A NOTE ON TOP-K LISTS: AVERAGE DISTANCE BETWEEN TWO TOP-K LISTS

Antoine Rolland

Univ Lyon, laboratoire ERIC, université Lyon 2, 69676 BRON,
antoine.rolland@univ-lyon2.fr

Résumé. Ce papier présente un complément à l'étude des top-k listes proposée par R. Fagin, R. Kumar and D. Sivakumar dans "Comparing top k lists" (J. Discrete Mathematics, 2003). Nous commenons par introduire quelques métriques de rang pour comparer des top-k listes, i.e. des classements où seuls les k premiers éléments sont pris en compte. Puis nous nous concentrerons sur la valeur maximale et la valeur moyenne prise par ces métriques sur l'ensemble des classements possibles.

Mots-clés. mesure de distance, top-k list, comparaison de rang.

Abstract. This short paper presents a complement to the top k list study proposed by R. Fagin, R. Kumar and D. Sivakumar in their paper "Comparing top k lists" (J. Discrete Mathematics, 2003). In this paper, we first introduce some ranking metrics which aim at comparing two top k lists, i.e. two rankings where only the first k elements are a matter of interest. Then we focus on the maximum values and the average values that can be obtained with these metrics.

Keywords. distance measures, top k list, rank aggregation

1 The top-k list problem

Fagin, Kumar and Sivakumar introduced in [1] a large study about the comparison of top k lists, i.e. ranking with only a subset of k members in the ordering, whereas the total number of potential members of the ordering is greater than k, and possibly infinite. However, the average value of the distance of two random top k lists is not included in this study. Therefore, we present in the paper a complement to the study in [1] focusing on these questions.

2 Ranking metrics

Characterizing the differences between two rankings is an old issue that have been particularly studied by Kendall. Different metrics have been proposed to determine the degree of similitude / difference between two rankings. Inspired by the previous works of Kendall [2] and Diaconis [3], we present here three of the best-known ranking correlation index.

- **Kendall's metric (Kendall's τ)**: the Kendall's τ index between two ranking R_1 and R_2 is obtained by counting the number of pairs of item which are ranked in opposite order in R_1 and R_2:

$$\tau_(R_1, R_2) = \\ \mid \{(i,j) : i < j, (r_1(i) < r_1(j) \wedge r_2(i) > r_2(j)) \vee (r_1(i) > r_1(j) \wedge r_2(i) < r_2(j))\} \mid$$

  where $r_1(i)$ and $r_2(i)$ are the rankings of the element i in R_1 and R_2 respectively. Usually these index is normalized by the total number of pairs, which is $n(n-1)/2$ with n the number of items in the ranking. Therefore the normalized index $\tau_N(R_1, R_2)$ is defined as follows:

$$\tau_N(R_1, R_2) = \\ \frac{|\{(i,j):i<j,(r_1(i)<r_1(j)\wedge r_2(i)>r_2(j))\vee(r_1(i)>r_1(j)\wedge r_2(i)<r_2(j))\}|}{n(n-1)/2}$$

  In [2] it is mentioned that the distribution of $\tau_N(R_1, R_2)$ is symmetric in the interval $[0, 1]$ and then that the average τ_N between two distribution is equal to 0.5.

- **Spearman's metrics (Spearman's ρ)**: the Spearman's ρ index between two ranking R_1 and R_2 is obtained by summing the rank difference in R_1 and R_2 for each element. Two variants of Spearman's ρ are possible:

$$\rho_{abs}(R_1, R_2) = \sum_{i=1}^{n} |r_1(i) - r_2(i)|$$

$$\rho_{sqr}(R_1, R_2) = \sum_{i=1}^{n} (r_1(i) - r_2(i))^2$$

  where $r_1(i)$ and $r_2(i)$ are the rankings of the element i in R_1 and R_2 respectively.

  As in the Kendall's τ case, we can defined normalized Spearman's ρ_N dividing ρ by the maximum available score.

