The College Mathematics Journal

ISSN: 0746-8342 (Print) 1931-1346 (Online) Journal homepage: http://www.tandfonline.com/loi/ucmj20

A Concise Proof of the Triangle Inequality for the Jaccard Distance

Artur Grygorian & Ionut E. Iacob

To cite this article: Artur Grygorian & Ionut E. Iacob (2018) A Concise Proof of the Triangle Inequality for the Jaccard Distance, The College Mathematics Journal, 49:5, 363-365

To link to this article: https://doi.org/10.1080/07468342.2018.1526020

Published online: 16 Nov 2018.

A Concise Proof of the Triangle Inequality for the Jaccard Distance

Artur Grygorian and Ionut E. Iacob

Artur Grygorian (ag06543@georgiasouthern.edu, MR ID 1276359) received his M.S. in mathematics from Georgia Southern University after studies in Ukraine. He is currently working as a senior data analyst in the supply chain analytics department at Home Depot.

Emil Iacob (ieiacob@georgiasouthern.edu, MR ID 966214) received his Ph.D. in computer science and an M.S. in mathematics from the University of Kentucky. He teaches mathematics and various data science topics at Georgia Southern University.

The Jaccard similarity measure is widely used in many applications, such as data mining and information retrieval [3] or even comparing DNA sequences [6]. The Jaccard similarity coefficient [2] of two sets A and B (not both empty) is defined as

$$J_{\text{sim}}(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

The higher the coefficient value (between 0 and 1), the more similar the two sets are.

The associated function

$$J_{\text{dist}}(A, B) = 1 - J_{\text{sim}}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

(with $J_{\text{dist}}(\varnothing, \varnothing) = 1$ by definition) is a distance function. While this is not difficult to prove, the first proof that J_{dist} satisfies the triangle inequality [4] was surprisingly tedious. Simpler proofs were given [1,5] and recently there were two new proofs using submodular functions as well as references for indirect proofs that use other distance functions [3].

In this article, we prove the triangle inequality for Jaccard distance with a simple, short proof by contradiction. But first, we give an example of Jaccard similarity for texts. This is used, for instance, in detecting plagiarism and identifying articles from the same source.

Example. Suppose we have three documents,

A: The falcons are the largest birds in the Falconinae subfamily of Falconidae.

B: The Patriots overcame a 25-point deficit with Falcons.

C: Falcons are large birds from the family of Falconidae.

doi.org/10.1080/07468342.2018.1526020
MSC: 97E50, 03E20

First, we construct the corresponding sets of (significant) words in each document,

$$A = \{\textit{falcon, large, bird, falconinae, family, falconidae}\},$$
$$B = \{\textit{patriot, overcome, point, deficit, falcon}\},$$
$$C = \{\textit{falcon, large, bird, family, falconidae}\}.$$

The typical approach to making these "bags of words" for a document is to take the list of all word roots while disregarding articles, pronouns, common verbs, punctuation, and grammar.

The Jaccard similarities for all pairs of these three documents are

$$\frac{|A \cap B|}{|A \cup B|} = \frac{1}{10}, \quad \frac{|A \cap C|}{|A \cup C|} = \frac{5}{6}, \quad \frac{|B \cap C|}{|B \cup C|} = \frac{1}{9}.$$

Clearly, documents A and C are similar while the other two pairs are not.

We now proceed to our proof of the triangle inequality for the Jaccard distance. This is a nice example of how a well-chosen proof technique, here proof by contradiction, can greatly simplify a proof.

Theorem. For sets A, B, C,

$$J_{\text{dist}}(A, B) \leq J_{\text{dist}}(A, C) + J_{\text{dist}}(C, B).$$

Proof. The case of two or all sets empty is trivial, so assume that at least two are nonempty.

We proceed by contradiction. Assume there are sets A, B, C, at least two nonempty, such that

$$1 - \frac{|A \cap B|}{|A \cup B|} > 1 - \frac{|A \cap C|}{|A \cup C|} + 1 - \frac{|C \cap B|}{|C \cup B|}$$

or, equivalently,

$$\frac{|A \cap B|}{|A \cup B|} < \frac{|A \cap C|}{|A \cup C|} + \frac{|C \cap B|}{|C \cup B|} - 1. \tag{1}$$

As shown in Figure 1, let a, b, c, x, y, z, w be nonnegative numbers representing cardinalities of the indicated nonoverlapping regions. Note that at most one of a, b, c can be zero.

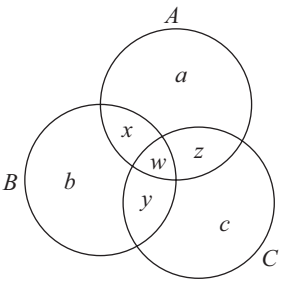

Figure 1. Venn diagram of sets A, B, C with the various intersection cardinalities labeled.

Using these, the inequality (1) becomes

$$\frac{x+w}{a+b+x+y+z+w} < \frac{z+w}{a+c+x+y+z+w} + \frac{y+w}{b+c+x+y+z+w} - 1. \tag{2}$$

Now clearly

$$\frac{x+w}{a+b+c+x+y+z+w} \leq \frac{x+w}{a+b+x+y+z+w}$$

and, since $\frac{m}{n} \leq \frac{m+k}{n+k}$ for any $0 \leq m < n$ and $0 \leq k$,

$$\frac{z+w}{a+c+x+y+z+w} \leq \frac{z+w+b}{a+c+x+y+z+w+b},$$
$$\frac{y+w}{b+c+x+y+z+w} \leq \frac{y+w+a}{b+c+x+y+z+w+a}.$$

Then (2) implies

$$\frac{x+w}{a+b+c+x+y+z+w} < \frac{b+z+w}{a+b+c+x+y+z+w} + \frac{a+y+w}{a+b+c+x+y+z+w} - 1$$

which, after simplification, reduces to

$$x+w < b+z+w+a+y+w-a-b-c-x-y-z-w$$

which implies $x < -c - x \leq 0$, a contradiction. ■

Acknowledgments. The authors wish to thank the reviewers for insightful suggestions which improved this note.

Summary. We give a short proof by contradiction of the triangle inequality for the Jaccard distance. The direct proof is not hard, but surprisingly tedious. This could serve as an example in an introduction to proofs class of a different method of proof significantly reducing a proof's length.

References

[1] Gilbert, G. (1972). Distance between sets. *Nature.* 239: 174. doi.org/10.1038/239174c0.
[2] Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytol.* 11: 37–50. doi.org/10.1111/j.1469-8137.1912.tb05611.x.
[3] Kosub, S. (2016). A note on the triangle inequality for the Jaccard distance. arxiv.org/abs/1612.02696.
[4] Levandowsky, M., Winter, D. (1971). Distance between sets. *Nature.* 234: 34–35. doi.org/10.1038/234034a0.
[5] Lipkus, A. H. (1999). A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.* 26: 263–265. doi.org/10.1023/A:101915443.
[6] Vorontsov, I. E., Kulakovskiy, I. V., Makeev, V. J. (2013). Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol. Biol.* 8: 23. doi.org/10.1186/1748-7188-8-23.