Efficient Determination of Neighborhoods Defined in Terms of Cosine Similarity Measure

Marzena Kryszkiewicz

Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
mkr@ii.pw.edu.pl

Abstract. Cosine similarity measure is often and successfully used in the area of information retrieval, text classification, clustering, and ranking, where documents are usually represented as term frequency vectors (or its variants such as tf_idf vectors). In these tasks, the most time-consuming operation is the calculation of most similar vectors. This operation is commonly believed to be inefficient, if viable at all, for large high dimensional datasets. Recently, the usage of the triangle inequality property was offered to efficiently deal with this problem in the case of applying a distance metric. Although the cosine similarity measure is not a distance metric and, in particular, does not satisfy the triangle inequality property, in this paper we propose how to use this property to equally efficiently determine vectors similar with regard to the cosine similarity measure as in the case of a distance metric. In the paper, we address three types of sets of cosine similar vectors: all vectors the similarity of which to a given vector is not less than an ε threshold value and two variants of k-nearest neighbors of a given vector.

Keywords: k-nearest neighbors, ε-neighborhoods, cosine similarity measure, triangle inequality property, data and text clustering, high dimensional data

1 Introduction

Cosine similarity measure is often and successfully used in the area of information retrieval, text classification, clustering, and ranking, where documents are usually represented as term frequency vectors (or its variants such as tf_idf vectors) [2]. 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. In the above tasks, basic and most time-consuming operation is the calculation of most similar vectors. The operation is commonly believed to be inefficient, if viable at all, for large high dimensional datasets. Recently, the usage of the triangle inequality property was offered to efficiently deal with this problem in the case of applying a distance metric [1, 3, 4, 5, 6, 7, 8].

Although the cosine similarity measure is not a distance metric and, in particular, does not satisfy the triangle inequality property, in this paper we propose how to use this property to equally efficiently determine vectors similar with regard to the cosine similarity measure as in the case of a distance metric. In the paper, we address three types of sets of cosine similar vectors: all vectors the similarity of which to a given vector is not less than an ε threshold value and two variants of k-nearest neighbours of a given vector.

Our paper has the following layout. Section 2 provides basic notions and properties used in the paper. In particular, we examine properties and relationships among the three types of neighborhoods. In Section 3, we recall methods we offered in [3, 4, 5, 6], which use the triangle inequality property to calculate neighborhoods based on a distance metric efficiently also in the case of large high dimensional datasets. Sections 4 and 5 contain our main contribution. In Section 4, we formulate and prove the equivalence of neighborhoods defined in terms of the cosine similarity measure and some neighborhoods defined in terms of Euclidean distance. In Section 5, we provide an illustration of usefulness of the theoretical results offered in Section 4. Section 6 concludes our work.

2 Basic Notions and Properties

2.1 Basic Operations on Vectors

In the paper, we will consider vectors of a same dimensionality, say n. A vector u will be sometimes denoted as $[u_1, \ldots, u_n]$, where u_i is the value of the i-th dimension of u, $i = 1..n$. In Table 2.1.1, we provide a specification of basic operations on vectors.

Table 2.1.1. Specification of basic operators on vectors

Name of operation	Notation	Specification
sum of vectors u and v	$u + v$	$[u_1 + v_1, \ldots, u_n + v_n]$
subtraction of vectors u and v	$u - v$	$[u_1 - v_1, \ldots, u_n - v_n]$
multiplication of vector u by scalar α	αu	$[\alpha u_1, \ldots, \alpha u_n]$
division of vector by scalar α	$\frac{u}{\alpha}$	$\frac{1}{\alpha} u$
standard vector dot product of vectors u and v	$u \cdot v$	$\Sigma_{i=1..n} u_i v_i$
length of vector u	$\| u \|$	$\sqrt{u \cdot u}$

Table 2.1.2 presents their properties, which we will use in the paper.

Table 2.1.2. Properties of operations on vectors

Properties of operations on vectors
$\| u \|^2 = u \cdot u = \Sigma_{i=1..n} u_i^2$
$(u + v) \cdot (u + v) = \Sigma_{i=1..n} (u_i + v_i)^2 = (u \cdot u) + (v \cdot v) + 2(u \cdot v)$
$(u - v) \cdot (u - v) = \Sigma_{i=1..n} (u_i - v_i)^2 = (u \cdot u) + (v \cdot v) - 2(u \cdot v)$

2.2 Vector Dissimilarity and Similarity Measures

In the sequel, dissimilarity between two vectors p and q will be denoted by $dis(p, q)$. A vector q is considered as *less dissimilar* from vector p than vector r if $dis(q, p) < dis(r, p)$. In order to compare vectors, one may use a variety of dissimilarity measures among which an important class constitute *distance metrics*.

A *distance metric* (or shortly, *distance*) is defined as a dissimilarity measure that satisfies the following three conditions:

1) $dis(p, p) = 0$ for any vector p;
2) $dis(p, q) = dis(q, p)$ for any vectors p and q;
3) $dis(p, r) \leq dis(p, q) + dis(q, r)$ for any vectors p, q, and r.

The third condition is known as the *triangle inequality property*. Often, an alternative form of this property, presented below, is more useful.

Property 2.2.1. (Triangle inequality property). For any three vectors p, q, r:

$$dis(p, q) \geq dis(p, r) - dis(q, r).$$

It was shown in [1, 3, 4, 5, 6, 7, 8] how to use it for efficient clustering of both low and high dimensional datasets.

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

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

Property 2.2.2. $Euclidean(u, v) = \sqrt{(u-v)(u-v)}$.

Sometimes similarity measures are used rather than dissimilarity measures to compare vectors. In the following, the similarity between two vectors p and q will be denoted by $sim(p, q)$. A vector q is considered as *more similar* to vector p than vector r if $sim(q, p) > sim(r, p)$. Please note that $-sim(q, p)$ could be interpreted as a measure of dissimilarity between q and p.

In many applications, especially in text mining, a *cosine similarity measure*, which is a function of the angle between two vectors, is applied.

A cosine similarity measure between vectors u and v is denoted by $cosSim(u, v)$ and defined as the cosine of the angle between them; that is,

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

Example 2.2.1. Table 2.2.1 presents sample three vectors p, q, r. They are also presented graphically in Figure 2.2.1. One may note that the Euclidean distance between p and q is greater than the Euclidean distance between r and q. On the other hand, in terms of the cosine similarity measure, p is more similar to q than r, as the cosine of the angle between p and q ($cosSim(p, q) = \cos\alpha$) is greater than the cosine of the angle between r and q ($cosSim(r, q) = \cos\beta$).

