J Math Chem (2010) 48:287–289
DOI 10.1007/s10910-010-9668-4

ORIGINAL PAPER

Transforming Tanimoto queries on real valued vectors to range queries in Euclidian space

Thomas G. Kristensen

Received: 14 January 2010 / Accepted: 21 February 2010 / Published online: 7 March 2010

Abstract The Tanimoto coefficient has previously been proven to be a metric, but only in the case of binary valued vectors. Moreover, it has been proven that the Tanimoto coefficient for real valued vectors is not a metric. This means that it is not immediately possible to use metric based data structures for accelerating Tanimoto queries. This note presents a method for transforming Tanimoto queries into range queries in Euclidian space, making it possible to use metric data structures, as well as data structures designed for Euclidian space.

Keywords Tanimoto coefficient · Real valued descriptors · Range query

Vectors are often used for representing molecular structures when searching for chemical compounds with similar properties. The entries of these vectors can be binary, in which case they are referred to as fingerprints, or real valued, in which case they are referred to as descriptors. A diverse set of similarity measures are available for dealing with these vectors [1]. This note focuses on the Tanimoto coefficient, which is applicable to fingerprints as well as descriptors.

Previous studies have focused on decreasing the query time into databases of fingerprint vectors when the Tanimoto coefficient is used [2–4]. In this note we focus on decreasing the time for Tanimoto queries into databases of real valued descriptors.

The *Tanimoto coefficient* $T(A, B)$ between two vectors $A, B \in \mathbb{R}^n$ is calculated as

$$T(A, B) = \frac{AB}{\|A\|^2 + \|B\|^2 - AB}$$

T. G. Kristensen (✉)
Bioinformatics Research Center, Aarhus University, 8000 Århus C., Denmark
e-mail: tgk@cs.au.dk

A *Tanimoto query* consists of a target vector A and a minimum coefficient t. The result of a Tanimoto query is the set of vectors B in a database for which $T(A, B) \geq t$.

If A and B are binary, that is if their entries take on values either zero or some entry specific value, $T(A, B)$ will lie in the interval $[0, 1]$. In that case it has been proven that the Tanimoto distance, defined as $1 - T$, is a metric [5,6]. This means that the triangle inequality holds and standard data structures, such as μ-, vp-, M- and GNAT-trees [7–10], can be used for accelerating Tanimoto queries.

If the entries of the vectors are allowed to take on arbitrary values, the codomain of T extends to $[-\frac{1}{3}, 1]$ [1] and $1 - T$ ceases to be a metric [6]. We can therefore no longer apply standard data structures for accelerating queries. However, it is possible to convert the query into that of a range query in Euclidian space, probably the best known metric. This not only allows the use of data structures based on the triangle inequality, but it also enables the use of data structures tailored for Euclidian space, such as the kD-tree [11], even for binary valued vectors.

The reduction is as follows: Let $A, B \in \mathbb{R}^n$ and $t \in]0, 1]$. Then $T(A, B) \geq t$ iff

$$\left\| \frac{t+1}{2t} A - B \right\| \leq \frac{\sqrt{-4t^2 + (t+1)^2}}{2t} \|A\|.$$

The proof is by simple rewriting; be aware that we use the fact that $\|A\|^2 + \|B\|^2 - AB \geq 0$.

$$\begin{aligned}
\left\| \frac{t+1}{2t} A - B \right\| &\leq \frac{\sqrt{-4t^2 + (t+1)^2}}{2t} \|A\| \\
\frac{(t+1)^2}{4t^2} \|A\|^2 + \|B\|^2 - 2\frac{t+1}{2t} AB &\leq \frac{-4t^2 + (t+1)^2}{4t^2} \|A\|^2 \\
\|A\|^2 + \|B\|^2 &\leq \frac{t+1}{t} AB \\
t\left(\|A\|^2 + \|B\|^2\right) &\leq tAB + AB \\
t &\leq \frac{AB}{\|A\|^2 + \|B\|^2 - AB}
\end{aligned}$$

The previous result only covers the case in which $t \in]0, 1]$. If $t \in [-\frac{1}{3}; 0[$ then $T(A, B) \geq t$ iff

$$\left\| \frac{t+1}{2t} A - B \right\| \geq \frac{\sqrt{-4t^2 + (t+1)^2}}{2t} \|A\|.$$

The proof is analogous to that on the interval $]0; 1]$. Figure 1 illustrates both relations in the plane.

The results are applicable when all vectors with a coefficient above some threshold t are sought. If instead the n nearest neighbours are sought, decreasing values of t can be used until the number of vectors reaches n.

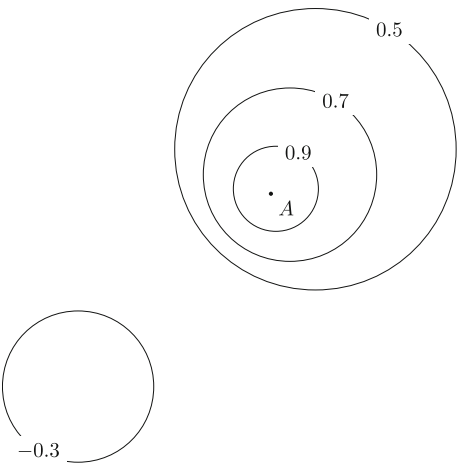

Fig. 1 An example in the plane. All the vectors inside the circle labelled 0.9 have a Tanimoto coefficient larger than 0.9 to A. All vectors outside the circle labelled -0.3 have a Tanimoto coefficient larger than -0.3 to A

Acknowledgments The author thanks Pierre Baldi for comments concerning the manuscript.

References

1. P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching. J. Chem. Inf. Comput. Sci. **38**(6), 983–996 (1998)
2. S.J. Swamidass, P. Baldi, Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. J. Chem. Inf. Model. **47**(2), 302–317 (2007)
3. P. Baldi, D.S. Hirschberg, R.J. Nasr, Speeding up chemical database searches using a proximity filter based on the logical exclusive or. J. Chem. Inf. Model. **48**(7), 1367–1378 (2008)
4. T.G. Kristensen, J. Nielsen, C.N.S. Pedersen, A tree-based method for the rapid screening of chemical fingerprints. Algorithms Mol Biol **5**(1), 9 (2010)
5. H. Späth, *Cluster Analysis Algorithms for Data Reduction and Classification of Objects* (Ellis Horwood, Chicester, 1980)
6. A.H. Lipkus, A proof of the triangle inequality for the tanimoto distance. J. Math. Chem. **26**(1–3), 263–265 (1999)
7. H. Xu, D.K. Agrafiotis, Nearest neighbor search in general metric spaces using a tree data structure with a simple heuristic. J. Chem. Inf. Model. **43**(6), 1933–1941 (2003)
8. P.N. Yianilos, Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the Fourth ACM-SIAM Symposium on Discrete Algorithms (1993)*
9. P. Ciaccia, M. Patella, P. Zezula, M-tree: An efficient access method for similarity search in metric spaces. In *VLDB'97: Proceedings of 23rd International Conference on Very Large Data Bases (August 25–29, 1997, Athens, Greece)*, ed. by M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochovsky, P. Loucopoulos, M.A. Jeusfeld (Morgan Kaufmann, 1997), pp. 426–435
10. S. Brin, Near neighbor search in large metric spaces. VLDB J. 574–584 (1995)
11. J.L. Bentley, Multidimensional binary search trees used for associative searching. Commun. ACM **18**(9), 509–517 (1975)