Chapter 1
Determining Cosine Similarity Neighborhoods by Means of the Euclidean Distance

Marzena Kryszkiewicz

Abstract Cosine similarity measure is often applied 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 or, alternatively, least dissimilar vectors. This operation has been commonly believed to be inefficient for large high dimensional datasets. However, using the triangle inequality to determine neighborhoods based on a distance metric, offered recently, makes this operation feasible for such datasets. Although the cosine similarity measure is not a distance metric and, in particular, violates the triangle inequality, in this chapter, we present how to determine cosine similarity neighborhoods of vectors by means of the Euclidean distance applied to ($\alpha-$)normalized forms of these vectors and by using the triangle inequality. 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.

Key words: k-nearest neighbors, ε-neighborhood, the cosine similarity measure, the Euclidean distance, the triangle inequality, normalized vector, data clustering, text clustering, high dimensional data

1.1 Introduction

Cosine similarity measure is often applied in the area of information retrieval, text classification, clustering, and ranking, where documents are usu-

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

ally 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 or, alternatively, least dissimilar vectors. This operation has been commonly believed to be inefficient for large high dimensional datasets. However, using the triangle inequality to determine neighborhoods based on a distance metric makes this operation feasible for such datasets [1, 5, 6, 7, 8, 10, 11].

Although, the cosine similarity measure is not a distance metric and, in particular, violates the triangle inequality, recently we have offered in the research report [9] how to determine cosine similarity neighborhoods of vectors by means of the Euclidean distance applied to normalized forms of these vectors and by using the triangle inequality. We have addressed there 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. This chapter is an extendend version of report [9]. In particular, we show here that cosine similarity neighborhoods may be determined by means of the Euclidean distance applied to $\alpha-$normalized forms of vectors, where $\alpha \neq 0$, and by using the triangle inequality. We also discuss here possible variants of the approach to calculating cosine similarity neighborhoods that was proposed in [9].

Our chapter has the following layout. Section 1.2 provides basic notions and properties used in the chapter. In particular, we examine properties and relationships among the three types of neighborhoods. In Section 1.3, we recall methods offered in [5, 6, 7, 8], which apply the triangle inequality property to efficiently calculate neighborhoods using a distance metric. The theoretical results which we derived in [9] for calculating cosine similarity neighborhoods based on the Euclidean distance as well as their current generalizations and consequences are presented in Section 1.4. In Section 1.5, we discuss and illustrate the ways of using these results for calculating the considered three types of cosine similarity neighborhoods. Section 1.6 concludes our work.

1.2 Basic Notions and Properties

1.2.1 Basic Operations on Vectors and Their Properties

In the chapter, 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 1.1, we recall definitions of basic operations on vectors. Table 1.2 presents their properties, which we will use in the chapter.

Table 1.1 Definitions of basic operations on vectors

Name of operation	Notation	Definition
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	αv	$[\alpha u_1,\ldots,\alpha u_n]$
division of vector u by scalar α	$\frac{u}{\alpha}$	$\frac{1}{\alpha}u$
standard vector dot product of vectors u and v	$u \cdot v$	$\sum_{i=1..n} u_i v_i$
length of vector u	$\|u\|$	$\sqrt{u \cdot u}$
normalized form of a vector u	$NF(u)$	$\frac{u}{\|u\|}$

Table 1.2 Properties of operations on vectors

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

A *normalized form of a vector u* will be denoted by $NF(u)$ and will be defined as the ratio of u to its length $\mid u \mid$. A *vector u* will be called a *normalized vector* (or alternatively, a *unit vector*) if $u = NF(u)$. Clearly, the length of a normalized vector is equal to 1.

