# Review of Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores*

Ziying Yang
ziyingy@student.unimelb.edu.au
692959

This paper explained that scoring functions of information retrieval (IR) systems may assign the same score for several documents, called ties, when retrieving and forming the ranking list. Documents receiving the same scores can be placed in distinct and arbitrary orders inside of the *tied block*, which may lead to fluctuations of scores evaluated by metrics for system performances. The authors of this paper intend to explore methods to compute the average score across all the permutations of possible orderings for the metrics Recall, Precision, F1, Average Precision (AP), Reciprocal Rank (RR) and Normalized Discounted Cumulative Gain (NDCG). These tie-aware measures will calculate expected evaluation scores of systems and so avoid situations such that some systems receive higher scores because the relevant documents are luckily ranked at the top and irrelevant ones ranked at the bottom of tied blocks.

The proposals in this paper are not substantial but still can be used for providing more accurate ranking of systems since the measures becomes tie-aware (although, it may not be necessary in some practical cases). There are several problems in this paper, including:

- The literature review fails to summarize the previous important work about considering and breaking ties and explained why creating tie-aware measures are worthy for IR evaluation.
- The value of their contributions that calculate the expected scores using tie-aware measures for IR evaluation lacks justification.
- The math notation is not well designed. Moreover, the tie-aware formula for RR is not correct based on probability theory.
- The work is straightforward and not terribly novel. The impact for IR evaluation results using their proposed methods is not tested or explained.

These issues are detailed considered in order as below:

Section 1 of this paper describes a lot about the background of IR evaluation but previous works about ties are only briefly reviewed. The impact of ties for IR evaluation is mentioned in the fifth paragraph, but there are no experiment results or cited conclusions that can strongly support the significance of authors' work.

An example of how TREC competitions dealing with ties is given in the sixth paragraph. Authors state that the ordering pick by TREC software (`trec_eval`) is arbitrary. But they may

*Frank McSherry and Marc Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In *Proc. ECIR*, pages 414–421, 2008

fail to look into the implementation of the software. Because in fact, even the system picks random orderings for tied documents, the software will still re-sort the documents with same scores by their document ID after reading the submitted input file. I suggest authors to review and compare the strategies of handling ties used by evaluation softwares developed for different text collections.

The last part of the introduction analyzes the simplicity of implementations and drawbacks of the proposed methods. But the reason why these methods have great advantages compared to the currently employed strategies is not discussed. In other words, they should have explained why computing the average of all permutations of orderings is worthy and better than randomly picking orderings for tied documents.

Although the idea of this paper is not quite novel, the proposed calculation formulas are useful. In section 2, authors describe the notations and formulas for different metrics. However, their notations are hard to be understood and not well explained. For example, in the first paragraph on the forth page, statements of vectors, elements and numbers are quite confusing. Authors denote the result vector $V_i$ as $\langle v_{t_i+1}, ..., v_{t_{i+1}} \rangle$, but in this set, giving only the first and the last elements accompanied by multiple levels of subscripts combined with computation symbols may cause misunderstandings. In section 2.4, the binding for $i$ and $j$ is stated neither by a compact and complete mathematical expression nor via detailed examples, which is difficult to be followed.

I suggest an alternative set of notations: denote documents in *equivalence class* (that is tie block, scores of documents inside of a block are same) $E_g$ as $[b_g...e_g]$ instead of using *tie-vector* $T$ defined in this paper, since only the first document $b_g$ and last document $e_g$ need to be noted in the context. And so we have $b_{g+1} = e_g + 1$. With these notations, cumulations in formulas described later can be explained clearer and simpler.

Authors should add equation numbers for all the formulas they describe in this paper.

For the Recall, Precision, F1 and NDCG, authors describe the principals of original metric formulas and then extend them to be tie-aware by identifying new equations for these metrics. Their methods are mostly based on the probability theory and defined correctly. However, when they propose the formula for AP in section 2.4, there is a unnecessary right bracket at the end of the formula for $AP@k(V)$. More than that, when $r_1 = n_1 = 1$ (that is, when the first document in the ranking list is relevant and not tied with any other document), the fraction $\frac{r_i - 1}{n_i - 1}$ equals to $\frac{0}{0}$ which is undefined. Therefore, a special case for this formula is required. Alternately, authors need to reform the formula.

The equation proposed for calculating the expected score of RR in section 2.5 is not correct. Denote $Pr$(A) as the probability that event A happens. The idea behind the method proposed by authors can be interpreted as below:

$$RR@k(V) = \sum_{j=t_i+1}^{min(t_{i+1},k)} \frac{1}{j} \times Pr \text{ (documents ranked before } j \text{ are irrelevant)} \quad (1)$$

$$\times Pr \text{ (documents ranked at } j \text{ is relevant)} \quad (2)$$

The fraction $f(x, r, n)$ defined by authors computes the $Pr$ (documents ranked before $j$ are irrelevant) in line (1) above. But the final bound equation:

$$RR@k(V) = \sum_{j=t_i+1}^{min(t_{i+1},k)} \frac{f(j - t_i, r_i, n_i)}{j} \quad (3)$$

for RR misses the probability described in line (2). Thus the bound formula for RR should have been:

$$RR@k(V) = \sum_{j=t_i+1}^{min(t_{i+1},k)} \frac{f(j-t_i, r_i, n_i)}{j} \times \frac{r}{n_i - (j-t_i)}. \quad (4)$$

In section 3, authors build implementations for both original metrics and their proposed tie-aware methods. They only measure the running time and memory used by two approaches and fail to prove that calculating the average score of all permutations, which takes longer execution time and greater memory, leads more accurate results for evaluating IR systems. Their experiments are too shallow and obvious to verify the value of proposed methods. There are neither table nor graph comparing experiment results. The test report is very brief and has few detail. A table containing the data of running time and memory used by both approaches for each metric mentioned in this paper will make authors' conclusions more convincing.

In addition, they should have tested whether rankings of systems will vary if the ordering of tied documents is ignored or not. The number of ties generated by systems is also a critical factor that need to be measured. Moreover, authors may need to construct graphs showing the best and worst scores that systems can receive regarding to the ordering of tied documents. The intersection of score ranges indicate the ranking variation of systems. And it is important to develop experiments using different well-known text collections with submitted inputs (systems) to show that this ranking variations occur for top systems as well (because top-weightedness is crucial for IR), which can testify the necessity of considering the ordering of tied documents for IR evaluation.