The cosine similarities between these vectors are presented in Table 2.2.2. Table 2.2.3 shows that neither $cosSim$ nor $-cosSim$ satisfy the triangle inequality property. □

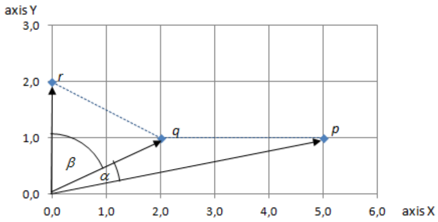

Fig. 2.2.1. Euclidean distance and the cosine similarity measure

Table 2.2.1. Sample three vectors

Vector	X	Y
p	5	1
q	2	1
r	0	2

Table 2.2.2. cosine similarity

(*u*, *v*)	*cosSim*(*u*, *v*)
(*p*, *q*)	0,964763821
(*p*, *r*)	0,196116135
(*q*, *r*)	0,447213595

Table 2.2.3. *cosSim* and *–cosSim* versus triangle inequality property

cosSim(*p*,*q*) ≤ *cosSim*(*p*,*r*) + *cosSim*(*r*,*q*)	*cosSim*(*p*,*r*) ≤ *cosSim*(*p*,*q*) + *cosSim*(*q*,*r*)	*cosSim*(*q*,*r*) ≤ *cosSim*(*q*,*p*) + *cosSim*(*p*,*r*)	–*cosSim*(*p*,*q*) ≤ –*cosSim*(*p*,*r*) + (–*cosSim*(*r*,*q*))	–*cosSim*(*p*,*r*) ≤ –*cosSim*(*p*,*q*) + (–*cosSim*(*q*,*r*))	–*cosSim*(*q*,*r*) ≤ –*cosSim*(*q*,*p*) + (–*cosSim*(*p*,*r*))
No	Yes	Yes	Yes	No	No

Corollary 2.2.1. Neither *cosSim* nor *–cosSim* satisfy the triangle inequality property.

2.3 Neighbourhoods Based on Dissimilarity Measures

Below, we provide definitions of neighborhoods in a given dataset *D* with regard to a given dissimilarity measure *dis*.

ε-neighborhood of a vector p in D is denoted by $\varepsilon\text{-}NB_{dis}^{D}(p)$ and defined as the set of all vectors in dataset *D* that are different from *p* and dissimilar from *p* by no more than *ε*; that is,

$$\varepsilon\text{-}NB_{dis}^{D}(p) = \{q \in D \mid q \neq p \wedge dis(p, q) \leq \varepsilon\}.$$

The set of all vectors in *D* that are different from *p* and less dissimilar from *p* than *q* will be denoted by $LessDissimilar_{dis}^{D}(p, q)$; that is,

$$LessDissimilar_{dis}^{D}(p, q) = \{s \in D \mid s \neq p \wedge dis(s, p) < dis(q, p)\}.$$

k-neighborhood of a vector p in D is denoted by $k\text{-}NB_{dis}^{D}(p)$ and defined as the set of all vectors *q* in *D*, $q \neq p$, such that the number of vectors that are different from *p* and less dissimilar from *p* than *q* is less than *k*; that is,

$$k\text{-}NB_{dis}^{D}(p) = \{q \in D \mid q \neq p \wedge |LessDissimilar_{dis}^{D}(p, q)| < k\}.$$

Please, note that for any value *k* and for each vector *p*, one may determine a value of parameter *ε* in such a way that $\varepsilon\text{-}NB_{dis}^{D}(p) = k\text{-}NB_{dis}^{D}(p)$. In the following, the least value of *ε* such that $\varepsilon\text{-}NB_{dis}^{D}(p) = k\text{-}NB_{dis}^{D}(p)$ will be called the *radius of* $k\text{-}NB_{dis}^{D}(p)$.

Proposition 2.3.1 [6]. Let $\varepsilon = \max(\{dis(q, p) \mid q \in k\text{-}NB_{dis}^{D}(p)\})$. Then $k\text{-}NB_{dis}^{D}(p) = \varepsilon\text{-}NB_{dis}^{D}(p)$ and *ε* is the radius of $k\text{-}NB_{dis}^{D}(p)$.

Proposition 2.3.2 [6]. If $|\varepsilon\text{-}NB_{dis}^{D}(p)| \geq k$, then $\varepsilon\text{-}NB_{dis}^{D}(p) \supseteq k\text{-}NB_{dis}^{D}(p)$.

Please, note that $k\text{-}NB_{dis}^{D}(p)$ may contain more than k vectors. In some applications, it is of interest to determine a set of exactly k “nearest” vectors (neighbors) instead of $k\text{-}NB_{dis}^{D}(p)$.

k-nearest neighbors of a vector p in D are defined as a set of k vectors q in D, $q \neq p$, such that the number of vectors that are different from p and less dissimilar from p than q is less than k.

Let $k\text{-}NN_{dis}^{D}(p)$ be a set of k-nearest neighbors of a vector p in D. Then the least value of ε such that $k\text{-}NN_{dis}^{D}(p) \subseteq \varepsilon\text{-}NB_{dis}^{D}(p)$ will be called the *radius of* $k\text{-}NN_{dis}^{D}(p)$.

Proposition 2.3.3. Let $k\text{-}NN_{dis}^{D}(p)$ be a set of k-nearest neighbors of a vector p in D and ε be the radius of $k\text{-}NN_{dis}^{D}(p)$. Then:

a) $k\text{-}NN_{dis}^{D}(p) \subseteq k\text{-}NB_{dis}^{D}(p)$;

b) $\forall q \in k\text{-}NB_{dis}^{D}(p) \setminus k\text{-}NN_{dis}^{D}(p)$, $dis(q, p) = \varepsilon$;

c) $k\text{-}NB_{dis}^{D}(p) = \varepsilon\text{-}NB_{dis}^{D}(p)$;

d) ε is the radius of $k\text{-}NB_{dis}^{D}(p)$;

e) ε is the radius of any set of k-nearest neighbors of vector p in D.

Corollary 2.3.1. Let $k\text{-}NN_{dis}^{D}(p)$ be a set of k-nearest neighbors of a vector p in D. If $|\varepsilon\text{-}NB_{dis}^{D}(p)| \geq k$, then $\varepsilon\text{-}NB_{dis}^{D}(p) \supseteq k\text{-}NB_{dis}^{D}(p) \supseteq k\text{-}NN_{dis}^{D}(p)$.

By Corollary 2.3.1, if ε-neighborhood of a vector p contains at least k vectors in D, then ε-neighborhood of p in D contains k-neighborhood of p in D, which in turn contains k-nearest neighbors of p in D.

2.4 Neighbourhoods Based on Similarity Measures

In this subsection, we provide alternative definitions of neighborhoods in a given set D in terms of a similarity measure *sim*.

