THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 7, Number 2/2006, pp. 000-000

ON THE BEHAVIOR OF CERTAIN METRIC ON THE PERMUTATIONS GROUP

Liviu P. DINU

University of Bucharest, Faculty of Mathematics and Computer Science
Academiei 14, 010014, Bucharest, Romania
E-mail: ldinu@funinf.cs.unibuc.ro

We investigate the expected and the maximum values of the rank distance on the permutations group. We compute these values and show when the maximum value is reached.

1 INTRODUCTION

When a new metric is introduced, often there is a "hidden variable" in the similarity relation (frequently depends on the specific area of research), so that we should always speak of similarity with respect to some property, and there is a plethora of measures in part because researchers are often inexplicit on this point.

On the other hand, one should have some knowledge about the nature of the problem to be solved. A result that is mathematically sound may be highly implausible and might not reflect what is known about the analyzed process.

In information sciences, in computational linguistics or in some automatically categorization problems (especially based on combining rankings) we have to take into account the natural tendency of the objects to place the most important information in the first part of the message. So, if the differences between two objects are at the top (i.e., in essential points), the distance has to have a bigger value then when the differences are at the bottom of the objects. A similar situation can be found in genomics, where the difference on the first positions between two codons is more important than the difference on the last positions (Marcus, 1974).

Following the upper motivations, we have introduced the rank distance [Dinu, 2003] as a similarity measure on rankings with linguistics and biological motivations. In some related papers we have investigated the behavior of this metrics in topics like aggregation of classifications, categorization [Dinu 2003], computational linguistics (especially the similarity of Romance languages, [Dinu and Dinu 2005]), the DNA sequence comparison [Dinu and Sgarro 2006], the similarity of trees structures [Dinu 2005]. From a computational point of view, rank distance has a good behavior w.r.t. the so-called "median string problem", i.e. the median string can be computed in a polynomial time by using rank distance [Dinu and Manea, 2006] (we remind that the same problem is NP-hard via other metrics, like Levenshtein, Kendall, etc.)

In many cases (computational linguistics, statistics, coding, classification task, etc.), it is enough to use metrics that work on string without repetitions (i.e. rankings or permutations).

In this paper we investigate some mathematical properties of the rank distance related on its expected and maximum values on the permutations group. We compute these values and show when they are reached.

2 PRELIMINARIES

Let $U = \{a, b,\ldots,z\}$ be a nonempty finite set, called the *universe of objects*. A *full ranking* over U is a one-to-one map $f : \{1,2,\ldots,|U|\} \to U$, which associate a rank to each object from U.

Recommended by Solomon MARCUS, member of the Romanian Academy

A *partial ranking* of length *i* is an injective map $f:\{1,2,\ldots,i\}\rightarrow U$, $i\leq|U|$, which associate a rank to each object from a subset of U. We denote by |*f*| the length of a partial ranking *f*. We denote by *V* (*f*) the set {*f*(1), *f*(2),…,*f*(*i*)}.

Given a (partial) ranking *f*, we associate a order *ord*(*x*|*f*) to each object *x* from this ranking as it follows:

$$ord(x\,|\,f)=|\,f\,|+1-f^{-1}(x) \tag{1}$$

Follows the upper motivations, we define the rank distance between two rankings as it follows:

Definition 1 *Let U be an alphabet and let f, g two partial rankings over U. The rank distance between f and g is defined by:*

$$\Delta(f,g)=\sum_{x\in V(f)\cap V(g)}|\,ord(x\,|\,f)-ord(x\,|\,g)\,|+\sum_{x\in V(f)\setminus V(g)}ord(x\,|\,f)+\sum_{x\in V(g)\setminus V(f)}ord(x\,|\,g) \tag{2}$$

Theorem 1 (Dinu, 2003) Δ *is a metric.*

One may be concerned in the motivation for the usage of objects' order in a ranking, instead of the rank itself. This motivation comes from two directions. First, we consider that the distance between two rankings should be greater if they differ at their top (on the high ranked objects), since in many applications the low ranked objects are neglected; consequently the objects with high ranks should have a greater weight. Second, the length of the rankings is also important: if a ranking is longer we consider that the criterion that produced it performed a more profound analyze of the objects, hence, it is more reliable than another criterion that produced a shorter ranking. Consequently, although, for example, two rankings of different length may have the same object on the first position, there is a difference between that object's orders, and, this difference should be reflected in the total distance.

