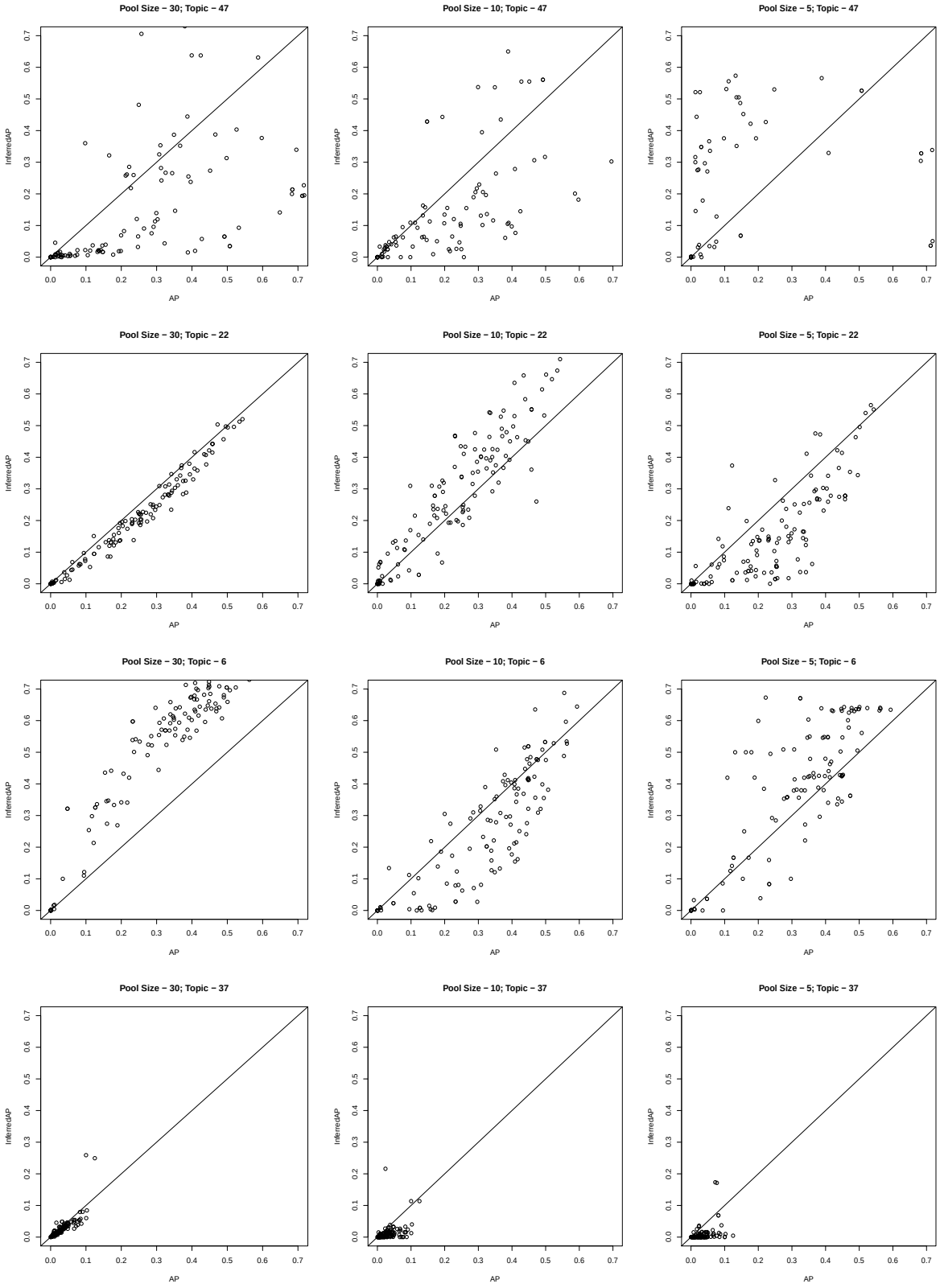

Figure 7.8: TREC-8 inferred AP as judgement set is reduced to 30, 10, 5 percent vs. actual AP for topics 47, 22, 6 and 37