ε-similarity neighborhood of a vector p in D is denoted by $\varepsilon\text{-}SNB_{sim}^{D}(p)$ and defined as the set of all vectors in dataset D that are different from p and similar to p by no less than ε; that is,

$$\varepsilon\text{-}SNB_{sim}^{D}(p) = \{q \in D |\ q \neq p \wedge sim(p, q) \geq \varepsilon\}.$$

The set of all vectors in D that are different from p and more similar to p than q will be denoted by $MoreSimilar_{sim}^{D}(p, q)$; that is,

$$MoreSimilar_{sim}^{D}(p, q) = \{s \in D |\ s \neq p \wedge sim(s, p) > sim(q, p)\}.$$

k-similarity neighborhood of a vector p in D is denoted by $k\text{-}SNB_{sim}^{D}(p)$ and defined as the set of all vectors q in D, $q \neq p$, such that the number of vectors that are different from p and more similar to p than q is less than k; that is,

$$k\text{-}SNB_{sim}^{D}(p) = \{q \in D |\ q \neq p \wedge |MoreSimilar_{sim}^{D}(p, q)| < k\}.$$

Please, note that for any value k and for each vector p, one may determine a value of parameter ε in such a way that $\varepsilon\text{-}SNB_{sim}^{D}(p) = k\text{-}SNB_{sim}^{D}(p)$. In the sequel, the greatest value of ε such that $\varepsilon\text{-}SNB_{sim}^{D}(p) = k\text{-}SNB_{sim}^{D}(p)$ will be called the *radius of* $k\text{-}SNB_{sim}^{D}(p)$.

Proposition 2.4.1. Let $\varepsilon = \min(\{sim(q, p)|\ q \in k\text{-}SNB_{sim}^{D}(p)\})$. Then $k\text{-}SNB_{sim}^{D}(p) = \varepsilon\text{-}SNB_{sim}^{D}(p)$ and ε is the radius of $k\text{-}SNB_{sim}^{D}(p)$.

Proposition 2.4.2. If $|\varepsilon\text{-}SNB_{sim}^{D}(p)| \geq k$, then $\varepsilon\text{-}SNB_{sim}^{D}(p) \supseteq k\text{-}SNB_{sim}^{D}(p)$.

k-similarity nearest neighbors of a vector p in D are defined as a set of k vectors q in D, $q \neq p$, such that the number of vectors that are different from p and more similar to p than q is less than k.

Let $k\text{-}SNN_{sim}^{D}(p)$ be a set of k-similarity nearest neighbors of a vector p in D. Then the greatest value of ε such that $k\text{-}SNN_{sim}^{D}(p) \subseteq \varepsilon\text{-}SNB_{sim}^{D}(p)$ will be called the *radius of* $k\text{-}SNN_{sim}^{D}(p)$.

Proposition 2.4.3. Let $k\text{-}SNN_{sim}^{D}(p)$ be a set of k-similarity nearest neighbors of a vector p in D and ε be the radius of $k\text{-}SNN_{sim}^{D}(p)$. Then:

a) $k\text{-}SNN_{sim}^{D}(p) \subseteq k\text{-}SNB_{sim}^{D}(p)$;

b) $\forall q \in k\text{-}SNB_{sim}^{D}(p) \setminus k\text{-}SNN_{sim}^{D}(p),\ sim(q, p) = \varepsilon$;

c) $k\text{-}SNB_{sim}^{D}(p) = \varepsilon\text{-}SNB_{sim}^{D}(p)$;

d) ε is the radius of $k\text{-}SNB_{sim}^{D}(p)$;

e) ε is the radius of any set of k-nearest neighbors of vector p in D.

Corollary 2.4.1. Let $k\text{-}SNN_{sim}^{D}(p)$ be a set of k-similarity nearest neighbors of a vector p in D. If $|\varepsilon\text{-}SNB_{sim}^{D}(p)| \geq k$, then $\varepsilon\text{-}SNB_{sim}^{D}(p) \supseteq k\text{-}SNB_{sim}^{D}(p) \supseteq k\text{-}SNN_{sim}^{D}(p)$.

By Corollary 2.4.1, if ε-similarity neighborhood of a vector p in D contains at least k vectors, then ε-similarity neighborhood of p in D contains k-similarity neighborhood of p in D, which in turn contains k-similarity nearest neighbors of p in D.

3 Triangle Inequality as a Mean for Efficient Determining of Neighborhoods Based on Distance Metrics

3.1 Efficient Determination of ε-Neighborhoods

In this subsection, we recall the method of determining ε-neighborhoods efficiently, which we proposed in [3, 4].

Lemma 3.1.1 [3, 4]. Let dis be a distance metric and D be a set of vectors. For any two vectors p, q in D and any vector r:

$$dis(p, r) - dis(q, r) > \varepsilon \Rightarrow q \notin \varepsilon\text{-}NB_{dis}^{D}(p) \wedge p \notin \varepsilon\text{-}NB_{dis}^{D}(q).$$

Lemma 3.1.1 comes from the fact that $dis(p, r) - dis(q, r) > \varepsilon$ (by assumption) and $dis(p, q) \geq dis(p, r) - dis(q, r)$ (by the triangle inequality property,). Hence,

$dis(p, q) > \varepsilon$. Thus, the fact that the difference of distances from two vectors p and q to some vector r is greater than ε implies that $q \notin \varepsilon\text{-}NB_{dis}^{D}(p)$ and $p \notin \varepsilon\text{-}NB_{dis}^{D}(q)$.

Now, let us consider a vector v such that $dis(v, r) - dis(q, r) > dis(p, r) - dis(q, r)$. If we know that $dis(p, r) - dis(q, r) > \varepsilon$, we may conclude that $dis(v, r) - dis(q, r) > \varepsilon$, and thus, $q \notin \varepsilon\text{-}NB_{dis}^{D}(v) \wedge v \notin \varepsilon\text{-}NB_{dis}^{D}(q)$ without calculating the real distance between v and q. This observation provides intuition behind Theorem 3.1.1, which we offered in [3, 4].

Theorem 3.1.1 [3, 4]. Let dis be a distance metric, r be any vector and D be a set of vectors ordered in a non-decreasing way with regard to their distances to r. Let p be any vector in D, q_f be a vector following vector p in D such that $dis(q_f, r) - dis(p, r) > \varepsilon$, and q_b be a vector preceding vector p in D such that $dis(p, r) - dis(q_b, r) > \varepsilon$. Then:

a) q_f and all vectors following q_f in D do not belong to $\varepsilon\text{-}NB_{dis}^{D}(p)$;

b) q_b and all vectors preceding q_b in D do not belong to $\varepsilon\text{-}NB_{dis}^{D}(p)$.

