A graphical view of distance between rankings: The Point and Area measures (Extended Abstract)

Giorgio Maria Di Nunzio and Gianmaria Silvello

Department of Information Engineering – University of Padua
{dinunzio,silvello}@dei.unipd.it

Abstract. In Information Retrieval (IR), measuring the distance between rankings is a way for comparing evaluation measures and assess system rankings. In this paper, we present a variation of the Spearman foot rule which allows us to define two measures that have nice analytical and geometrical properties that can be effectively used to compare different rankings and to evaluate IR experiments. A Web application that shows how these measures behave from the graphical point of view is available at: `https://gmdn.shinyapps.io/ShinyPointArea/`

1 Introduction

Search engines effectiveness can be measured by analyzing their visible outcomes which are lists of documents ranked in descending order of relevance to a given topic. In this context, the correlation among rankings can be used to assess the search engines effectiveness; in fact, when one of the two rankings is the ideal one – i.e. the best obtainable result in a laboratory based evaluation – the correlation between two rankings becomes a measure of effectiveness. A high correlation between a ranking under evaluation and the ideal indicates a good behavior of the search engine being tested. Two standard de-facto measures of distance are the *Spearman foot rule* [6] and the *Kendall rank distance* [5]. Both measures are very easy to calculate and to interpret, but they lack some properties when it comes to evaluate search engines effectiveness. A review and a classification of ranking similarity measures (including extensions of the Spearman foot rule and the Kendall rank distance) has been presented by Webber et alii in [7]. This classification is based on two main properties: *conjointness* and *weightedness*. Conjointness is the property of dealing with complete (conjoint) or partial (non-conjoint) rankings; weightedness is the property of being able (weighted) or not being able (unweighted) to weight misplacements at the top of the list more than at its bottom.

In this paper, we present a variation of the Spearman foot rule leading to the definition of two new measures: a point-wise (qualitative) measure and an area-wise (quantitative) measure which can be classified as non-conjoint and unweighted. [1] Point-wise and area-wise measures present analytical and geometrical properties that can be effectively used to compare different rankings and to evaluate IR experiments. Furthermore, the point-wise measure provides an intuitive and effective graphical interpretation that can be used for performing a qualitative analysis on rankings comparison. Whereas, the area-wise measure can be used for a quantitative analysis given that its normalized version offers a simple measure of correlation between rankings. Moreover, as a by-product of this approach, we also obtain an original reformulation of the Kendall rank distance computed at each rank.

2 Methodological Approach

Given a set of documents $\mathcal{D} = \{d_1, \ldots, d_i, \ldots, d_n\}$, we consider two rankings r_α and r_β as two permutations without repetitions of $\mathcal{D}$. [2] We can use a method $idx_\alpha(d_i)$ to extract the index of document d_i within r_α. In general, the problem is to find the index of a document of r_α in the another ranking r_β. To this purpose, we define the function $\mathcal{F}_{\alpha,\beta}(k)$ as:

$$\mathcal{F}_{\alpha,\beta}(k) = idx_\beta(r_\alpha(k))) \tag{1}$$

This function translates the k-th document in r_α and returns the index of that document in the ranking r_β. For example, let $r_a = [d_2, d_1, d_4, d_3]$ and $r_b = [d_1, d_4, d_2, d_3]$ be two instances of r_α and r_β. Then, for $k = 1$, $\mathcal{F}_{a,b}(1) = idx_b(r_a(1)) = idx_b(d_2) = 3$.

The definition of Spearman footrule can be rewritten using Eq. 1 as:

$$S_{\alpha,\beta}(i) = \sum_{k=1}^{i} |\mathcal{F}_{\alpha,\beta}(k) - k| \ . \tag{2}$$

which is the total element-wise misplacements between the two lists r_α and r_β.

3 Point and Area Measures

The two measures we present in this paper derive from a variation of the Spearman foot rule. By removing the absolute value from the Spearman foot rule, we obtain the point-wise measure:

$$P_{\alpha,\beta}(i) = \sum_{k=1}^{i} (\mathcal{F}_{\alpha,\beta}(k) - k) \ . \tag{3}$$

[1] The point-wise and area-wise measures already contain a parameter for weighting the top and the bottom of a ranking list differently. For the purpose of this paper, we present the unweighted version of the measures.

[2] There are important research areas in IR which require distance measures on incomplete permutations (see Webber et alii [7], Fagin et alii [3] for Web search results and Angelini et alii [2] for search engine failure analysis), also known as the property of conjointness of a distance measure. For space reasons, we cannot address this issue in this paper and leave the solution of this problem to future works.

As a result, we make a distinction between negative and positive misplacements: when $\mathcal{F}_{\alpha,\beta}(k) > k$, we obtain a positive error because the document at rank k should have been ranked higher in r_β; when $\mathcal{F}_{\alpha,\beta}(k) < k$, we obtain a negative value because we find a highly relevant document lower in the list, which means that we are recovering a misplacement occurred earlier. Ultimately, the point-wise goes to zero when the last element of the list is computed. When the point-wise measure returns a positive number at any rank i, it means that there is still some non-recovered misplacement in r_b. On the other hand, every time $P_{\alpha,\beta}(i) = 0$ it means that the two rankings have retrieved the same elements at rank i. Compared to Spearman, the measure $P_{\alpha,\beta}(i)$ gives us some additional information about the distance between a given ranking and the ideal one at rank i: we can tell how far we are from the ideal ranking. However, the point-wise measure does not tell how bad a ranking is before rank i.