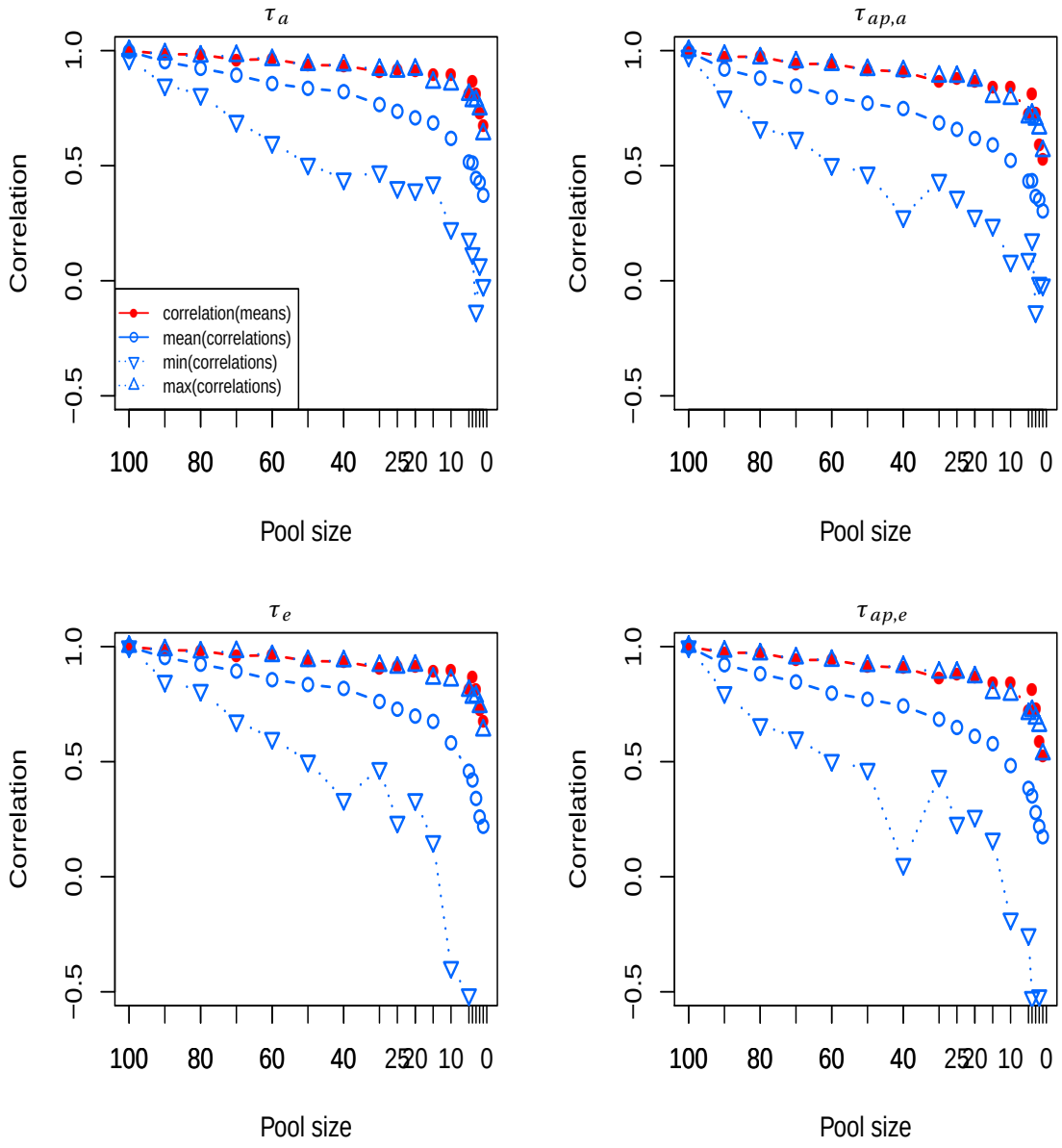

Figure 7.9: Change in a, e variants for inferred AP vs actual AP as the judgement sets are reduced.