As follows from Theorem 3.1.1, it makes sense to order all vectors in a given dataset D with regard to some reference vector, say r, as this enables simple elimination of a potentially large subset of vectors that certainly do not belong to an ε-neighborhood of an analyzed vector. The experiments reported in [3, 4], which related to density based clustering based on determination of ε-neighborhood for each point in D by means of the class of *TI-DBSCAN* algorithms, we proposed there, showed that the usage of Theorem 3.1.1 fastened the process by up to two orders of magnitude even for high dimensional large datasets (up to hundreds of vectors and up to tens of thousands of vectors).

3.2 Efficient Determination of *k*-Neighborhoods

In this subsection, we recall the method of determining k-neighborhood efficiently, which we proposed in [5, 6]. Also in this case, all vectors in a given dataset D are assumed to be ordered with regard to some reference vector r. Then, for each vector p in D, its k-neighborhood can be determined in the following steps:

1) The radius, say ε, of k-neighborhood of p is estimated based on the real distances of k vectors located directly before and after p in the ordered set D.

2) Next, ε-neighborhood is determined as described in Subsection 3.1 (the only difference is that the real distances to p from vectors considered in phase 1, are not calculated again).

3) k-neighborhood of p is determined as a subset of ε-neighborhood found in step 2.

The above description is a bit simplified. In [5, 6], steps 2 and 3 were not split, and the value of ε was adapted (narrowed) with each new candidate vector having a chance to belong to k-neighborhood of p. Please, see [5, 6] for a more detailed description. The presented approach is justified by Theorem 3.2.1, which we offered in [5, 6].

Theorem 3.2.1 [5, 6]. Let dis be a distance metric, r be any vector and D be a set of vectors ordered in a non-decreasing way with regard to their distances to r. Let p be any vector in D and ε be a value such that $|\varepsilon\text{-}NB_{dis}^{D}(p)| \geq k$, q_f be a vector following vector p in D such that $dis(q_f, r) - dis(p, r) > \varepsilon$, and q_b be a vector preceding vector p in D such that $dis(p, r) - dis(q_b, r) > \varepsilon$. Then:

a) q_f and all vectors following q_f in D do not belong to $k\text{-}NB_{dis}^{D}(p)$;

b) q_b and all vectors preceding q_b in D do not belong to $k\text{-}NB_{dis}^{D}(p)$.

As follows from the experiments reported in [5, 6], which related to the determination of *k*-neighborhood for each point in *D* by means of the class of *TI-k-Neighborhood-Index* algorithms, we proposed there, showed that the usage of Theorem 3.2.1 fastened the process by up to two orders of magnitude even for high dimensional large datasets.

3.3 Efficient Determination of *k*-Nearest Neighbors

Let $k\text{-}NN_{dis}^{D}(p)$ be a set of *k*-similarity nearest neighbors of a vector p in D. Since $k\text{-}NB_{dis}^{D}(p) \supseteq k\text{-}NN_{dis}^{D}(p)$, then the vectors that do not belong to $k\text{-}NB_{dis}^{D}(p)$ do not belong $k\text{-}NN_{dis}^{D}(p)$ either. This observation allows us to derive Proposition 3.3.1 from Theorem 3.2.1.

Proposition 3.3.1. Let *dis* be a distance metric, r be any vector and D be a set of vectors ordered in a non-decreasing way with regard to their distances to r. Let p be any vector in D, $k\text{-}NN_{dis}^{D}(p)$ be a set of *k*-similarity nearest neighbors of vector p in D and ε be a value such that $|\varepsilon\text{-}NB_{dis}^{D}(p)| \geq k$, q_f be a vector following vector p in D such that $dis(q_f, r) - dis(p, r) > \varepsilon$, and q_b be a vector preceding vector p in D such that $dis(p, r) - dis(q_b, r) > \varepsilon$. Then:

a) q_f and all vectors following q_f in D do not belong to $k\text{-}NN_{dis}^{D}(p)$;

b) q_b and all vectors preceding q_b in D do not belong to $k\text{-}NN_{dis}^{D}(p)$.

As follows from Proposition 3.3.1, determination of *k*-similarity nearest neighbors $k\text{-}NN_{dis}^{D}(p)$ of a vector p can be carried out in a similar way as the determination of $k\text{-}NB_{dis}^{D}(p)$ described in Subsection 3.2. The two procedures will differ only in extracting $k\text{-}NN_{dis}^{D}(p)$ and $k\text{-}NB_{dis}^{D}(p)$ from an ε-neighborhood containing at least k vectors; namely, $k\text{-}NN_{dis}^{D}(p) \subseteq k\text{-}NB_{dis}^{D}(p)$.

4 Cosine Similarity Measure versus Euclidean Distance and Neighborhoods Defined in Their Terms

In Section 3, we described how neighborhoods expressed in terms of a distance metric can be calculated efficiently by using the triangle inequality property for skipping vectors that do not have a chance to belong to these neighborhoods. On the other hand, as we showed in Example 2.2.1, the cosine similarity measure is not a distance metric as it does not satisfy the triangle inequality property. However, in this section, we will show that neighborhoods defined in terms of the cosine similarity measure are equivalent to some neighborhoods defined in terms of Euclidean distance. Hence, we will transform the problem of calculating neighborhoods defined in terms of the cosine similarity measure to the problem of calculating neighborhoods defined in terms of Euclidean distance, which is efficiently solvable by means of the triangle inequality property. To this end, we will start with recalling properties of, so called, normalized vectors. Then, we will show the relationship between the cosine similarity measure and Euclidean distance. Finally, we will use this relationship to formulate and prove the equivalence of neighborhoods defined in terms of the cosine similarity measure and some neighborhoods defined in terms of Euclidean distance.

4.1 Properties of Normalized Vectors

A *normalized form of a vector u* is denoted by $NF(u)$ and defined as the ratio of u to its length $|u|$; that is,

$$NF(u) = \frac{u}{|u|}.$$

A *vector u* is called a *normalized vector* (or alternatively, a *unit vector*) if $u = NF(u)$.

Property 4.1.1. Let u be a vector. Then:

a) $NF(u) \cdot NF(u) = 1$;

b) $|NF(u)| = 1$.

Proof. Ad a) $NF(u) \cdot NF(u) = \frac{u}{|u|} \cdot \frac{u}{|u|} = \frac{u \cdot u}{|u|^2} = \frac{|u|^2}{|u|^2} = 1$.

Ad b) $|NF(u)| = \sqrt{NF(u) \cdot NF(u)} = \sqrt{1} = 1$; □

As follows from Property 4.1.1, the standard vector dot product of a normalized vector with itself is equal to 1 and the length of a normalized vector is equal to 1.

