

CS599: Structure and Dynamics of Networked Information (Spring 2005)

03/02/2005: Rank Aggregation

Scribes: Nupur Kothari and Ranjit Raveendran

Last class, we stated and proved that there is no social choice function satisfying all the four conditions of monotonicity, non-triviality, independence of irrelevant alternatives, and non-dictatorship (Arrow's Theorem 1951). We then moved away from the axiomatic approach to rank aggregation, and started considering it as an optimization problem instead. We looked at two distance metrics on orderings: Spearman's Footrule, and Kendall's τ .

- **Kendall's τ** is the number of adjacent transpositions to move \langle_1 to \langle_2 . This is identical to the number of adjacent swaps performed by BubbleSort when transforming \langle_1 to \langle_2 .
- **Spearman's Footrule F** is the sum of distances in positions over all elements between \langle_1 and \langle_2 . That is, it is the L_1 distance of the vectors whose i^{th} entry is the position of the i^{th} element in the respective ordering.

It is easy to see that the two metrics are not the same. One obvious example is when $\langle_1 = 1\ 2$, and $\langle_2 = 2\ 1$. Then, the Spearman Footrule distance is $F = 2$, while the Kendall distance is $\tau = 1$. While the two can be different, we can prove that they are not *very* different.

Lemma 1 For any orderings \langle_1 and \langle_2 , we have $\tau(\langle_1, \langle_2) \leq F(\langle_1, \langle_2) \leq 2\tau(\langle_1, \langle_2)$.

Proof. We first show that $F(\langle_1, \langle_2) \leq 2\tau(\langle_1, \langle_2)$. We do this by induction on the value of $\tau(\langle_1, \langle_2)$. In the base case, when $\tau(\langle_1, \langle_2) = 0$, both the orderings are the same, and hence $F(\langle_1, \langle_2) = 0$.

In the induction step, we look at \langle_1 and \langle_2 such that $\tau(\langle_1, \langle_2) > 0$. Let \langle' be obtained from \langle_2 by one switch towards \langle_1 . Then, $\tau(\langle_1, \langle') = \tau(\langle_1, \langle_2) - 1$, and $\tau(\langle', \langle_2) = 1$, $F(\langle', \langle_2) = 2$. By the Triangle Inequality, applied to the metric F , we have that $F(\langle_1, \langle_2) \leq F(\langle_1, \langle') + F(\langle', \langle_2) = F(\langle_1, \langle') + 2\tau(\langle', \langle_2)$. So we can apply the Induction Hypothesis to \langle_1 and \langle' , obtaining that $F(\langle_1, \langle') \leq 2\tau(\langle_1, \langle')$. Thus,

$$F(\langle_1, \langle_2) \leq F(\langle_1, \langle') + 2\tau(\langle', \langle_2) \leq 2\tau(\langle_1, \langle') + 2\tau(\langle', \langle_2) = 2\tau(\langle_1, \langle_2)$$

completing the inductive proof.

For the other inequality, $\tau(\langle_1, \langle_2) \leq F(\langle_1, \langle_2)$, we use induction on $F(\langle_1, \langle_2)$. In the base case $F(\langle_1, \langle_2) = 0$, both the orderings are the same, so $\tau(\langle_1, \langle_2) = 0$.

For the inductive step, we have the problem that simple switches will not necessarily improve the value of F . As an example, we can look at $\langle_1 = 1\ 2\ 3\ 4$ and $\langle_2 = 4\ 3\ 2\ 1$. In this case, a switch towards \langle_1 will result in the ordering $\langle' = 4\ 3\ 1\ 2$, which has $F(\langle_1, \langle') = F(\langle_1, \langle_2)$. Thus, here we try to "meta-swap" two elements such that one is too far to the left of its position in \langle_1 and the other is too far right of its position in \langle_1 .

Without loss of generality, we may assume that $\langle_1 = 1 \dots n$ (i.e., the elements are sorted), and \langle_2 has element i in position a_i . Thus $F(\langle_1, \langle_2) = \sum_i |a_i - i|$.

Let i be maximal such that $a_i \neq i$, i.e., the rightmost element that is out of position. Notice that this implies that $a_i < i$, for otherwise, the element a_i (the one that is in position a_i in the ordering \langle_1) would be a larger index out of place. Let $j \leq a_i$ be the largest index with $a_j > a_i$ and $a_j > j$, i.e., the rightmost element to the left of i in the order \langle_2 which is too far right. Notice that such an element j must exist, as only $a_i - 1$ of the elements $1, \dots, a_i$ can be in positions $1, \dots, a_i - 1$. Also notice that $a_j \leq i$, as otherwise, the element a_j , which is out of position, would show that i is not maximal.

Let \langle' be the ordering obtained by swapping i and j in the ordering \langle_2 . Notice that in \langle' , neither i nor j “overshoot” their actual positions, because $j \leq a_i$, and $a_j \leq i$. Thus, both i and j move $|a_i - a_j|$ positions closer to their destination, and all other elements stay in position. We obtain that $F(\langle_1, \langle_2) = 2|a_i - a_j| + F(\langle_1, \langle')$. Applying the Induction Hypothesis, we thus have that $F(\langle_1, \langle_2) \geq 2|a_i - a_j| + \tau(\langle_1, \langle')$.