From equation (1) and Definition 1 we obtain an alternative formula for rank distance:

$$\Delta(f,g)=\sum_{x\in V(f)\cap V(g)}||\,f\,|-f^{-1}(x)-|\,g\,|+g^{-1}(x)\,|+\sum_{x\in V(f)\setminus V(g)}||\,f\,|+1-f^{-1}(x)\,|+\sum_{x\in V(g)\setminus V(f)}||\,g\,|+1-g^{-1}(x)\,| \tag{3}$$

Remark 1 *If* |*f*|=|*g*| *and V* (*f*) = *V* (*g*) *(i.e. f and g are rankings of the same subset of objects), then*

$$\Delta(f,g)=\sum_{x\in V(f)}|\,f^{-1}(x)-g^{-1}(x)\,| \tag{4}$$

3 RANK DISTANCE ON PERMUTATIONS

Without loss of generality, we can assume that *U* ={1,2,…,n}. Let *f* and *g* be two full rankings over *U*.

Since *f* and *g* are, in this case, permutations of degree *n*, relation (4) gives us enough reasons to investigate the rank distance as a measure of similarity on permutations.

3.1 Max rank distance

In this subsection we investigate the following problems: given a permutation σ from S_n, which are the permutations τ from S_n such that the rank distance between σ and τ is maximum? Which is this maximum?

Let *n* be a positive number and let $\sigma\in S_n$ be a permutation (we denote by S_n the group of permutations with n elements). We say that an integer from σ is *low* if its position is $\leq n/2$ and it is *high* if its position is $>n/2$.

Let Θ_σ be as following:

$$\Theta_\sigma=\{\tau\in S_n\,|\,\forall x\in\{1,\ldots,n\}, x\ is\ low\ in\ \tau\ iff\ \ x\ is\ high\ in\ \sigma\ and\ viceversa\} \tag{5}$$

Proposition 1 *For each* $\sigma \in S_n$ *and every two permutations* τ , π *in* Θ_σ *we have:* $\Delta(\sigma,\tau)=\Delta(\sigma,\pi)$.

Proof: We have:

$$\Delta(\sigma,\tau)=\sum_{x\ is\ high\ in\ \sigma}|ord(x|\sigma)-ord(x|\tau)|+\sum_{x\ is\ low\ in\ \sigma}|ord(x|\sigma)-ord(x|\tau)|=$$
$$=\sum_{x\ is\ high\ in\ \sigma}(ord(x|\sigma)-ord(x|\tau))+\sum_{x\ is\ low\ in\ \sigma}(ord(x|\tau)-ord(x|\sigma))=$$
$$=\sum_{i=\frac{n}{2}}^{n}i-\sum_{x\ is\ low\ in\ \tau}ord(x|\tau)+\sum_{x\ is\ high\ in\ \tau}ord(x|\tau)-\sum_{i=1}^{\frac{n}{2}}i= \tag{6}$$
$$=\sum_{i=\frac{n}{2}}^{n}i-\sum_{x\ is\ low\ in\ \pi}ord(x|\pi)+\sum_{x\ is\ high\ in\ \pi}ord(x|\pi)-\sum_{i=1}^{\frac{n}{2}}i=\Delta(\sigma,\pi)$$

Lemma 1 (Dinu, 2003b) *If a and b are two real numbers, a > b, then the function f : R*$\rightarrow$ *R, f(x)* = |x-b|-|x-a| *is not decreasing.*