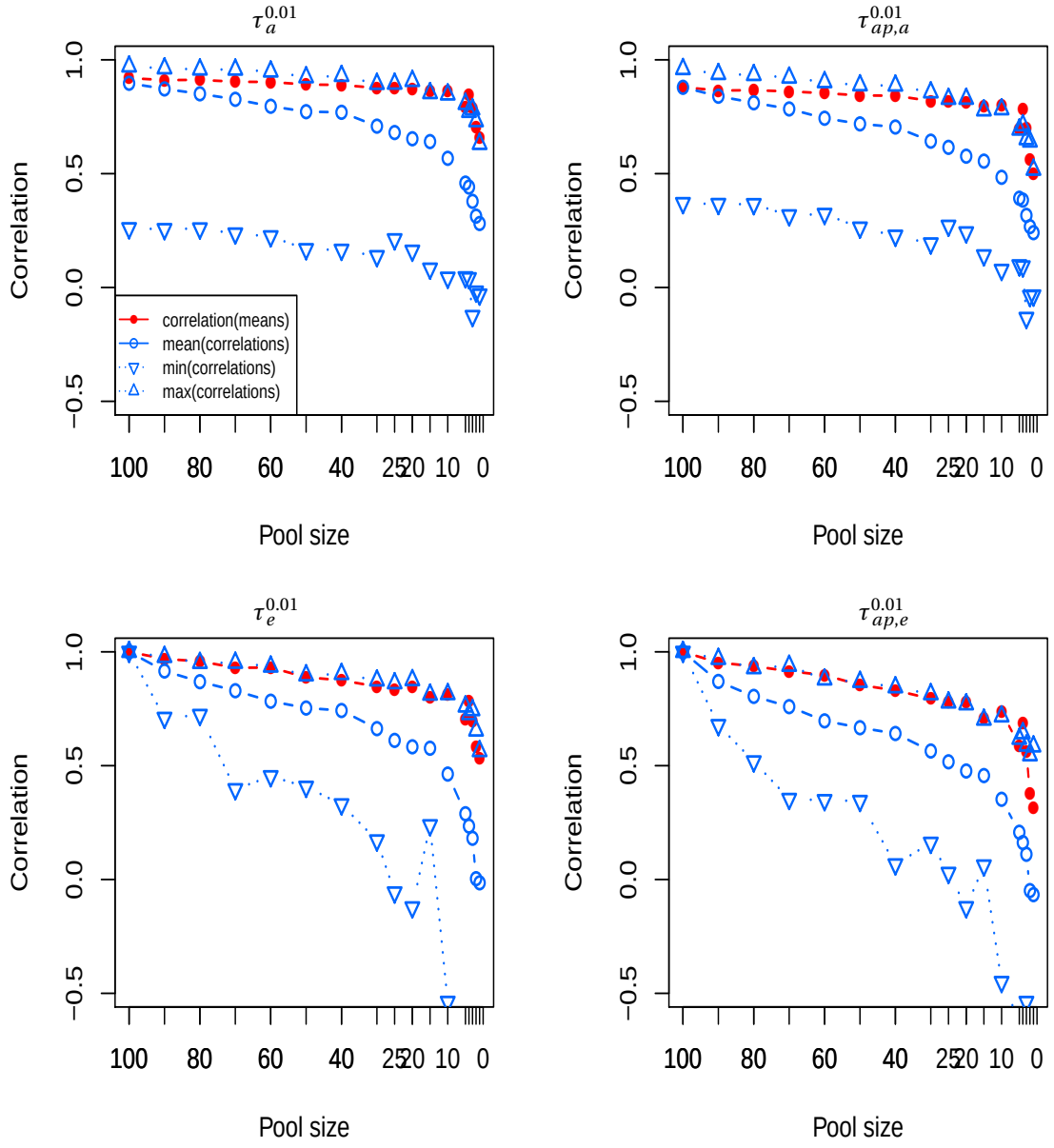

Figure 7.10: Change in a, e variants with threshold $w = 0.01$ for inferred AP vs actual AP as the judgement sets are reduced.

