Electronic Communications of the EASST
Volume 63 (2014)

Proceedings of the Eighth International Workshop on Software Clones (IWSC 2014)

About Metrics for Clone Detection

— Position Paper —

Thierry Lavoie, Ettore Merlo

5 pages

Guest Editors: Nils Göde, Yoshiki Higo, Rainer Koschke
Managing Editors: Tiziana Margaria, Julia Padberg, Gabriele Taentzer
ECEASST Home Page: http://www.easst.org/eceasst/ ISSN 1863-2122

About Metrics for Clone Detection

Thierry Lavoie, Ettore Merlo

Departement de genie informatique et logiciel
Ecole Polytechnique de Montreal
Montreal, Canada
thierry-m.lavoie@polymtl.ca, ettore.merlo@polymtl.ca

Abstract: This paper presents some results about some metrics and their possible impact on clone detectors. It also discusses some advantages of why we should use metrics instead of arbitrary measures.

Keywords: Clones, Metrics

1 Introduction

Clone detectors rely on the concept of similarity and distance measures to identify cloned fragments. The choice of a specific distance function in a clone detector is arbitrary up to some extent. However, with a deeper knowledge of similarity measures, we can condition this choice to have some properties that can help improve scalability and quality of tools. This paper presents some interesting results, insights and questions about similarity and distance measures, including a somehow counter-intuitive result on the cosine distance.

For a comprehensive survey of many distances and metrics, the reader is invited to read [2].

This paper covers the following topics:

- A link between the Jaccard measure on sets and the Manhattan distance in euclidean space
- Limitations of the cosine distance on normalized vectors and approaches to overcome them
- Unanswered questions on some similarity detection techniques

As a convention in this paper, we set $\mathbb{R}_+^n = [0, \infty)^n$.

2 An Equivalence of the Jaccard Metric and the Manhattan Distance

Sometimes it is useful to link two known metrics to get a desired behavior or a better understanding of one of the two. In our case, we link the Jaccard metric on sets with the Manhattan distance on the space $\mathbb{R}_+^n$. Our equivalence holds for arbitrary sets, but different representations in the space $\mathbb{R}_+^\ltimes$ may be possible, which leads to different applications. We first prove the result, then we will explain how the result may be used.

Recall the Jaccard measure on two finite sets U, V is defined as:

$$\mathcal{J}(U,V) = \frac{|U \bigcap V|}{|U \bigcup V|}$$

and from this we define the Jaccard metric as:

$$\delta(U,V) = 1 - \frac{|U \bigcap V|}{|U \bigcup V|} = \frac{|U \bigcup V| - |U \bigcap V|}{|U \bigcup V|}$$

which is known to satisfy all the metric axioms (see section 3 for a brief recall). The Manhattan distance, noted l_1, on two vectors[1] u, v from $\mathbb{R}^n_+$ is:

$$l_1(u,v) = \sum_{i=1}^{n} |u_i - v_i|$$

Now, lets associate a unique integer $1 \leq i \leq |U \bigcup V|$ to every element μ in U and V. Now, choose $n = |U \bigcup V|$ and take $u, v \in \mathbb{R}^n_+$ such as $u_i = 1 \leftrightarrow \mu_i \in U$ otherwise $u_i = 0$, and $v_i = 1 \leftrightarrow \mu_i \in V$ otherwise $v_i = 0$. From this, we draw:

$$|U \bigcup V| = \sum_{i=1}^{n} max(u_i, v_i)$$
$$|U \bigcap V| = \sum_{i=1}^{n} min(u_i, v_i)$$

The *min* and *max* terms in the preceding equations are only equal if $u_i = v_i$, otherwise one of the two is u_i and the other is v_i and because $|u_i - v_i| = |v_i - u_i|$ we must have:

$$\begin{aligned} |U \bigcup V| - |U \bigcap V| &= |\sum_{i=1}^{n} max(u_i, v_i) - \sum_{i=1}^{n} min(u_i, v_i)| \\ &= \sum_{i=1}^{n} |u_i - v_i| \end{aligned}$$

Replacing the last two equalities in the original Jaccard metric leads to:

$$\delta(U,V) = \frac{|U \bigcup V| - |U \bigcap V|}{|U \bigcup V|} = \frac{\sum_{i=1}^{n} |u_i - v_i|}{\sum_{i=1}^{n} max(u_i, v_i)}$$

with the numerator of the last term being the Manhattan distance between u and v. This proves the existence of an equivalence between a normalization of the Manhattan distance between certain vectors and the Jaccard similarity between sets. It is easy to generalize this result to allow any positive integer for u_i and v_i instead of $0, 1$. We simply need to project multiple elements of the sets U and V onto a single coordinate i. The proof is then almost identical to the one presented here.

Why is this result interesting ? In practice, clone detectors use a lot of similarity and distance measures on sets. Most of them are not metrics (like the Dice coefficient, the Tanimoto distance,

[1] Actually, to define the Manhattan distance , we do not need the full power of a vector space but only requires the n-tuples in $\mathbb{R}^n_+$. However, it is now a custom to name everything in $\mathbb{R}^n_+$ a vector and we shall follow the custom.

etc.) and thus have a behavior less understood. Moreover, metrics lead to known opportunity of optimizations in search spaces that arbitrary distances do not offer [1]. Thus, even if the choice is ultimately arbitrary, there exist some arguments that favor metrics over arbitrary distances and knowing the link between some of them can help making a better choice. In this case, a simple distance between vectors gives us a useful interpretation as a distance between sets.

3 Properties of Some Angular Distances

We start by proving a result on the sine function.