Proposition 2 *Let* $\sigma \in S_n$ *be a permutation. The maximum rank- distance is reached by the permutation* τ *where ord (x|*τ *) = n + 1- ord (x|*σ *) ,* $\forall x \in \{1,2,...,n\}$. *Under these conditions, the maximum rank distance is:* $\max_{\tau\in S_n}\Delta(\sigma,\tau)=\frac{n^2}{2}$ *if n is even, respectively* $\frac{n^2-1}{2}$ *if n is odd.*

Proof: Assume without loss of generality that σ = e_n (identical permutation of S_n).

We suppose that there is an $i \in V(\tau)$ such that:

1. ord(i|σ)=n+1-i
2. ord(i|τ)≠ *i.*

Under these circumstances, there is $j \in V(\tau)$ such that ord(j|τ)=*i* and ord(i|τ)=*x*.

We have three cases:

- Case 1. *x=j*;
- Case 2. *i<x*;
- Case 3. *i>x*.

Case 1. *x = j*; If we interchange the elements *i* and *j* in τ , we obtain a new permutation τ' such that $\Delta(e_n,\tau')\geq\Delta(e_n,\tau)$.

$$\begin{array}{lllllllll} e_n=(1 & 2 & \dots & i & \dots & j & \dots & n) \\ \tau=\ (\tau(1) & \tau(2) & \dots & \tau(n+1-j)=i & \dots & \tau(n+1-i)=j & \dots & \tau(n)) \\ \tau'=(\tau(1) & \tau(2) & \dots & \tau'(n+1-j)=j & \dots & \tau'(n+1-i)=i & \dots & \tau(n)) \end{array}$$

It is enough to prove that:

$$|(n+1-i)-j|+|(n+1-j)-i|\leq|(n+1-i)-i|+|(n+1-j)-j| \tag{7}$$

Since

$$|(n+1-i)-i|+|(n+1-j)-j|\geq|2(n+1)-2(i+j)|,$$

(7) is proved.

Case 2. *i < x*. We suppose that *i < x*. If we interchange the elements *i* and *j* in τ , we obtain a new permutation τ' such that $\Delta(e_n,\tau')\geq\Delta(e_n,\tau)$.

It is enough to proof that:

$$|(n+1-i)-x|+|(n+1-j)-i| \le |(n+1-i)-i|+|(n+1-j)-x|,$$

which is equivalent to:

$$|i-(n+1-j)|-|i-(n+1-i)| \le |x-(n+1-j)|-|x-(n+1-i)| \qquad (8)$$

Since $n+1-j < n+1-i,$ by using Lemma1, we obtain (8).

Case 3. $i > x$. $\exists k$ such that $ord(k \mid \sigma) = n+1-k$ and $ord(k \mid \tau) > k$. We denote $ord(k \mid \tau) = y$. Let k be the first element that satisfies the above condition. Because $ord(k \mid \tau) > k$, there is l such that $ord(l \mid \tau) = k$. We will show that $l > k$. If $l < k$, we are in the following situation:

$$ord(l \mid \sigma) = n+1-l, \quad ord(l \mid \tau) = k > l.$$

But this is a contradiction with the hypothesis (k is the first element in σ for which $ord(k \mid \tau) > k$). So, we have reduced this case to the precedent one.

Proposition 3 *For each* $\sigma \in S_n$ *and every two permutations* τ, π *such that* $\pi \in \Theta_\sigma$ *and* $\tau \notin \Theta_\sigma$, *we have:* $\Delta(\sigma,\tau) < \Delta(\sigma,\pi)$.