Property 4.1.2. Let u and v be vectors. Then:

a) $cosSim(NF(u), NF(v)) = NF(u) \cdot NF(v)$;

b) $cosSim(u, v) = cosSim(NF(u), NF(v))$;

c) $cosSim(u, v) = NF(u) \cdot NF(v)$.

Proof. Ad a) $cosSim(NF(u), NF(v)) = \frac{NF(u) \cdot NF(v)}{|NF(u)||NF(v)|} =$

/* by Property 4.1.1b, $|NF(u)| = |NF(v)| = 1$ */ $= NF(u) \cdot NF(v)$.

Ad b) $cosSim(u,v) = \frac{u \cdot v}{|u||v|} = \frac{u}{|u|} \cdot \frac{v}{|v|} = NF(u) \cdot NF(v) =$ /* by Property 4.1.2a */ $=$ $cosSim(NF(u), NF(v))$.

Ad c) Follows immediately from Property 4.1.2a,b. □

As follows from Property 4.1.2a, the cosine similarity between two normalized vectors is equal to the standard vector dot product of these vectors. In addition, by Property 4.1.2b, the cosine similarity between two vectors is equal to the cosine similarity between their normalized forms. By Property 4.1.2c, the cosine similarity between two vectors is equal to the standard vector dot product of their normalized forms.

4.2 Relationship between Cosine Similarity Measure and Euclidean Distance

We start with formulating and proving two lemmas showing that the cosine similarity between two vectors can be expressed as a function of their lengths and Euclidean distance between them (Lemma 4.2.1) and that the cosine similarity between two normalized (forms of) vectors can be expressed as a function of solely their Euclidean distance (Lemma 4.2.2).

Lemma 4.2.1. Let u and v be vectors. Then:

a) $cosSim(u,v) = \dfrac{(u \cdot u) + (v \cdot v) - (u - v) \cdot (u - v)}{2\,|u|\,|v|}$;

b) $cosSim(u,v) = \dfrac{|u|^2 + |v|^2 - Euclidean^2(u,v)}{2\,|u|\,|v|}$.

Proof. Ad a) Since $(u - v) \cdot (u - v) = (u \cdot u) + (v \cdot v) - 2(u \cdot v)$ and $cosSim(u,v) = \dfrac{u \cdot v}{|u|\,|v|}$, then $(u - v) \cdot (u - v) = (u \cdot u) + (v \cdot v) - 2(cosSim(u,v)|\,u\,||\,v\,|)$.

Hence, $cosSim(u,v) = \dfrac{(u \cdot u) + (v \cdot v) - (u - v) \cdot (u - v)}{2\,|u|\,|v|}$.

Ad b) Follows immediately from Lemma 4.2.1a, the fact that $u \cdot u = |\,u\,|^2$, $v \cdot v = |\,v\,|^2$ and $Euclidean(u, v) = \sqrt{(u - v) \cdot (u - v)}$. □

Lemma 4.2.2. Let u and v be vectors. Then:

$$cosSim(NF(u), NF(v)) = \frac{2 - Euclidean^2(NF(u), NF(v))}{2}.$$

Proof. $cosSim(NF(u), NF(v))$ = /* by Lemma 4.2.1b */ =

$$\frac{|\,NF(u)\,|^2 + |\,NF(v)\,|^2 - Euclidean^2(NF(u), NF(v))}{2\,|\,NF(u)\,|\,|\,NF(v)\,|} =$$

/* by Property 4.1.1b, $|\,NF(u)\,| = |\,NF(v)\,| = 1$ */ =

$$\frac{2 - Euclidean^2(NF(u), NF(v))}{2}.\ \square$$

Taking into account that the cosine similarity between two vectors is equal to the cosine similarity between their normalized forms (by Property 4.1.2b), we conclude from Lemma 4.2.2 that the cosine similarity between two vectors can be expressed as a function of solely Euclidean distance of their normalized forms (Theorem 4.2.1).

Theorem 4.2.1. Let u and v be vectors. Then:

$$cosSim(u, v) = \frac{2 - Euclidean^2(NF(u), NF(v))}{2}.$$

4.3 Neighborhoods Based on Cosine Similarity Measure and Neighborhoods Based on Euclidean Distance

In this subsection, we will use Theorem 4.2.1 to derive relationships between neighborhoods based on the cosine similarity measure and some neighborhoods based on Euclidean distance. First, we start with Lemma 4.3.1a, in which we formulate and prove that a comparison of the cosine similarity between two vectors with an ε threshold is equivalent to a comparison of Euclidean distance between their normalized forms with some ε' threshold. In Lemma 4.3.1b, we formulate and prove that a comparison of the cosine similarity between any two vectors s and p with the cosine similarity between vector q and any vector p is equivalent to a comparison of Euclidean distances between their normalized forms.

Lemma 4.3.1. Let $\varepsilon \in [-1, 1]$ and $\varepsilon' = \sqrt{2-2\varepsilon}$. Then:

a) $cosSim(u, v) \geq \varepsilon$ iff $Euclidean(NF(u), NF(v)) \leq \varepsilon'$;

b) $cosSim(s,p) > cosSim(q,p)$ iff $Euclidean(NF(s), NF(p)) < Euclidean(NF(q), NF(p))$.

Proof. Ad a) $cosSim(u, v) \geq \varepsilon$ iff /* by Theorem 4.2.1 */

$$\frac{2 - Euclidean^2(NF(u), NF(v))}{2} \geq \varepsilon \text{ iff}$$

$$Euclidean(NF(u), NF(v)) \leq \sqrt{2-2\varepsilon} = \varepsilon'.$$

Ad b) $cosSim(s, p) > cosSim(q, p)$ iff /* by Theorem 4.2.1 */

$$\frac{2 - Euclidean^2(NF(s), NF(p))}{2} > \frac{2 - Euclidean^2(NF(q), NF(p))}{2} \text{ iff}$$

$$Euclidean(NF(s), NF(p)) < Euclidean(NF(q), NF(p)).$$ □

Lemma 4.3.1 enables us to formulate and prove Lemma 4.3.2, in which we present the problem of determining vectors $p_{(l)}$ in D that are more cosine similar to a given vector $p_{(i)}$ than another vector $p_{(j)}$ by means of normalized forms of vectors and Euclidean distance.