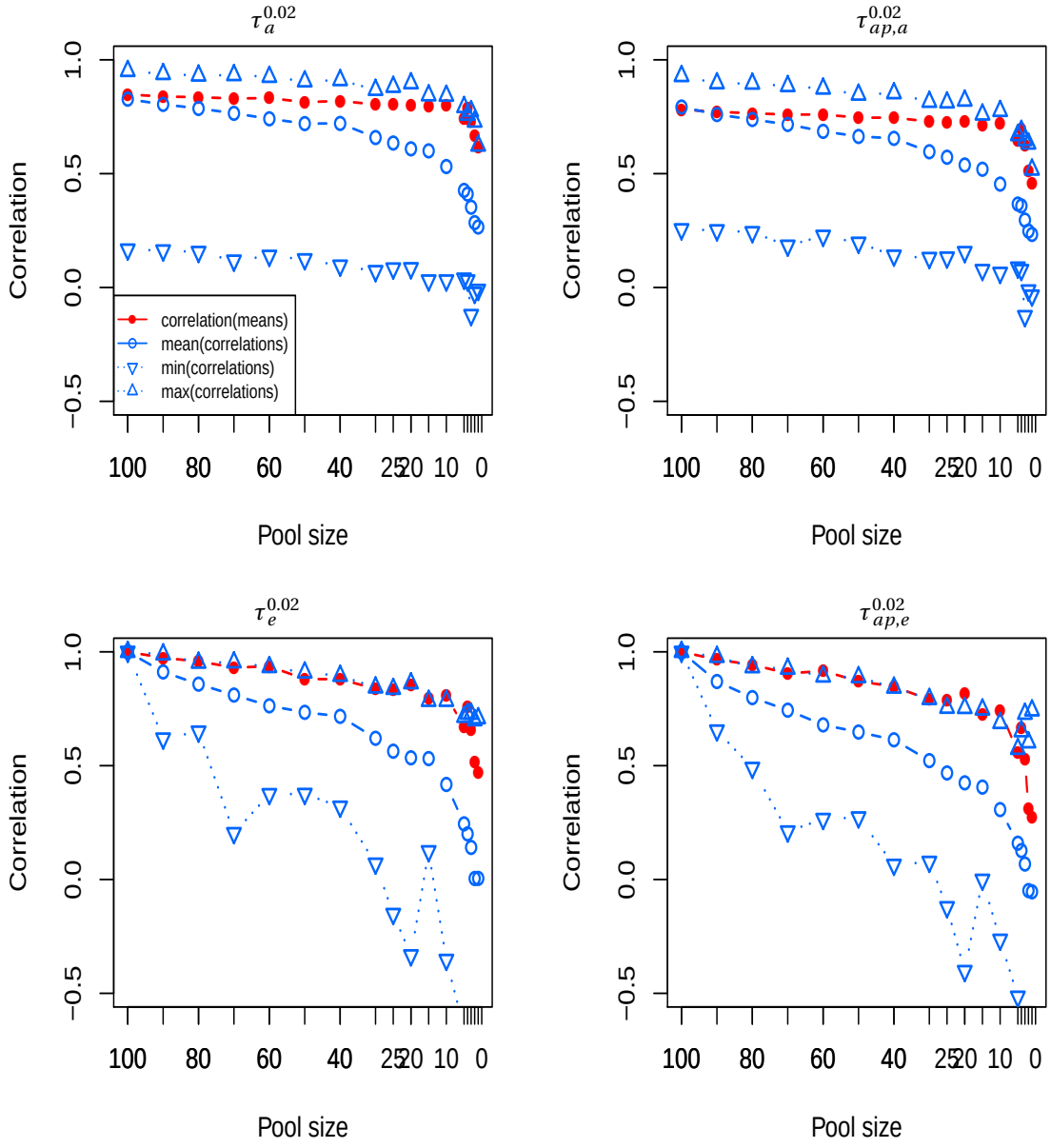

Figure 7.11: Change in a, e variants with threshold $w = 0.02$ for inferred AP vs actual AP as the judgement sets are reduced.