In the chapter, we will also refer to notions of an α-*normalized form of a vector* and an α-*normalized vector*. An α-*normalized form of a vector u*, where $\alpha \neq 0$, is defined as $\alpha NF(u)$. A *vector u* is called an α-*normalized vector*, where $\alpha \neq 0$, if $u = \alpha NF(u)$. Clearly, the length of an α-normalized vector is equal to $\mid \alpha \mid$. If $\alpha = 1$, then obviously an α-normalized form of a vector is a normalized form of a vector and an α-normalized vector is a normalized vector.

1.2.2 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 are *distance metrics*.

A *distance metric* (or shortly, *distance*) in a set of vectors D is defined as a dissimilarity measure $dis : D \times D \rightarrow [0, +\infty)$ that satisfies the following three conditions for all vectors p, q, and r in D:

1) $dis(p,q) = 0 \Leftrightarrow p = q$;
2) $dis(p,q) = dis(q,p)$;
3) $dis(p,r) \leq dis(p,q) + dis(q,r)$.

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

Property 1.2.2.1(Alternative form of the triangle inequality). For any three vectors p, q, and r in a vector set D:

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

It was shown in [1, 5, 6, 7, 8, 10, 11] how to use this property for efficient clustering of both low and high dimensional data.

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:

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

Property 1.2.2.2. $Euclidean(u,v) = \sqrt{(u-v) \cdot (u-v)}$.

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, for example, $-sim(q,p)$ or $1 - sim(q,p)$ could be interpreted as a measure of dissimilarity between q and p.

In many applications, especially in text mining, *cosine similarity measure*, which is a function of the angle between two vectors, is applied. The *cosine similarity measure* between vectors u and v is denoted by $cosSim(u,v)$ and is defined as the cosine of the angle between them; that is,

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

Example 1.2.2.1 (The Euclidean distance and the cosine similarity). Figure 1.1 presents sample three vectors p, q, r. 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, since 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 of these vectors are presented in Table 1.3. One may note that neither $cosSim(p,q) \leq cosSim(p,r) + cosSim(r,q)$ nor $-cosSim(p,r) \leq -cosSim(p,q) + (-cosSim(q,r))$ nor $(1 - cosSim(p,r)) \leq (1 - cosSim(p,q)) + (1 - cosSim(q,r))$ (please, see Table 1.4). □

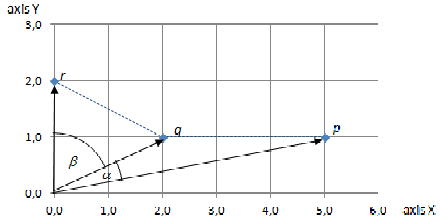

Fig. 1.1 The Euclidean distance and the cosine similarity

Table 1.3 The cosine similarity between vectors from Figure 1.1

(u,v)	$cosSim(u,v)$
(p,q)	0.964763821
(p,r)	0.196116135
(q,r)	0.447213595

Table 1.4 The cosine similarity versus the triangle inequality for vectors from Figure 1.1

the triangle inequality property	fulfilled?
$\boldsymbol{cosSim(p,\ q)\ \leq cosSim(p,\ r)\ +\ cosSim(r,\ q)}$	**no**
$cosSim(p,r) \leq cosSim(p,q) + cosSim(q,r)$	yes
$cosSim(q,r) \leq cosSim(q,p) + cosSim(p,r)$	yes
$-cosSim(p,q) \leq -cosSim(p,r) + (-cosSim(r,q))$	yes
$\boldsymbol{-cosSim(p,\ r) \leq -cosSim(p,\ q)\ +\ (-cosSim(q,\ r))}$	**no**
$\boldsymbol{-cosSim(q,\ r) \leq -cosSim(q,\ p)\ +\ (-cosSim(p,\ r))}$	**no**
$(1 - cosSim(p,q)) \leq (1 - cosSim(p,r)) + (1 - cosSim(r,q))$	yes
$\boldsymbol{(1-cosSim(p,\ r)) \leq (1-cosSim(p,\ q))\ +\ (1-cosSim(q,\ r)))}$	**no**
$(1 - cosSim(q,r)) \leq (1 - cosSim(q,p)) + (1 - cosSim(p,r))$	yes

Corollary 1.2.2.1.

a) It is false that the triangle inequality holds for $cosSim$ in each vector set.
b) It is false that the triangle inequality holds for $-cosSim$ in each vector set.
c) It is false that the triangle inequality holds for $(1-cosSim)$ in each vector set.