Lemma 4.3.2. Let D be an ordered set of m vectors $(p_{(1)}, \ldots, p_{(m)})$, D' be the ordered set of m vectors $(u_{(1)}, \ldots, u_{(m)})$ such that $u_{(i)} = NF(p_{(i)})$, $i = 1..m$, $\varepsilon \in [-1, 1]$ and $\varepsilon' = \sqrt{2-2\varepsilon}$. Then:

a) $MoreSimilar_{cosSim}^{D}(p_{(i)}, p_{(j)}) = \{p_{(l)} \in D \mid u_{(l)} \in LessDissimilar_{Euclidean}^{D'}(u_{(i)}, u_{(j)})\}$;

b) $|MoreSimilar_{cosSim}^{D}(p_{(i)}, p_{(j)})| < k$ iff $|LessDissimilar_{Euclidean}^{D'}(u_{(i)}, u_{(j)})| < k$.

Proof. Ad a) $MoreSimilar_{cosSim}^{D}(p_{(i)}, p_{(j)}) =$

$\{p_{(l)} \in D \mid p_{(l)} \neq p_{(i)} \wedge cosSim(p_{(l)}, p_{(i)}) > cosSim(p_{(j)}, p_{(i)})\}$ = /* by Lemma 4.3.1b */

$\{p_{(l)} \in D \mid u_{(l)} \in D' \wedge u_{(l)} \neq u_{(i)} \wedge Euclidean(u_{(l)}, u_{(i)}) < Euclidean(u_{(j)}, u_{(i)})\} =$

$\{p_{(l)} \in D \mid u_{(l)} \in LessDissimilar_{Euclidean}^{D'}(u_{(i)}, u_{(j)})\}$.

Ad b) Follows immediately from Lemma 4.3.2a. □

Now, we are ready to formulate and prove the equivalence of neighborhoods defined in terms of the cosine similarity measure and some neighborhoods defined in terms of Euclidean distance.

Theorem 4.3.1. Let D be an ordered set of m vectors $(p_{(1)}, \ldots, p_{(m)})$, D' be the ordered set of m vectors $(u_{(1)}, \ldots, u_{(m)})$ such that $u_{(i)} = NF(p_{(i)})$, $i = 1..m$, $\varepsilon \in [-1, 1]$ and $\varepsilon' = \sqrt{2-2\varepsilon}$. Then:

a) $\varepsilon\text{-}SNB_{cosSim}^{D}(p_{(i)}) = \{p_{(j)} \in D|\ u_{(j)} \in \varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(i)})\}$;

b) $k\text{-}SNB_{cosSim}^{D}(p_{(i)}) = \{p_{(j)} \in D|\ u_{(j)} \in k\text{-}NB_{Euclidean}^{D'}(u_{(i)})\}$;

c) If $k\text{-}NN_{Euclidean}^{D'}(u_{(i)})$ is a set of k nearest neighbours of $u_{(i)}$ in D', then $\{p_{(j)} \in D|\ u_{(j)} \in k\text{-}NN_{Euclidean}^{D'}(u_{(i)})\}$ is a set of k similarity nearest neighbors of $p_{(i)}$ in D.

Proof. Ad a) $\varepsilon\text{-}SNB_{cosSim}^{D}(p_{(i)}) = \{p_{(j)} \in D|\ p_{(j)} \neq p_{(i)} \wedge cosSim(p_{(i)}, p_{(j)}) \geq \varepsilon\}$ =
/* by Lemma 4.3.1a */
$\{p_{(j)} \in D|\ u_{(j)} \in D' \wedge u_{(j)} \neq u_{(i)} \wedge Euclidean(u_{(i)}, u_{(j)}) \leq \varepsilon'\}$ =
$\{p_{(j)} \in D|\ u_{(j)} \in \varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(i)})\}$.

Ad b) $k\text{-}SNB_{cosSim}^{D}(p_{(i)})$ =
$\{p_{(j)} \in D|\ p_{(j)} \neq p_{(i)} \wedge |MoreSimilar_{cosSim}^{D}(p_{(i)}, p_{(j)})| < k\}$ = /* by Lemma 4.3.2b */
$\{p_{(j)} \in D|\ u_{(j)} \in D' \wedge u_{(j)} \neq u_{(i)} \wedge |LessDissimilar_{Euclidean}^{D'}(u_{(i)}, u_{(j)})| < k\}$ =
$\{p_{(j)} \in D|\ u_{(j)} \in k\text{-}NB_{Euclidean}^{D'}(u_{(i)})\}$.

Ad c) Let $k\text{-}NN_{Euclidean}^{D'}(u_{(i)})$ be a set of k nearest neigbours of $u_{(i)}$ in D'.
Then, $\forall u_{(j)} \in k\text{-}NN_{Euclidean}^{D'}(u_{(i)})$, $u_{(j)} \in D'$ $\wedge$ $u_{(j)} \neq u_{(i)}$ $\wedge$ $|LessDissimilar_{Euclidean}^{D'}(u_{(i)}, u_{(j)})| < k$. Hence, by Lemma 4.3.2b, $\forall u_{(j)} \in k\text{-}NN_{Euclidean}^{D'}(u_{(i)})$, $p_{(j)} \in D \wedge p_{(j)} \neq p_{(i)} \wedge |MoreSimilar_{cosSim}^{D}(p_{(i)}, p_{(j)})| < k$. Thus, $\{p_{(j)} \in D|\ u_{(j)} \in k\text{-}NN_{Euclidean}^{D'}(u_{(i)})\} \subseteq k\text{-}SNN_{cosSim}^{D}(p_{(i)})$ and $|\{p_{(j)} \in D|\ u_{(j)} \in k\text{-}NN_{Euclidean}^{D'}(u_{(i)})\}| = k$. Therefore, $\{p_{(j)} \in D|\ u_{(j)} \in k\text{-}NN_{Euclidean}^{D'}(u_{(i)})\}$ is a set of k similarity nearest neighbors of $p_{(i)}$ in D. □

5 Determination of Neighborhoods Based on Cosine Similarity Measure as Determination of Neighborhoods Based on Euclidean Distance

Theorem 4.3.1 tells us that neighborhoods defined in terms of the cosine similarity measure can be determined as respective neighborhoods defined in terms of Euclidean distance. Thus, we propose the following approach to determination of neighborhoods defined in terms of the cosine similarity measure:

First, the original vectors, say $D = (p_{(1)}, \ldots, p_{(m)})$, should be transformed to their normalized forms $D' = (u_{(1)}, \ldots, u_{(m)})$. Then, ε-similarity neighborhoods (or alternatively, k-similarity neighborhoods or k-similarity nearest neighbors) in dataset D with regard to the cosine similarity measure should be found as ε'-neighborhoods, where $\varepsilon' = \sqrt{2-2\varepsilon}$, (or alternatively, k- neighborhoods or k-nearest neighbors) in dataset *D*' with regard to Euclidean distance by means of the triangle inequality property as described in Section 3.