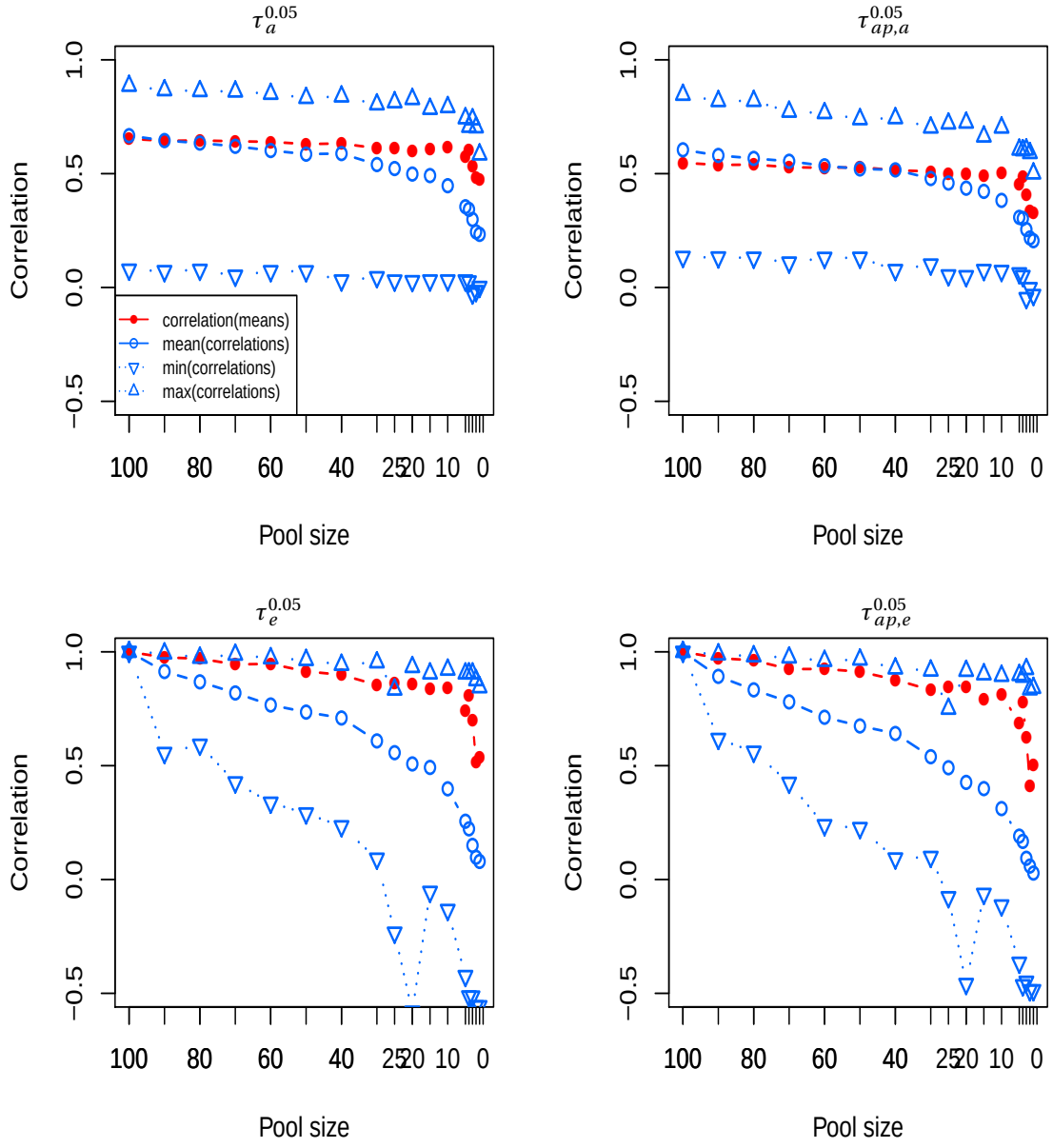

Figure 7.12: Change in a, e variants with threshold $w = 0.05$ for inferred AP vs actual AP as the judgement sets are reduced.