Proof: Since $\tau \notin \Theta_\sigma$, there is $i \in High(\sigma)$ such that $i \notin Low(\tau) \Rightarrow i \in High(\tau)$. So, there is $j \in Low(\tau)$ such that $j \in Low(\sigma)$. Let $x = ord(i \mid \tau)$ be the rank of i in τ and let $y = ord(j \mid \tau)$ be the rank of j in τ. We interchange i and j in τ, obtaining a new permutation τ'. So, $ord(i \mid \tau') = y$ and $ord(j \mid \tau') = x$.

$$|ord(i \mid \sigma) - ord(i \mid \tau)| + |ord(j \mid \sigma) - ord(j \mid \tau)| < |ord(i \mid \sigma) - ord(i \mid \tau')| + |ord(j \mid \sigma) - ord(j \mid \tau')|$$

$$|ord(i \mid \sigma) - x| + |ord(j \mid \sigma) - y| < |ord(i \mid \sigma) - y| + |ord(j \mid \sigma) - x| \qquad (9)$$

which is equivalent to:

$$|y - ord(j \mid \sigma)| - |y - ord(i \mid \sigma)| < |x - ord(j \mid \sigma)| - |ord(i \mid \sigma) - x| \qquad (10)$$

Since (10) is true from Lemma 1, it result that $\Delta(\sigma,\tau) < \Delta(\sigma,\tau')$.

Theorem 2 *For a given permutation* σ *the maximum rank distance is achieved by all permutations from* Θ_σ.

Proof: It results from Propositions (1), (2) and (3).

3.2 On the average rank distance

Here we investigate the problem of finding the average value of the rank-distance on the group of permutations S_n.

The average rank-distance of S_n is defined by:

Definition 2 *We denote the average rank distance of Sn by* $E_\Delta(S_n)$ *and it is defined as it follows:*

$$E_\Delta(S_n) = \frac{1}{\binom{n!}{2}} \sum_{\sigma \in S_n} \sum_{\tau \in S_n} \Delta(\sigma,\tau).$$

Remark 2 *An equivalent definition is:*

$$E_\Delta(S_n)=\frac{1}{n!}\sum_{\sigma\in S_n}\Delta(\sigma,S_n).$$

Example 1 *For n* = 3, *we have: S*₃ = {(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)}. *So,* $E_\Delta(S_3)=\frac{1}{6}16=2.66$.

We are interested by an analytical formula of the average value of the rank distance.

Lemma 2 $\Delta(\sigma,S_n)=\Delta(\tau,S_n)$, *for every two permutations* $\sigma,\tau\in S_n$.

Using this result, we obtain that the average value is equal to:

$$E_\Delta(S_n)=\frac{1}{n!}\sum_{\sigma\in S_n}\Delta(e_n,S_n), \tag{11}$$

where $e_n=(1,2,\ldots,n)$.

Theorem 3 *The rank distance between* S_n *and its identical permutation* e_n *is equal to:*

$$\Delta(e_n,S_n)=\frac{(n+1)!(n-1)}{3} \tag{12}$$

Proof: For an easy comprehension of our proof we'll use a matrix *M* with *n* rows and *n*! columns, constructed as it follows: each column represents one of the *n*! permutations of S_n. They are lexicographically ordered:

$$M=\begin{pmatrix} 1 & 1 & 1 & \cdots & n & n & n \\ 2 & 2 & 2 & \cdots & n-1 & n-1 & n-1 \\ 3 & 3 & 3 & \cdots & n-2 & n-2 & n-2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n-2 & n-2 & n-1 & \cdots & 2 & 3 & 3 \\ n-1 & n & n-2 & \cdots & 3 & 1 & 2 \\ n & n-1 & n & \cdots & 1 & 2 & 1 \end{pmatrix}$$

We can see that each number appears in the matrix exactly (*n*-1)! times on first position, (*n*-1)! times on second one, and so on, on (*n*-1)! times on the *n*'th position.