Example 5.1 (Determination of an ε-similarity neighborhood). In this example, we will consider determination of ε-similarity neighborhood $\varepsilon\text{-}SNB_{cosSim}{}^{D}(p_{(3)})$ of two dimensional vector $p_{(3)}$ in dataset $D = (p_{(1)}, \ldots, p_{(8)})$ from Figure 5.1 (and Table 5.1). We assume that the cosine similarity threshold $\varepsilon = 0.9855$ (which roughly corresponds to the angle of 10°; that is, cos(10°) ≈ 0.9855). Figure 5.2 shows set $D' = (u_{(1)}, \ldots, u_{(8)})$ that contains normalized forms of vectors from *D*. Clearly, the lengths of all of them are equal to 1. Now, we will determine the corresponding Euclidean distance threshold ε' as $\sqrt{2-2\varepsilon}$ (according to Theorem 4.3.1a). Hence, $\varepsilon' = 0.17$. At this moment, we may start the procedure of determining ε-similarity neighborhood for vector $p_{(3)}$ in dataset *D* with the *cosSim* measure as the procedure of determining ε'-neighborhood for vector $u_{(3)}$ in dataset *D*' of normalized vectors as proposed in [3, 4] (and recalled in Subsection 3.1) with regard to Euclidean distance.

The vectors in *D*' need to be sorted with regard to their Euclidean distances to a same reference vector. For the sake of the example, we choose $r = [1, 0]$ as a reference vector. Table 5.2 shows set *D*' ordered in a non-decreasing way with regard to the distances of its vectors to vector *r*.

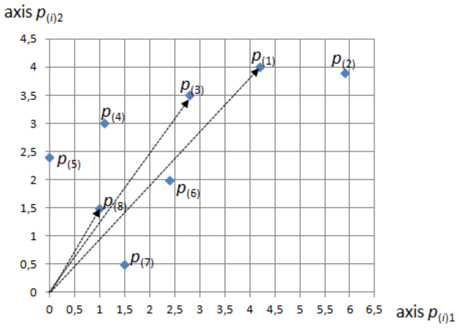

Fig. 5.1. Sample set *D* of vectors

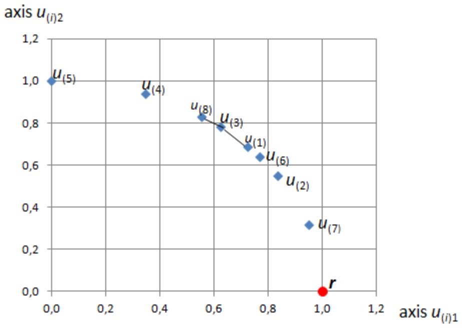

Fig. 5.2. Set *D*' containing normalized forms of vectors from *D*

Table 5.1. Sample set *D*

Vector $p_{(i)}$	$p_{(i)1}$	$p_{(i)2}$
$p_{(1)}$	4,20	4,00
$p_{(2)}$	5,90	3,90
$p_{(3)}$	2,80	3,50
$p_{(4)}$	1,10	3,00
$p_{(5)}$	0,00	2,40
$p_{(6)}$	2,40	2,00
$p_{(7)}$	1,00	0,50
$p_{(8)}$	1,00	1,50

Table 5.2. Sorted set *D*' ($r = [1, 0]$)

Vector $u_{(i)}$	$u_{(i)1}$	$u_{(i)2}$	*Euclidean* $(u_{(i)}, r)$
$u_{(7)}$	0,89	0,45	0,46
$u_{(2)}$	0,83	0,55	0,58
$u_{(6)}$	0,77	0,64	0,68
$u_{(1)}$	0,72	0,69	0,74
$u_{(3)}$	0,62	0,78	0,87
$u_{(8)}$	0,55	0,83	0,94
$u_{(4)}$	0,34	0,94	1,15
$u_{(5)}$	0,00	1,00	1,41

As follows from Table 5.2, the first vector q_f following vector $u_{(3)}$ in D' for which $\mathit{Euclidean}(q_f, r) - \mathit{Euclidean}(u_{(3)}, r) > \varepsilon'$ is vector $u_{(4)}$ ($\mathit{Euclidean}(u_{(4)}, r) - \mathit{Euclidean}(u_{(3)}, r) = 1.15 - 0.87 = 0.28 > \varepsilon'$), and the first vector q_b preceding vector $u_{(3)}$ in D' for which $\mathit{Euclidean}(u_{(3)}, r) - \mathit{Euclidean}(q_b, r) > \varepsilon'$ is vector $u_{(6)}$ ($\mathit{Euclidean}(u_{(3)}, r) - \mathit{Euclidean}(u_{(6)}, r) = 0.87 - 0.68 = 0.19 > \varepsilon'$). Thus, by Theorem 3.1.1, neither vector $u_{(4)}$ nor the vectors following $u_{(4)}$ in D' as well as neither vector $u_{(6)}$ nor the vectors preceding $u_{(6)}$ in D' belong to $\varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(3)})$. Hence, only vectors $u_{(1)}$ and $u_{(8)}$ may belong to $\varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(3)})$ and only for these vectors it is necessary to calculate their real Euclidean distances to $u_{(3)}$. These distances are as follows, $\mathit{Euclidean}(u_{(1)}, u_{(3)}) = 0.13$ and $\mathit{Euclidean}(u_{(8)}, u_{(3)}) = 0.09$. Since, both values are less than ε', then $\varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(3)}) = \{u_{(1)}, u_{(8)}\}$, and by Theorem 4.3.1a, $\varepsilon\text{-}SNB_{cosSim}^{D}(p_{(3)}) = \{p_{(1)}, p_{(8)}\}$. Similarly, one may determine ε-similarity neighborhood for the remaining vectors in D using already sorted set D'. □

Example 5.2 (Determination of k-similarity nearest neighbors). In this example, we will consider determination of k-similarity nearest neighbors $k\text{-}SNN_{cosSim}^{D}(p_{(3)})$, where $k = 2$, of vector $p_{(3)}$ in the same dataset $D = (p_{(1)}, \ldots, p_{(8)})$ as in Example 5.1 (see Figure 5.1 or Table 5.1). First, set $D' = (u_{(1)}, \ldots, u_{(8)})$ containing normalized forms of vectors from D needs to be determined. Figure 5.2 presents D'. Now, the determination of k-similarity nearest neighbors for vector $p_{(3)}$ in dataset D with the *cosSim* measure can be performed as the determination of k-nearest neighbors $k\text{-}NN_{Euclidean}^{D'}(u_{(3)})$ of vector $u_{(3)}$ in dataset D' of normalized vectors as formulated in Section 3.3 with regard to Euclidean distance. This procedure starts with ordering D' with regard to Euclidean distances of its vectors to a same reference vector r. In the example, we assume $r = [1, 0]$. Table 5.2 shows the result of sorting D' with regard to the distances of its vectors to r.