We can also calculate the Kendall τ distance between \langle_2 and \langle' quite easily. By making $|a_i - a_j|$ switches to the right of element i , we get it into position. Then, another $|a_i - a_j| - 1$ switches to the left of element j move it to position a_i . All elements in between are moved once to the left and once to the right, and thus end up in the same position. Hence, we just proved that $\tau(\langle', \langle_2) \leq 2|a_i - a_j| - 1$. Now, by the Triangle Inequality for the metric τ , we obtain that

$$\tau(\langle_1, \langle_2) \leq \tau(\langle_1, \langle') + \tau(\langle', \langle_2) \leq \tau(\langle_1, \langle') + 2|a_i - a_j| - 1 < F(\langle_1, \langle_2).$$

This completes the inductive proof. ■

Now that we have metrics to measure the difference between two orderings, given a list of several orderings $\langle_1, \langle_2, \dots, \langle_k$, we want to find an ordering \langle “close to all” of them. There are multiple concrete optimization criteria we could be trying to minimize, for example:

1. The average τ distance $\frac{1}{k} \sum_i \tau(\langle, \langle_i)$.
2. The average Footrule distance $\frac{1}{k} \sum_i F(\langle, \langle_i)$.
3. The maximum τ distance $\max_i \tau(\langle, \langle_i)$.
4. The maximum Footrule distance $\max_i F(\langle, \langle_i)$.

Of these, the first criterion has the nice property that an ordering minimizing it guarantees to satisfy the extended Condorcet property.

Lemma 2 *If \langle minimizes $\sum_i \tau(\langle, \langle_i)$, then it satisfies the extended Condorcet property.*

Proof. We prove the lemma by contradiction. Assume that we have an ordering \langle which minimizes $\sum_i \tau(\langle, \langle_i)$, but does not satisfy the extended Condorcet property. Then, there is a partition (S, \bar{S}) of the alternatives $1, \dots, n$ such that every alternative in S beats all those in \bar{S} ¹, yet \langle ranks some $j \in \bar{S}$ ahead of some $i \in S$. Therefore, there must be such a pair such that i and j are adjacent in the ordering \langle . Now, swapping i and j improves $\sum_i \tau(\langle, \langle_i)$ by i 's margin of victory. Since the relative orderings of no other elements are affected, the objective function actually improves, contradicting the optimality of \langle . ■

Unluckily, minimizing the average τ distance is NP-hard, as proved by Dwork et al. [1] (we don't give the proof here).

Fact 3 *If $k \geq 4$, then minimizing $\sum_i \tau(\langle, \langle_i)$ is NP-hard.*

Instead, we may focus on the second objective function: minimizing $\sum_i F(\langle, \langle_i)$. This one can actually be minimized in polynomial time. The intuition is to look at an ordering/permutation as a matching between elements and positions. Then, we can express the total objective function value as a sum of “penalties” incurred by each element for the position it is in. Specifically, we assign a penalty $\phi_{j,p} = \sum_i |p - \langle_i(j)|$ for putting element j in position p (recall that $\langle_i(j)$ denotes the position in which element j appears in the order \langle_i). For each j , we thus need to assign a (distinct) $p(j)$ to minimize $\sum_j \phi_{j,p(j)}$. Since feasible assignments are exactly perfect matchings between elements and positions, our goal is to find the *cheapest* perfect matching in the complete bipartite graph where the edge (j, p) has cost $\phi_{j,p}$. Finding minimum-cost perfect matchings is known to be solvable in polynomial time, for instance by first computing any perfect matching, and then repeatedly eliminating negative cycles in the residual graph.

¹Recall that i is said to *beat* j if there are more orderings in which i precedes j than vice versa.

Notice that the ordering minimizing the average Footrule distance is also a 2-approximation to the problem of minimizing the average Kendall's τ distance. For if OPT denotes the best Kendall ordering, and \langle the best Footrule ordering, then

$$\sum_i \tau(\langle, \langle_i) \leq \sum_i F(\langle, \langle_i) \leq \sum_i F(\text{OPT}, \langle_i) \leq 2 \sum_i \tau(\text{OPT}, \langle_i).$$

In all of the previous discussion, we assumed that the orderings were actually complete. What happens when we are not given full orderings? What techniques can we then use to compare these orderings? Fagin et al. [2] study the issue of how to compare top k lists. If the lists don't all contain all of the elements, one needs to extend (or replace) the distance notions defined above.

Fagin et al. suggest several techniques to deal with this. One is to augment all the lists such that all the elements appear in all the lists, by appending the missing elements at the end of each list (since they were clearly not considered to be in the top k by that list). This raises the question in which order the extra elements should be appended to the lists. An "optimistic" view would define the distance between lists based on the assumption that the missing elements appear in the same relative order as in the other list. Another solution is to append the elements in a random order, and define the distance as the average. [2] shows that neither of these approaches for pairwise comparison of lists satisfies the triangle inequality, and hence, we do not obtain a metric. However, they show that the distance measures can be both upper and lower bounded by a metric.

References

- [1] Cynthia Dwork and S. Ravi Kumar and Moni Naor and D. Sivakumar, *Rank aggregation methods for the Web*, WWW, 2001.
- [2] R. Fagin and R. Kumar and D. Sivakumar, *Comparing top k lists*, SIAM Journal on Discrete Mathematics 17/1, 2003.