Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes

Alonso Gragera [a,d], Vorapong Suppakitpaisarn [a,b,c,*]

[a] Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, 113-0033 Tokyo, Japan
[b] Global Research Center for Big Data Mathematics, National Institute of Informatics, 2-1-2 Hitotsubashi Chiyoda-ku, 101-8430 Tokyo, Japan
[c] ERATO Kawarabayashi Large Graph Project, Japan Science and Technology Agency, Japan
[d] International School for Postgraduate Studies, University of Granada, Spain

ARTICLE INFO

Article history:
Received 20 May 2016
Received in revised form 13 December 2016
Accepted 1 January 2017
Available online xxxx

Keywords:
Distance
Metric
Near-metric
Approximate triangle inequality
Sørensen–Dice index
Tversky index

ABSTRACT

In this work, we calculate a tight relaxed triangle inequality ratio for some of the most well-known indexes used in finding dissimilarities between two finite sets known as the Sørensen–Dice and Tversky indexes. This relaxed triangle inequality ratio affects efficiency and approximation ratios of recent algorithms for many combinatorial problems such as traveling salesman and nearest neighbor search. Because of that, there are many works providing ratios for several other indexes. In this work, we focus on the Tversky index, which is a generalization of many dissimilarity indexes commonly used in practice. We provide the tight ratio of the Tversky index in this paper. Because the Sørensen–Dice index is a special case of the Tversky index, we know from the results that the tight ratio for the Sørensen–Dice index is equal to 1.5.

1. Introduction

In this work, we focus on dissimilarity indexes between two finite sets. Many well-known ones are special cases of the Tversky index [1,2]. Let α, β be a non-negative rational number no larger than 1. The Tversky dissimilarity between a finite set A and a finite set B, $d^T_{\alpha,\beta}(A, B)$, can be defined as follows:

$$d^T_{\alpha,\beta}(A, B) := 1 - \frac{|A \cap B|}{|A \cap B| + \alpha \cdot |A \backslash B| + \beta \cdot |B \backslash A|}.$$

When $\alpha > \beta$, each element in $A \backslash B$ contributes more to the value of Tversky index than each element in $B \backslash A$. In that case, each element in $A \backslash B$ is less expected and should be given more importance than an element in $B \backslash A$ [3].

The Tversky dissimilarity index was proposed to be used on psychological experiments [1,2]. However, it is also often used on many other research fields including document or image retrieval [4,5], software engineering [6,7], and cheminformatics [8,9]. Also, it is a generalized form of the most commonly used set dissimilarity indexes, Jaccard–Tanimoto [10,11] and Sørensen–Dice [12,13]. When $\alpha = \beta = 1$, the Tversky index is equal to the Jaccard–Tanimoto index, and when $\alpha = \beta = 0.5$, the Tversky index is equal to the Sørensen–Dice index.

* Corresponding author at: Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, 113-0033 Tokyo, Japan.
E-mail addresses: alonso@is.s.u-tokyo.ac.jp (A. Gragera), vorapong@is.s.u-tokyo.ac.jp (V. Suppakitpaisarn).

http://dx.doi.org/10.1016/j.tcs.2017.01.004
0304-3975/

Please cite this article in press as: A. Gragera, V. Suppakitpaisarn, Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes, Theoret. Comput. Sci. (2017), http://dx.doi.org/10.1016/j.tcs.2017.01.004

When the Tversky index is equal to Jaccard–Tanimoto index ($\alpha = 1$ and $\beta = 1$), it was shown by Lipkus that the dissimilarity is metric [14], and when the Tversky index is equal to the Sørensen–Dice index ($\alpha = 0.5$ and $\beta = 0.5$) it is know to be a near-metric [15]. Aside from the in $\alpha = 1$ and $\beta = 1$ case, Tversky dissimilarity index is not a metric, that is because the dissimilarity does not satisfy the triangle inequality [16], i.e. there exists A, B, C such that $d^T_{\alpha,\beta}(A, C) > d^T_{\alpha,\beta}(A, B) + d^T_{\alpha,\beta}(B, C)$. For example, when $A = \{1\}$, $B = \{1, 2\}$ and $C = \{3\}$, we have

$$d^T_{\alpha,\beta}(A, C) = 1 - \frac{0}{0 + \alpha + \beta} = 1,$$
$$d^T_{\alpha,\beta}(A, B) = 1 - \frac{1}{1 + \beta} \leq \frac{1}{2},$$
$$d^T_{\alpha,\beta}(B, C) = 1 - \frac{1}{1 + \alpha} \leq \frac{1}{2}.$$

When $\alpha < 1$ or $\beta < 1$, we know that either $d^T_{\alpha,\beta}(A, B)$ or $d^T_{\alpha,\beta}(B, C)$ is strictly less than $\frac{1}{2}$, and $d^T_{\alpha,\beta}(A, B) + d^T_{\alpha,\beta}(B, C) < 1 = d^T_{\alpha,\beta}(A, C)$. Besides, the index is not symmetric when $\alpha \neq \beta$, i.e. there exists A, B such that $d^T_{\alpha,\beta}(A, B) \neq d^T_{\alpha,\beta}(B, A)$.