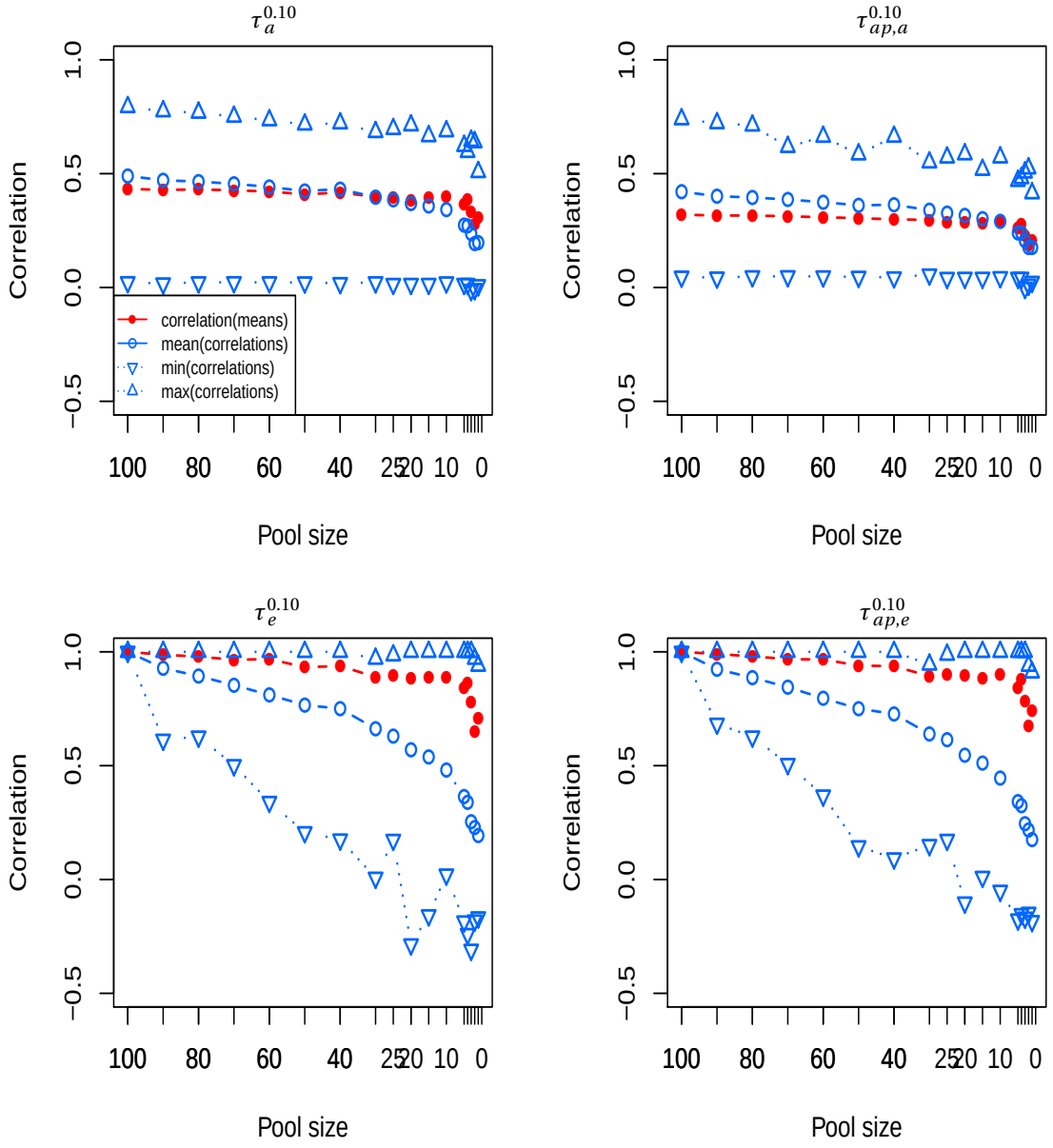

Figure 7.13: Change in a, e variants with threshold $w = 0.10$ for inferred AP vs actual AP as the judgement sets are reduced.

shows that the systems are in fact performing lower than expected by correlation(means). This is a compelling point to consider the topic level variations of the system as by the standard averaging technique (RoS), the systems are overestimated.

8

CONCLUSION

Since rank correlation coefficients are extensively used in IR, it is important to be able to distinguish the correct variant of the correlation coefficient applicable for the given scenario. The presence of ties in the rankings are also highly probable and the coefficient corresponding to the tie under consideration must be applied. For this reason, the different scenarios to be considered while computing correlation were extensively studied as a part of this thesis. As a result of which, different variants of the Kendall and AP correlation coefficients were formulated, applicable to the different scenarios studied. The appealing flexibility in introducing artificial ties by the person performing the correlation, different from the rankers, was also considered and customizable threshold values to determine whether items in a ranking are ties was also successfully formulated.

With the newly formulated variants of both τ and τ_{ap}, the necessity to continue to have these different variants was justified with a practical experiment to show that the different variants capture different information. Therefore, conclusions from previous research, may not hold on application of the different alternative variants of the correlation coefficients proposed. This was shown for the results of the experiment carried out by W. Webber et al in [1], in which it was concluded that reporting the simple evaluation measures is redundant. On repeating this experiment for the coefficients proposed in this thesis, we reach different conclusions. Hence, the practical assessment stresses the importance of using the correct coefficient based on the purpose for which the correlation is computed.

The shortcoming of the standard averaging of system performances over all topics, adopted so far in many IR research, to perform rank correlation analysis was experimentally shown to emphasize the need to compute the correlation on per topic level. The per topic level comparison of systems were not possible previously, due to the presence of ties in the IR system rankings which the correlation coefficients did not account for. With the different variants of τ and τ_{ap} coefficients formulated in this thesis to handle ties in rankings, per topic level comparisons are also made possible. Now by computing the correlation on topic level, we indeed observe different results than the correlation over the mean system scores for all topics. The topic level correlations between system rankings also helps in error analysis where systems can be compared on topic level to find out on which topics, a given system needs to be improved. This is highly beneficial to improve over-all system performances.

