Tanimoto index: reminder. In many practical applications, efficient algorithms use the value \( r = \frac{AB}{A^2 + B^2 - AB} \) – or some function of this value -- to describe the similarity between the vector A and B. Here, AB is a scalar (dot) product of the two vectors, and \( A^2 \) is the same as AA, i.e., the square of the vector’s length. The expression \( r \) is known as Tanimoto index.

Formulation of the problem. Usually, the main motivation for using the Tanimoto index is that it helps solve practical problems. However, the empirical success of this index prompt a natural question: why is this expression successful – and others are not as successful?

In this paper, we will provide a possible explanation for this success.

Preliminary analysis of the problem. Once we know the ratio \( r \), we can also compute the ratio \( R = \frac{1}{r + 1} = \frac{A^2 + B^2}{AB} \). Any function \( f(R) \) of the ratio \( R \) is also a function of \( r: f(R) = f(\frac{1}{r + 1}) \). Vice versa, since \( r = \frac{1}{(R - 1)} \), every function \( f(r) \) of the ratio \( R \) is also a function of the ratio \( R: f(r) = f(\frac{1}{(R - 1)}) \).

Thus, functions of the ratio \( r \) are exactly functions of the ratio \( R \), and usefulness of functions of \( r \) means usefulness of the functions of the ratio \( R \).

From this viewpoint, to explain the usefulness of Tanimoto index, it is sufficient to explain the usefulness of the ratio \( R \).

Towards an explanation: the simplest case. The simplest case is when each of the two vectors A and B has only one component. In other words, instead of the two vectors A and B, we had two numbers \( a \) and \( b \) – e.g., the two numbers obtained in two experiments. In this case, a natural data processing idea is:

- to estimate the mean \( E \), i.e., the average, the first order moment, as \( E = \frac{a + b}{2} \),
- and to estimate the second moment as \( M = \frac{a^2 + b^2}{2} \).

Then, we can process these two values, i.e., come up with some function of \( E \) and \( M \).

Comment. Another way to describe this situation is to consider the mean and the variance \( V \) (or the mean and the standard deviation – which is square root of \( V \)), but this is equivalent to estimating \( E \) and \( M \), since \( V = M - E^2 \) and thus,
$M = V + E^2$.

**Towards the general case.** What if we have vectors $A$ and $B$ instead of numbers? In this case, the sum $A + B$ is also a vector, and the sum $A^2 + B^2$ is a number.

So, we need a characteristic $c$ that depends on the sums $A + B$ and $A^2 + B^2$:

$$c = f(A + B, A^2 + B^2).$$

We want a numerical characteristic that will not change if we rotate a vector – i.e., equivalently, if we choose another basis in the corresponding space. The only characteristic of a vector that is preserved under all rotations is its length. The square of the length of the vector $A + B$ is equal to

$$(A+B)^2 = A^2 + 2AB + B^2.$$

So, a general rotation-invariant function of $A + B$ and $A^2 + B^2$ should depend on $(A + B)^2$ and $A^2 + B^2$:


One can easily check that this is equivalent to $c = g(AB, A^2 + B^2)$, where we denoted $g(x, y) = f(2x + y, y)$.

**Final requirement: scale-invariance.** We would also like to have a value that does not change if we change the measuring unit. If we replace the original measuring unit by a new unit which is $k$ times smaller, then all the numerical components of both vectors $A$ and $B$ are multiplied by $k$.

For example, if we go from feet to inches, then all the numerical values of length are multiplied by $k = 12$.

Under this transformation,

- $AB$ becomes $k^2 \cdot (AB)$, and
- $A^2 + B^2$ becomes $k^2 \cdot (A^2 + B^2)$.

So, invariance means that

$$g(k^2 \cdot (AB), k^2 \cdot (A^2 + B^2)) = g(AB, A^2 + B^2).$$

In particular, for $k^2 = 1 / (A^2 + B^2)$, we get

$$c = g(AB, A^2 + B^2) = g((AB) / (A^2 + B^2), 1) = g(1 / R, 1).$$

**Conclusion.** So, indeed any rotation- and scale-invariant characteristic of the two vectors which is based on the first two moments is a function of $R$, i.e., equivalently, a function of the Tanimoto index.

The use of the Tanimoto index is thus justified.