  In [3] are introduced the following values for the maximum values:

  - $max_{abs} = \begin{cases} (2m)^2 \text{ where } n = 2m \\ (2m)^2 + 2m \text{ where } n = 2m + 1 \end{cases}$
  - $max_{sqr} = \frac{1}{3}(n^3 - n)$

  In [3] are also introduced the following values for the average values of ρ:

  - $\overline{\rho}_{abs} = \frac{1}{3}(n^2 - 1)$
  - $\overline{\rho}_{sqr} = \frac{1}{6}(n^3 - n)$

3 Average distance value for top k lists

3.1 Top k list definition

Following [1], we now define top k list, i.e. ranking when we only have the top k members of the ordering. As proposed in [1], a top k list R is a bijection from a domain D (intuitively, the members of the top k list) to $[k]$. In our paper, we extend this definition assuming that D is a subset of a discrete and possibly infinite set N of size $n \in \mathbb{N} \cup +\infty$. In order to formalize this presentation of a top k list into a set of n elements, we then choose what is called the "optimistic approach" in [1] and suppose that all the $n-k$ elements that are not in the top k list are ranked *ex aequo* at the $k+1$ position. Therefore, a top k list R is a bijection from N to $\{1, 2, 3..., k, k+1, ..., k+1\}$.

3.2 Computing distance between two top k lists

In [1], Fagin *et al.* propose several options to compute a distance between two top k lists, based either on Kendall's τ or Spearman's footrule. We introduce here three different distances:

- **Kendall** τ: as previously stated, the chosen option is to keep what is called the "optimistic approach" in [1] as adaptation of Kendall's metric for top k lists. Let $N = \{1, \dots n\}$ and R_1 and R_2 two top k ranking on N.

$$d_{Kendall}(R_1, R_2) = \sum_{\{i,j\} \in N} K_{i,j}(R_1, R_2)$$

  where

  - $K_{i,j}(R_1, R_2) = 0$ if i and j appear in the same order in R_1 and R_2
  - $K_{i,j}(R_1, R_2) = 1$ if i and j appear in the opposite order in R_1 and R_2
  - $K_{i,j}(R_1, R_2) = 0$ if both i and j appear in position $k+1$ in a ranking (i.e. not in the top k) and in positions ahead $k+1$ in the other ranking.

- **Absolute Spearman's** ρ: the absolute Spearman's ρ index between two ranking R_1 and R_2 is obtained by summing the absolute rank difference in R_1 and R_2 for each element, again with each element not ranked in the first k having a rank equal to $k+1$.

$$\rho_{abs}(R_1, R_2) = \sum_{i=1}^{n} |r_1(i) - r_2(i)|$$

- **Squared Spearman's** ρ: the squared Spearman's ρ index between two ranking R_1 and R_2 is obtained by summing the squared rank difference in R_1 and R_2 for each element, again with each element not ranked in the first k having a rank equal to $k+1$.

$$\rho_{sqr}(R_1, R_2) = \sum_{i=1}^{n} (r_1(i) - r_2(i))^2$$

3.3 Computing the maximum distance value between two top k lists

The maximum distance value between two top k lists from a set n of size n will be denoted $\max(d_{n,k})$. It is easy to compute as it is reached in the case of two opposite ranking. These values should be interesting in the case of a normalization of the distance values in a range between 0 and 1. If $n \geq 2k$, the maximum distance value does not depends on n, and is equal to $2k$ times the average distance of an element of rank $k+1$ to an element of rank $1, \ldots, k$.

	$\max(d_{n,k})$
Kendall	$k(k+1)$
Spearmann abs	$k(k+1)$
Spearmann sqr	$k(k+1)(2k+1)/3$

Table 1: Different maximum values ($n \geq 2k$)

3.4 Computing the average distance value between two top k lists

The aim of this section is to determine the average value $\overline{d}_{n,k}$ on the set of all possible couples of top k lists (R_1, R_2) on a set N of size n. We suppose in the following that $n \geq 2k$, i.e. the top k list is really a "top", and not just a small subset. Without loss of generality we can choose the canonical order $R = (1, 2, 3..., k, k+1, ..., k+1)$ for R_1, and compute $\overline{d}_{n,k}$ as the average value of $d(R, R')$, R' being a top k ranking on $\mathcal{X}$.