Clearly, the cosine similarity between any non-zero vectors u and v depends solely on the angle between them and does not depend on their lengths, hence the calculation of $cosSim(u,v)$ may be carried on $NF(u)$ and $NF(v)$ instead of u and v, respectively.

Property 1.2.2.3. Let u and v be non-zero vectors. Then:

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

1.2.3 Neighbourhoods Based on Dissimilarity Measures

Below, we provide definitions of neighborhoods in a given vector set D with respect to a given dissimilarity measure *dis*.

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

$$\varepsilon\text{-}NB^D_{dis}(p) = \{q \in D\backslash\{p\} \mid dis(p,q) \leq \varepsilon\}.$$

The set of all vectors in $D\backslash\{p\}$ that are less dissimilar from p than q will be denoted by $LessDissimilar^D_{dis}(p,q)$; that is,

$$LessDissimilar^D_{dis}(p,q) = \{s \in D\backslash\{p\} \mid dis(s,p) < dis(q,p)\}.$$

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

$$k\text{-}NB^D_{dis}(p) = \{q \in D\backslash\{p\} \mid \mid LessDissimilar^D_{dis}(p,q) \mid < k\}.$$

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

Proposition 1.2.3.1 [8]. Let $\varepsilon = max(\{dis(q,p) \mid q \in k\text{-}NB^D_{dis}(p)\})$. Then $k\text{-}NB^D_{dis}(p) = \varepsilon\text{-}NB^D_{dis}(p)$ and ε is the radius of $k\text{-}NB^D_{dis}(p)$.

Proposition 1.2.3.2 [8]. If $| \varepsilon\text{-}NB^D_{dis}(p) | \geq k$, then $\varepsilon\text{-}NB^D_{dis}(p) \supseteq k\text{-}NB^D_{dis}(p)$.

Please, note that $k\text{-}NB^D_{dis}(p)$ 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 $k\text{-}NB^D_{dis}(p)$.

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

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

Proposition 1.2.3.3 [9]. Let $k\text{-}NN^D_{dis}(p)$ be a set of k-nearest neighbors of a vector p in D and ε be the radius of $k\text{-}NN^D_{dis}(p)$. Then:

a) $k\text{-}NN^D_{dis}(p) \subseteq k\text{-}NB^D_{dis}(p)$;
b) $\forall q \in k\text{-}NB^D_{dis}(p) \backslash\ k\text{-}NN^D_{dis}(p), dis(q,p) = \varepsilon$;
c) $k\text{-}NB^D_{dis}(p) = \varepsilon\text{-}NB^D_{dis}(p)$;
d) ε is the radius of $k\text{-}NB^D_{dis}(p)$;
e) ε is the radius of each set of k-nearest neighbors of vector p in D.

Corollary 1.2.3.1 [9]. Let $k\text{-}NN^D_{dis}(p)$ be a set of k-nearest neighbors of a vector p in D.

a) If $| \varepsilon\text{-}NB^D_{dis}(p) | \geq k$, then $\varepsilon\text{-}NB^D_{dis}(p) \supseteq k\text{-}NB^D_{dis}(p) \supseteq k\text{-}NN^D_{dis}(p)$;
b) If $| \varepsilon\text{-}NB^D_{dis}(p) | = k$, then $\varepsilon\text{-}NB^D_{dis}(p) = k\text{-}NB^D_{dis}(p) = k\text{-}NN^D_{dis}(p)$.

By Corollary 1.2.3.1a, 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.