Let us assume that we have calculated the distances between $u_{(3)}$, and its directly preceding and following vectors in D'; that is, $u_{(1)}$ and $u_{(8)}$, respectively. These distances are as follows: $\mathit{Euclidean}(u_{(1)}, u_{(3)}) = 0.13$, $\mathit{Euclidean}(u_{(8)}, u_{(3)}) = 0.09$. Let $\varepsilon' = \max(\mathit{Euclidean}(u_{(1)}, u_{(3)}), \mathit{Euclidean}(u_{(8)}, u_{(3)}))$; that is, $\varepsilon' = 0.13$. Please, note that $u_{(1)}, u_{(8)} \in \varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(3)})$ and $|\varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(3)})| \geq k$. The latter fact implies that $\varepsilon'\text{-}NB_{Euclidean}^{D'}(u_{(3)})$ contains k-nearest neighbors of $u_{(3)}$ in D' (by Corollary 2.3.1). Nevertheless, it is not yet certain if $u_{(1)}$ and/or $u_{(8)}$ are these nearest neighbors of $u_{(3)}$ in D'.

As follows from Table 5.2, the first vector q_f following vector $u_{(3)}$ in D' for which $\mathit{Euclidean}(q_f, r) - \mathit{Euclidean}(u_{(3)}, r) > \varepsilon'$ is vector $u_{(4)}$ ($\mathit{Euclidean}(u_{(4)}, r) - \mathit{Euclidean}(u_{(3)}, r) = 1.15 - 0.87 = 0.28 > \varepsilon'$), and the first vector q_b preceding vector $u_{(3)}$ in D' for which $\mathit{Euclidean}(u_{(3)}, r) - \mathit{Euclidean}(q_b, r) > \varepsilon'$ is vector $u_{(6)}$ ($\mathit{Euclidean}(u_{(3)}, r) - \mathit{Euclidean}(u_{(6)}, r) = 0.87 - 0.68 = 0.19 > \varepsilon'$). Thus, by Proposition 3.3.1, neither vector $u_{(4)}$ nor the vectors following $u_{(4)}$ in D' as well as neither vector $u_{(6)}$ nor the vectors preceding $u_{(6)}$ in D' belong to $k\text{-}NN_{Euclidean}^{D'}(u_{(3)})$. Hence, only vectors $u_{(1)}$ and $u_{(8)}$ may belong to $k\text{-}NN_{Euclidean}^{D'}(u_{(3)})$ and only for these vectors it is necessary to know their real Euclidean distances to $u_{(3)}$. As $k = 2$, we should choose two vectors among those having a chance to belong to $k\text{-}NN_{Euclidean}^{D'}(u_{(3)})$ that are least distant from $u_{(3)}$ in D'. In our example, we have only $u_{(1)}$ and $u_{(8)}$ as such candidates, and they are real k-nearest neighbors of $u_{(3)}$ in D'. Thus, by Theorem 4.3.1c, $p_{(1)}$ and $p_{(8)}$ are k-similarity nearest neighbors of $p_{(3)}$ in D.

Please, note that in our example we had to calculate the Euclidean distance to vector $u_{(3)}$ only from two out of eight vectors in D'. □

Examples 5.1 and 5.2 illustrated determination of an ε-similarity neighborhood and k-similarity nearest neighbors of a given vector p in D. The determination of a k-similarity neighborhood of vector p in D would be very similar to the determination of its k-similarity nearest neighbors in D. There would be only slight difference in determining $k\text{-}NB_{Euclidean}^{D'}(u)$ and k-nearest neighbors $k\text{-}NN_{Euclidean}^{D'}(u)$ of u in D', where $u = NF(p)$ and D' is the set of normalized forms of vectors in D. Namely, if ε' was the radius of $k\text{-}NN_{Euclidean}^{D'}(u)$, then $k\text{-}NN_{Euclidean}^{D'}(u)$ would contain at least one vector in D', say v, such that $Euclidean(v, u) = \varepsilon'$, while $k\text{-}NB_{Euclidean}^{D'}(u)$ would contain all such vectors in D'. Please, see [5, 6] for the detailed description of the class of *TI-k-Neighborhood-Index* algorithms that enable efficient determination of $k\text{-}NB_{Euclidean}^{D'}(u)$ by means of the triangle inequality property.

6 Conclusions

In the paper, we have proposed a new solution to determining neighborhoods defined in terms of the cosine similarity measure. We have proposed and proved that this problem can be transformed to the problem of determining neighborhoods defined in terms of Euclidean distance, for which very efficient solutions based on the application of the triangle inequality property had been proposed recently in literature. More specifically, our approach to calculating neighborhoods defined in terms of the cosine similarity measure consists in introducing the preprocessing step in which the original vectors from set D are transformed to their normalized forms, which is done only once. In addition, a threshold value needs to be modified in the case of the task of determining ε-similarity neighborhoods. Once the preprocessing is carried out, the procedures of determining the neighborhoods can be carried out according to the methodology we proposed in [3, 4] and [5, 6], respectively. As a result, we conclude that the problem of determining neighborhoods defined in terms of the cosine similarity measure has the same complexity as the problem of determining neighborhoods defined in terms of a distance metric.

Acknowledgements

This work was supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: ,,Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

References

[1] Elkan, C: Using the Triangle Inequality to Accelerate k-Means. In. Proc. of ICML’03, Washington (2003) 147-153

[2] Han, J, Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann (2011)

[3] Kryszkiewicz, M., Lasek, P.: TI-DBSCAN: Clustering with DBSCAN by means of the triangle inequality. ICS Research Report 3/2010, Warsaw, April 2010

[4] Kryszkiewicz M., Lasek P.: TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality. RSCTC 2010: 60-69

[5] Kryszkiewicz M., Lasek P.: A Neighborhood-Based Clustering by Means of the Triangle Inequality. IDEAL 2010: 284-291

[6] Kryszkiewicz, M., Lasek, P.: A Neighborhood-Based Clustering by Means of the Triangle Inequality and Reference Points. ICS Research Report 3/2011, Warsaw, September 2011

[7] Moore, A. W., The Anchors Hierarchy: Using the Triangle Inequality to Survive High Dimensional Data. In: Proc. of UAI, Stanford (2000) 397–405

[8] Patra B.K., Hubballi N., Biswas S., Nandi S.: Distance Based Fast Hierarchical Clustering Method for Large Datasets. RSCTC 2010: 50-59