The average distance value between two top k lists depends of n, size of the set N, and k. Therefore we will denote $\overline{d}_{n,k}$ the average value of the distance between two top k lists in a set of n elements. In order to compute $\overline{d}_{n,k}$, we propose to divide it into several cases depending on the number of common values in the top k elements of R and R'. Let i be the number of elements of $\mathcal{X}$ which are in the first k elements of R but are not in the first k elements of R'. We will then consider all the different cases for i varying from 0 (the two top k lists have the same elements) to k (there is no common elements in the two top k lists). This last case is always possible as we supposed that $n \geq 2k$.

For a specific triplet (n, k, i), the number of different possible rankings, denoted by $N(n, k, i)$ is

$$N(n, k, i) = \binom{k}{i}\binom{n-k}{i}k!$$

There are $\binom{k}{i}$ different ways to find i elements into the k elements which are in the top-k of list R; there are also $\binom{n-k}{i}$ different ways to find i elements into the $n-k$ elements which are not in the top-k of list R. There are $k!$ different ways to rank k elements.

For a specific couple (k, i), the average distance between two top k lists with $k-i$ common elements, denoted by $\overline{d}_{k,i}$ is given by

$$\overline{d}_{k,i} = 2i.\overline{d'}_{k+1} + (k-i)\frac{\overline{d}_k}{k}$$

where $\overline{d'}_{k+1}$ is the average distance between element number $k+1$ and elements $1, \ldots, k$, and $\overline{d}_k$ is the average distance between two ranking of k elements. It is because there are i elements that are in the top k of R and not in R' and i elements that are in the top k of R' and not in R, with both an average distance of $\overline{d'}_{k+1}$ between their respective positions in R and R'. And there are $(k-i)$ elements that are both in top k list of R and R', with an average distance of $\frac{\overline{d}_k}{k}$ between their respective positions in R and R'.

The table 2 gives the formulas of $\overline{d'}_{k+1}$ and $\overline{d}_k$ for different ranking metrics. The average distance between two top k lists of set of n elements is then:

$$\overline{d}_{n,k} = \frac{\sum_{i=0}^{k} \overline{d}_{k,i} N(n, k, i)}{\sum_{i=0}^{k} N(n, k, i)}$$

	$\overline{d'}_{k+1}$	$\overline{d}_k$
Kendall	$\frac{k+1}{2}$	$\frac{k(k-1)}{2}$
Spearmann abs	$\frac{k+1}{2}$	$\frac{1}{3}(k^2-1)$
Spearmann sqr	$\frac{(k+1)(2k+1)}{6}$	$\frac{1}{6}(k^3-k)$

Table 2: Different average values

3.5 Average values for some k

In order to give examples, we propose in tables 3, 4 some values of $\overline{d}_{n,k}$ for $k = 3$, $k = 5$, and $k = 10$.

$k=3$	n							
	6	8	10	20	30	50	100	∞
Kendall	7.5	69/8	9,3	10.65	11.1	11.46	11.73	12
Spearmann abs	22/3	8.5	9.2	10.6	332/30	11.44	11.72	12
Spearmann sqr	16	19	20.8	24.4	25.6	26.56	27.28	28

Table 3: Different average values for $k = 3$

$k=5$	n						
	10	15	20	30	50	100	∞
Kendall	20	70	25	80/3	28	29	30
Spearmann abs	19	68/3	24.5	79/3	27.8	28.9	30
Spearmann sqr	65	80	87.5	95	101	105.5	110

Table 4: Different average values for $k = 5$

References

[1] R. Fagin, R. Kumar, D. Sivakumar, Comparing top k lists, J. Discrete Mathematics 17 (1) (2003) 134–160.

[2] M. Kendall, J. D. Gibbons, Rank Correlation Methods, Edward Arnold, London, 1990.

[3] P. Diaconis, Number 11 in ims lecture series, in: Group Representation in Probability and Statistics, Institute of Mathematical Statistics, 1988.