We have:

$$\Delta(e_n,S_n)=\underbrace{(n-1)![0+1+2+\ldots+(n-1)]}_{\sum_{\sigma\in S_n}|ord(1|e_n)-ord(1|\sigma)|}+\underbrace{(n-1)![1+0+1+2+\ldots+(n-2)]}_{\sum_{\sigma\in S_n}|ord(2|e_n)-ord(2|\sigma)|}+$$

$$+\underbrace{(n-1)![2+1+0+1+2+\ldots+(n-3)]}_{\sum_{\sigma\in S_n}|ord(3|e_n)-ord(3|\sigma)|}+\ldots+\underbrace{(n-1)![(i-1)+(i-2)+\ldots.+1+0+1+2+\ldots+(n-i)]}_{\sum_{\sigma\in S_n}|ord(i|e_n)-ord(i|\sigma)|}+$$

$$+\ldots+\underbrace{(n-1)![(n-1)+(n-2)+\ldots+1+0]}_{\sum_{\sigma\in S_n}|ord(n|e_n)-ord(n|\sigma)|}=(n-1)!\sum_{i=1}^{n}[(i-1)+(i-2)+\ldots+1+0+1+2+\ldots+(n-i)]=$$

$$=(n-1)!\sum_{i=1}^{n}\left[\frac{i(i-1)}{2}+\frac{(n-i)(n-i+1)}{2}\right]=\frac{(n-1)(n+1)!}{3}$$

Corollary 1 *The average value of the rank distance is equal to:*

$$E_{\Delta}(S_n) = \frac{(n-1)(n+1)}{3}.$$

Proof: We have:

$$E_{\Delta}(S_n) = \frac{1}{n!}\sum_{\sigma \in S_n} \Delta(e_n, S_n) = \frac{1}{n!}\frac{(n-1)(n+1)!}{3} = \frac{(n-1)(n+1)}{3}$$

4 CONCLUSIONS

In this paper we computed the maximum and the average values of the rank distance on the group of permutations S_n. Both values are useful in practical problems which involved rank distance when it is stringent to compare a quantitative value to the maximum or average value. Usually, this is realized by normalizing the rank distance.

ACKNOWLEDGEMENTS

MEdC-ANCS and CNR-NATO have supported this research.

REFERENCES

1. S.C. Chan, A.K.C. Wong and D.K.Y. Chiu, *A survey of multiple sequence comparison methods,* Bulletin of Math. Biology, Vol. 54 n4 (1992) pp 563-598.
2. T.H. Cormen, C.E. Leiserson, R.R. Rivest, *Introduction to Algorithms*, MIT Press, 1990.
3. M. Deza, T. Huang. *Metrics on permutations. A survey*, J. Combinatorics, Information and System Sciences, 23(1998), 173-185
4. P. Diaconis, R.L. Graham, *Spearman footrule as a Measure of Disarray*, Journal of Royal Statistical Society. Series B (Methodological), Vol. 39, No. 2(1977), 262-268.
5. L.P. Dinu, *On the classification and aggregation of hierarchies with different constitutive elements*, Fundamenta Informaticae, 55(1), 39-50, 2003b.
6. L.P. Dinu, *Rank distance with applications in similarity of natural languages*, Fundamenta Informaticae, 65(1-4), 135-149, 2005.
7. A. Dinu, L.P. Dinu. *On the Syllabic Similarities of Romance Languages*. Lecture Notes in Computer Science, Volume 3406, pp. 785-789, 2005
8. L.P. Dinu, A. Sgarro. A low-complexity distance for DNA strings, Fundamenta Informaticae (to appear)
9. C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, *Rank aggregation methods for the web,* Proceedings of the 10th International World Wide Web Conference, 2001, 613622.
10. C. de la Higuera, F. Casacuberta, *Topology of strings: Median string is NP- complete,* Theoretical Computer Science, 230:39-48, 2000.
11. Kohonen, *Median strings,* Pattern Recognition Letters, 3:309-313, 1985.
12. Marcus, S. Linguistic structures and generative devices in molecular genetics. *Cahiers Ling. Theor. Appl.,* 11, 77-104, 1974.

Received May 22, 2006