BIBLIOGRAPHY

[1] W. Webber, A. Moffat, J. Zobel, and T. Sakai, *Precision-at-ten considered redundant,* in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2008) pp. 695–696.

[2] E. Yilmaz and J. A. Aslam, *Estimating average precision with incomplete and imperfect judgments,* in *Proceedings of the 15th ACM international conference on Information and knowledge management* (ACM, 2006) pp. 102–111.

[3] M. Kendall, *Rank correlation methods* (C. Griffin, 1948).

[4] E. M. Voorhees, *Variations in relevance judgments and the measurement of retrieval effectiveness,* Information processing & management **36**, 697 (2000).

[5] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez, *Crawling a country: better strategies than breadth-first for web page ordering,* in *Special interest tracks and posters of the 14th international conference on World Wide Web* (ACM, 2005) pp. 864–872.

[6] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow, *Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval,* in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2005) pp. 512–519.

[7] I. Khennak, H. Drias, and H. Mosteghanemi, *An effective term-ranking function for query expansion based on information foraging assessment,* in *Mining Intelligence and Knowledge Exploration* (Springer, 2014) pp. 1–10.

[8] T. Sakai, *On the reliability of information retrieval metrics based on graded relevance,* Information processing & management **43**, 531 (2007).

[9] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas, *Million query track 2009 overview.* in *TREC* (2009).

[10] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas, *Do user preferences and evaluation measures line up?* in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2010) pp. 555–562.

[11] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz, *Relevance assessment: are judges exchangeable and does it matter,* in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2008) pp. 667–674.

[12] R. W. White, I. Ruthven, J. M. Jose, and C. Van Rijsbergen, *Evaluating implicit feedback models using searcher simulations,* ACM Transactions on Information Systems (TOIS) **23**, 325 (2005).

[13] J. Callan and M. Connell, *Query-based sampling of text databases,* ACM Transactions on Information Systems (TOIS) **19**, 97 (2001).

[14] R. Fagin, R. Kumar, and D. Sivakumar, *Comparing top k lists,* SIAM Journal on discrete mathematics **17**, 134 (2003).

[15] R. Kumar and S. Vassilvitskii, *Generalized distances between rankings,* in *Proceedings of the 19th international conference on World wide web* (ACM, 2010) pp. 571–580.

[16] G. M. Di Nunzio and G. Silvello, *A graphical view of distance between rankings: The point and area measures.* in *IIR* (2015).

[17] W. Webber, A. Moffat, and J. Zobel, *A similarity measure for indefinite rankings,* ACM Transactions on Information Systems (TOIS) **28**, 20 (2010).

[18] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Pearson correlation coefficient,* in *Noise reduction in speech processing* (Springer, 2009) pp. 1–4.

[19] N. Ferro, *What does affect the correlation among evaluation measures?* ACM Transactions on Information Systems (TOIS) **36**, 19 (2017).

[20] M. A. Woodbury, *Rank correlation when there are equal variates,* The Annals of Mathematical Statistics **11**, 358 (1940).

[21] Student, *An experimental determination of the probable error of dr spearman's correlation coefficients,* Biometrika , 263 (1921).

[22] J. Urbano and M. Marrero, *The treatment of ties in ap correlation,* in *ACM International Conference on the Theory of Information Retrieval* (2017) pp. 321–324.

[23] K. Sparck Jones, *Automatic indexing,* Journal of documentation **30**, 393 (1974).

[24] J. Urbano and M. Marrero, *How do gain and discount functions affect the correlation between dcg and user satisfaction?* in *European conference on information retrieval* (Springer, 2015) pp. 197–202.

[25] M. G. Kendall, *A new measure of rank correlation,* Biometrika **30**, 81 (1938).

[26] M. G. Kendall, *The treatment of ties in ranking problems,* Biometrika **33**, 239 (1945).

[27] E. Yilmaz, J. A. Aslam, and S. Robertson, *A new rank correlation coefficient for information retrieval,* in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2008) pp. 587–594.

[28] B. Carterette, *On rank correlation and the distance between rankings,* in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2009) pp. 436–443.