The Cosine Similarity in Terms of the Euclidean Distance

Marzena Kryszkiewicz
Warsaw University of Technology, Poland

INTRODUCTION

In many applications, especially in information retrieval, text mining, biomedical engineering and chemistry, the cosine similarity is often used to find objects most similar to a given one, so called nearest neighbors. Objects are typically represented as vectors. In particular, documents are often represented as term frequency vectors or its variants such as *tf_idf* vectors (Salton, Wong, & Yang 1975; Han, Kamber, & Pei 2011). The cosine similarity measure between vectors is interpreted as the cosine of the angle between them. According to this measure, two vectors are treated as similar if the angle between them is sufficiently small; that is, if its cosine is sufficiently close to 1.

The determination of nearest neighbors is challenging if analyzed vectors are high dimensional. In the case of distance metrics, one may apply the triangle inequality to quickly prune large numbers of objects that certainly are not nearest neighbors of a given vector (Uhlmann, 1991, Moore, 2000; Elkan, 2003; Kryszkiewicz & Lasek, 2010a; Kryszkiewicz & Lasek, 2010b; Kryszkiewicz & Lasek, 2010c; Patra, Hubballi, Biswas & Nandi, 2010; Kryszkiewicz & Lasek, 2011). Nevertheless, the cosine similarity is not a distance metric and, in particular, does not preserve the triangle inequality in general. In spite of this fact it was shown recently in (Kryszkiewicz, 2011; Kryszkiewicz, 2013a) that the problem of determining a cosine similarity neighborhood can be transformed to the problem of determining the Euclidean distance. This result allows applying the triangle inequality to make the determination of cosine similarity neighborhoods faster.

The objective of this chapter is to present the ways in which the relationship between the cosine similarity and the Euclidean distance can be used to determine cosine similar objects efficiently. We also outline further research directions related to this task.

BACKGROUND

Basic Notions and Properties

In this chapter, we assume that objects are represented by vectors of dimensionality n. Each vector u is understood as a sequence $[u_1, \ldots, u_n]$, where u_i is the value of the i-th dimension of u, $i = 1..n$. A vector all dimensions of which equal 0 will be called a *zero vector*. Otherwise, it will be called a *non-zero vector*.

Similarity or, respectively, dissimilarity of vectors can be defined in many ways. An important class of dissimilarity measures are distance metrics, which preserve the triangle inequality.

A measure d is said to *preserve the triangle inequality* if for any vectors u, v, and r, $d(u, r) \leq d(u, v) + d(v, r)$ or, alternatively $d(u, v) \geq d(u, r) - d(v, r)$.

The most popular distance metric is the *Euclidean distance*. The *Euclidean distance* between vectors u and v is denoted by $Euclidean(u, v)$ and is defined as follows:

$$Euclidean(u, v) = \sqrt{\sum_{i=1..n} (u_i - v_i)^2}.$$

Clearly, $Euclidean(u, v) = Euclidean(v, u)$.

DOI: 10.4018/978-1-4666-5202-6.ch223

Among most popular similarity measures is the *cosine similarity*. *The cosine similarity* of non-zero vectors u and v is denoted by $cosSim(u, v)$ and is defined as the cosine of the angle between them; that is,

$$cosSim(u, v) = \frac{u \cdot v}{|u||v|},$$

where:

- $u \cdot v$ is the *standard vector dot product of vectors u and v* and equals $\sum_{i=1..n} u_i v_i$;
- $|u|$ is *the length of vector u* and equals $\sqrt{u \cdot u}$.

Clearly, $cosSim(u, v) = cosSim(v, u)$ and the length of any vector equals the Euclidean distance between this vector and zero vector. Two non-zero vectors are treated as similar in the greatest possible degree if the angle between them equals 0; that is, if the cosine of the angle equals 1, regardless of particular lengths of the vectors.

Please note that, for example, $-cosSim(u, v)$ or $1 - cosSim(u, v)$ could be interpreted as a measure of dissimilarity between u and v.

Example 1: Figure 1 presents sample three vectors $u[0.2, 2.0]$, $r[3.0, 0.8]$ and $v[8.0, 1.0]$. Hence, $Euclidean(u, r) \approx 3.05$, $Euclidean(v, r) \approx 5.00$, $Euclidean(u, v) \approx 7.86$, whereas $cosSim(u, r) \approx 0.35$, $cosSim(v, r) \approx 0.99$, $cosSim(u, v) \approx 0.22$. One may note that the Euclidean distance between vectors v and r is greater than the Euclidean distance between vectors u and r, so v is more dissimilar from r than u in terms of the Euclidean distance. On the other hand, in terms of the cosine similarity measure, vector v is more similar to vector r than vector u, since the cosine of the angle between vectors v and r ($cosSim(v, r) = \cos\alpha$) is greater than the cosine of the angle between vectors u and r ($cosSim(u, r) = \cos\beta$).

One may easily check that neither $cosSim(v, r) \leq cosSim(v, u) + cosSim(u, r)$ nor $(-cosSim(u, v)) \leq (-cosSim(u, r)) + (-cosSim(r, v))$ nor $(1 - cosSim(u, v)) \leq (1 - cosSim(u, r)) + (1 - cosSim(r, v))$.

As follows from the above example, neither *cosSim* nor $-cosSim$ nor $1 - cosSim$ are guaranteed to satisfy the triangle inequality.

Figure 1. The Euclidean distance and the cosine similarity measure