The area-wise measure considers the area formed by the segments between two adjacent points $P_{\alpha,\beta}(k-1)$ and $P_{\alpha,\beta}(k)$ and the x-axis. As an approximation of the area to be measured, we use a linear interpolation between adjacent points and we define the $A_{\alpha,\beta}(i)$ as the as a sum of all the trapezoids formed with the x-axis from rank 1 to i:

$$A_{\alpha,\beta}(i) = \sum_{k=1}^{i} \frac{\big(P_{\alpha,\beta}(k-1) + P_{\alpha,\beta}(k)\big)}{2} h \qquad (4)$$

where h is the height of each trapezoid, and $P_{\alpha,\beta}(0) = 0$ by assumption. The height h of each trapezoid can be a constant (for example $h = 1$), or it can be used as a tuning parameter for weighting errors differently according to the rank of the elements.

The area-wise measure can be used effectively in two ways: to accumulate the information about misplacements until rank i, to compare two or more rankings given a reference ranking (for example, the ideal one). Moreover, we can also study when the same type of misplacements occur in one ranking or another (earlier or later in the ranking), for example when two adjacent relevance degrees are exchanged.

It may be convenient to normalise the value returned by the area-wise between 0 and 1. The normalisation can be done by dividing the area of a relevance list at rank i by the largest obtainable area given by the worst possible ranking, that is the ideal ranking taken in the inverse order. The normalisation of the area-wise measure is defined as:

$$\mathrm{nA}_{\alpha,\beta} = \frac{A_{\alpha,\beta}}{A^*_{\alpha,\beta}} \qquad (5)$$

where $A^*_{\alpha,\beta}$ is the area of the worst possible ranking. The normalized area $\mathrm{nA}_{\alpha,\beta}$ defines the distance between β and α where $\mathrm{nA}_{\alpha,\beta} = 0$ means that β and α are the same ranking and $\mathrm{nA}_{\alpha,\beta} = 1$ means that they are inverse rankings one of the other. From this it is straightforward to derive the correlation coefficient as: $\text{A-corr}_{\alpha,\beta} = 1 - \mathrm{nA}_{\alpha,\beta}$.

4 Conclusion and Future Works

In this paper, we have introduced two new measures of distance among rankings: the point-wise and the area-wise measure. The point-wise measure was derived as a modification of the original Spearman foot rule, while the area-wise measure is built on top of the point-wise. The normalisation of the area-wise measure and the correlation coefficient derived from it – i.e. A-corr – is a measure of correlation between rankings. We have discussed some properties of these two measures in terms of qualitative analysis and quantitative analysis. For space reasons, we could not give a complete formalisation of the fact that both measures are metrics – i.e. the reflexivity, non-negativity, symmetry and triangle inequality properties hold. The area-wise measure already incorporates a parameter h that can be used to weight misplacements that happen at top ranks more than at low ranks; a possibility is to define h as the inverse of the index $h(i) = 1/i$. In this way, we would obtain a non-conjoint weighted distance measure. In addition, the parameter h can be used to model users search intents; for example, we could adjust the height h as a tuning parameter to represent a particular aspect of user behaviour such as the level of patience in the RBO measure [7].

Furthermore, we want to investigate the use of the point-wise curve and the A-corr measure as a full-fledged effectiveness measures. The behavior of the point-wise curve resembles the Cumulative Relative Position (CRP) [1, 4] one, but it presents several key differences such as the fact that the point-wise curve does not cross the x-axis whereas the CRP one does and that CRP is defined as an effort-oriented measure and not as a correlation one.

References

1. M. Angelini, N. Ferro, K. Järvelin, H. Keskustalo, A. Pirkola, G. Santucci, and G. Silvello. Cumulated Relative Position: A Metric for Ranking Evaluation. In *Proc. of the 3rd Int. Conf. of the CLEF Initiative (CLEF 2012)*, pages 112–123. LNCS 7488, Springer, Heidelberg, Germany, 2012.
2. M. Angelini, N. Ferro, G. Santucci, and G. Silvello. VIRTUE: A Visual Tool for Information Retrieval Performance Evaluation and Failure Analysis. *Journal of Visual Languages & Computing*, 2014.
3. R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top k Lists. In *Proc. of the 14th annual ACM-SIAM symposium on Discrete algorithms*, SODA '03, pages 28–36. Society for Industrial and Applied Mathematics, 2003.
4. N. Ferro, G. Silvello, H. Keskustalo, A. Pirkola, and K. Järvelin. The Twist Measure for IR Evaluation: Taking User's Effort Into Account. *Journal of the American Society for Information Science and Technology (JASIST)*, (in print).
5. M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, 1938.
6. C. Spearman. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103, 1904.
7. W. Webber, A. Moffat, and J. Zobel. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.*, 28(4):20, 2010.