If the index does not satisfy the triangle inequality, we cannot use algorithms proposed for metric indexes. However, recently, there have been several algorithms [17–19] proposed for near-metrics (semi-metric indexes that satisfy the ρ-relaxed triangle inequality for some $\rho > 1$). A dissimilarity index d satisfies the inequality, if for any finite sets A, B, C,

$$d(A, C) \leq \rho \left(d(A, B) + d(B, C)\right).$$

Those algorithms are more efficient when the value of ρ is smaller. Knowing the value of ρ for a specific dissimilarity index can help in analyzing the efficiency of algorithms, when they are applied with that index. In this paper, we will refer to the value of ρ in the previous inequality as the *relaxed triangle inequality ratio*.

1.1. Our contribution

In section 2, we show that the relaxed triangle inequality of the Tversky index, $d^T_{\alpha,\beta}$ is equal to $\frac{\sqrt{1/(\alpha\cdot\beta)}+1}{2}$, i.e.

$$d^T_{\alpha,\beta}(A, C) \leq \frac{\sqrt{\frac{1}{\alpha\cdot\beta}} + 1}{2} \left(d^T_{\alpha,\beta}(A, B) + d^T_{\alpha,\beta}(B, C)\right).$$

Then, in section 3, we show that this ratio is tight. For any rational number c no larger than $\frac{\sqrt{1/(\alpha\cdot\beta)}+1}{2}$, we give A, B, C such that $d^T_{\alpha,\beta}(A, C) = c\left(d^T_{\alpha,\beta}(A, B) + d^T_{\alpha,\beta}(B, C)\right)$.

From the results, we know that the tight ratio is $\frac{\frac{1}{\alpha}+1}{2}$ when $\beta = \alpha$, and the tight ratio for the Sørensen–Dice index is 1.5. A part of the results in this work have been published in the proceeding of WALCOM 2016 [20]; in particular, a restricted version proof for the case where $\alpha = \beta$.

1.2. Implications of our results

The application of the relaxed triangle inequality ratio in computer science is firstly discussed in [21]. In that paper, the authors calculate the ratio for the dissimilarity in shapes of an image database system named IBM's QBIC, and discuss about how to use it for image shape retrieval. After their initial proposal, there are several works calculating the ratio for other dissimilarities (cf. [22]).

Other than the image shape retrieval, the relaxed triangle inequality ratio affects efficiency of many algorithms for traveling salesman problem (TSP). The approximation ratio of an algorithm proposed by Andreae [23] is $\rho^2 + \rho$ when ρ is the ratio. This approximation ratio is later improved to 4ρ in [24] for near-metrics where $\rho > 3$. Many works used TSP results on a graph of which each node is represented by a set and weight of each edge is defined by a set dissimilarity between two sets incident to it [25]. Our results can be applied to each of those algorithms.

Many online clustering algorithms' approximation ratio also depends on the relaxed triangle inequality ratio. In [26], Zhang et al. define a cost for each clustering, and show that an upper bound of the optimal cost depends on the relaxed triangle inequality ratio value. Furthermore, an algorithm for streaming k-means problem with approximation ratio equal to $3\rho + 1$ is proposed in [17] by Braverman et al., and an approximation algorithm for online k-median problem is proposed in [18] by Mettu and Plaxton. Beside the approximation ratio, some polynomial-time approximation schemes (PTAS) where the computation time depends on the value of ρ is given in [19,27]. Clustering algorithms based on indexes considered in this paper have been intensively considered in many machine learning applications (e.g. [28,29]).

Discussed in [8], all Tversky indexes such that $\alpha = \beta$ are known to be invariant, i.e. there is an increasing function f such that $d^T_{\alpha,\alpha}(A, B) = f\left(d^T_{\alpha',\alpha'}(A, B)\right)$ for any set A, B and any $0 \leq \alpha, \alpha' \leq 1$. Because of that, those indexes, which include Jaccard–Tanimoto and Sørensen–Dice indexes, are equivalent in applications such as ranking or nearest neighbor search.

Please cite this article in press as: A. Gragera, V. Suppakitpaisarn, Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes, Theoret. Comput. Sci. (2017), http://dx.doi.org/10.1016/j.tcs.2017.01.004