Theorem 1 *Let X be a subset of $\mathbb{R}^n_+ | x \in X \rightarrow ||x|| = 1$. Let $\theta_{x,y}$ be the angle between any two x and $y \in X$. Finally, let $\delta : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \longrightarrow \mathbb{R}$ be defined as $\delta(x,y) = \sin\theta_{x,y}$. Then, δ is a metric on X.*

Proof. We need to prove that δ satisfies the four properties of a metric.

(Non-negativity) $\delta(x,y) \geq 0$. This is true, since the all x, y have positive coordinates and the angle between such vectors must be in $[0, \frac{\pi}{2}]$.

(Nullity) $\delta(x,y) = 0 \leftrightarrow x = y$. This is true, since the sine of an angle restricted to $[0, \frac{\pi}{2}]$ is 0 if and only if that angle is 0, and the angle between x and y is 0 if and only if $x = y$.

(Symmetry) $\delta(x,y) = \delta(y,x)$. This is true since the angle between x and y equals the angle between y and x.

(Triangle inequality) $\delta(x,y) + \delta(y,z) \geq \delta(x,z)$. The property holds, but it is tricky to prove. First, observe that if the angle between x and y or the angle between y and z is greater than the angle between x and z, then the property must hold since the sine is monotonically increasing in $[0, \frac{\pi}{2}]$. It remains to prove that it holds if the bigger angle is between x and z.

Now, clearly the sum of the angles $\theta_{x,y}$ and $\theta_{y,z}$ is greater or equal than $\theta_{x,z}$. Because sine is monotonically increasing in the considered interval, we now have:

$$\sin(\theta_{x,y}) + \sin(\theta_{y,z}) \geq \sin(\theta_{x,y} + \theta_{y,z}) \geq \sin(\theta_{x,z}) \tag{1}$$

We develop the first two members of this inequality:

$$\sin(\theta_{x,y}) + \sin(\theta_{y,z}) \geq \sin(\theta_{x,y} + \theta_{y,z}) = \sin(\theta_{x,y})\cos(\theta_{y,z}) + \sin(\theta_{y,z})\cos(\theta_{x,y})$$

Subtracting the right member to the left leaves:

$$\sin(\theta_{x,y}) - \sin(\theta_{x,y})\cos(\theta_{y,z}) + \sin(\theta_{y,z}) - \sin(\theta_{y,z})\cos(\theta_{x,y}) \geq 0$$

Because the cosine of these angles lies in $[0,1]$, we have

$$\sin(\theta_{x,y}) - \sin(\theta_{x,y})\cos(\theta_{y,z}) \geq 0$$
$$\sin(\theta_{y,z}) - \sin(\theta_{y,z})\cos(\theta_{x,y}) \geq 0$$

and we conclude that the sum of the two must be greater than or equal to 0. The inequality between the first two members of equation 1 holds, and because of our definition of the second member, the inequality between the last two members already held. Thus, we conclude that:

$$\sin(\theta_{x,y}) + \sin(\theta_{y,z}) \geq \sin(\theta_{x,z})$$

and the triangle inequality is satisfied.
All four properties hold and we have secured the theorem. □

Why is this result interesting ? First, even if the cosine distance is a very popular similarity measure on vectors, it is not a metric. To preserve the non-linearity of the cosine function, this results actually states that you need its dual function, the sine, to get a metric, under certain restriction. The sine is not as straightforward to compute as the cosine on vectors, but it has the advantage of being a metric on the first quadrant of $\mathbb{R}^n$.

It is also worth questioning whether or not the non-linearity of the sine and cosine is a desirable property. The angle distance between two vectors is a metric (this fact is used in the above proof) and is linear on the arc distance between the vectors on a unit circle. It would be interesting to verify whether or not the cosine is actually better than the sine and the angular distance considering our arguments. In general, it would be safe to compare a measure with closely related one to assess which is better even if it means a small additional computational cost.

4 Further Questions and Research: Implicit Distances, Why Should we Recover them ?

The previous sections dealt with special cases of distances that can be easily converted to a metric in order to use all the properties and knowledge we can draw from the vast literature on the subject. All these observations are interesting to ponder in the context of clone detectors explicitly based on a similarity or a distance measure. However, many tools only use implicit distances that are not properly defined as a mapping between a pair or a cluster of objects onto the real line $\mathbb{R}$. What should be the course of action for these tools ?

As hard as it can be, it should be possible in practice to recover the distance function. The mapping might be hard to express, or too many parameters may interact together to produce a long and complex formula, but these are only additional reasons to support the need to formalize implicit or hidden distance functions. Complete understanding of clone detection technology is intertwined with our ability to encapsulate their behavior in mathematical formula: if this task proves to be tedious, what can we say of understanding their behavior or how can we even completely compare them ?

The following questions are a starting point to help investigate the mathematical foundation of clone detection tools:

- What distance function does the clone detector uses ?
- How many parameters does the distance depends upon ? Are some redundant ? Are some useless ?

- If no mathematical formula seems possible to exist to encapsulate the distance used, why is it so ?
- What key features does the tool use ? Are there other tools using those features ? Do those tools have a well-formulated distance ?
- Is there an already existing distance that approximates the tool's implicit function ? How good is this approximation ?

Nevertheless, unanswered questions do not hinder performances and tools built around implicit distances can produce good results. To shed a deeper light on why they do have good results might however help the general understanding of clone detectors' behavior.

Acknowledgement

The authors wish to thank Theotime Menguy for his quick review of the proof in section 3. They would also like to thank Mathieu Merineau for providing the excellent reference [2].

Bibliography

[1] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proc. of 23rd International Conference on Very Large Data Bases*, pages 426–435. Morgan Kaufmann Publishers, 1997.

[2] S. seok Choi and S. hyuk Cha. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, pages 43–48, 2010.