1.2.4 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 $\varepsilon\text{-}SNB^D_{sim}(p)$ and is defined as the set of all vectors in $D\backslash\{p\}$ that are similar to p by no less than ε; that is,

$$\varepsilon\text{-}SNB^D_{sim}(p) = \{q \in D\backslash\{p\} \mid sim(p,q) \geq \varepsilon\}.$$

The set of all vectors in $D\backslash\{p\}$ that are less similar to p than q will be denoted by $MoreSimilar^D_{sim}(p,q)$; that is,

$$MoreSimilar^D_{sim}(p,q) = \{s \in D\backslash\{p\} \mid sim(s,p) > sim(q,p)\}.$$

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

$$k\text{-}SNB^D_{sim}(p) = \{q \in D\backslash\{p\} \mid \mid MoreSimilar^D_{sim}(p,q) \mid < k\}.$$

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

Proposition 1.2.4.1 [9]. Let $\varepsilon = \min(\{sim(q,p) \mid q \in k\text{-}SNB^D_{sim}(p)\})$. Then $k\text{-}SNB^D_{sim}(p) = \varepsilon\text{-}SNB^D_{sim}(p)$ and ε is the similarity radius of $k\text{-}SNB^D_{sim}(p)$.

Proposition 1.2.4.2 [9]. If $\mid \varepsilon\text{-}SNB^D_{sim}(p)\mid \geq k$, then $\varepsilon\text{-}SNB^D_{sim}(p) \supseteq k\text{-}SNB^D_{sim}(p)$.

k-similarity nearest neighbors of a vector p in D are defined as a set of k vectors q in $D\backslash\{p\}$ such that the number of vectors in $D\backslash\{p\}$ that are more similar to p than q is less than k.

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

Proposition 1.2.4.3 [9]. Let $k\text{-}SNN^D_{sim}(p)$ be a set of k-similarity nearest neighbors of a vector p in D and ε be the similarity radius of $k\text{-}SNN^D_{sim}(p)$. Then:

a) $k\text{-}SNN^D_{sim}(p) \subseteq k\text{-}SNB^D_{sim}(p)$;
b) $\forall q \in k\text{-}SNB^D_{sim}(p)\backslash\ k\text{-}SNN^D_{sim}(p), sim(q,p) = \varepsilon$;
c) $k\text{-}SNB^D_{sim}(p) = \varepsilon\text{-}SNB^D_{sim}(p)$;
d) ε is the similarity radius of $k\text{-}SNB^D_{sim}(p)$;
e) ε is the similarity radius of each set of k-similarity nearest neighbors of vector p in D.

Corollary 1.2.4.1 [9]. Let $k\text{-}SNN^D_{sim}(p)$ be a set of k-similarity nearest neighbors of a vector p in D.

a) If $\mid \varepsilon\text{-}SNB^D_{sim}(p) \mid \geq k$, then $\varepsilon\text{-}SNB^D_{sim}(p) \supseteq k\text{-}SNB^D_{sim}(p) \supseteq k\text{-}SNN^D_{sim}(p)$;
b) If $\mid \varepsilon\text{-}SNB^D_{sim}(p) \mid = k$, then $\varepsilon\text{-}SNB^D_{sim}(p) = k\text{-}SNB^D_{sim}(p) = k\text{-}SNN^D_{sim}(p)$.

By Corollary 1.2.4.1a, if ε-similarity neighborhood of a vector p contains at least k vectors in D, 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.

1.3 The Triangle Inequality as a Mean for Efficient Determination of Neighborhoods Based on a Distance Metric

In this section, we present rudiments of the methods offered in [5, 6, 7, 8, 9] for speeding up the determination of the three types of neighboorhods based on a distance metric by employing the triangle inequality to efficiently prune non-promising candidates for neighbors. The methods guarantee that the pruning will not eliminate any true neighbor.

1.3.1 Efficient Determination of ε-Neighborhoods Based on a Distance Metric

In this subsection, we recall the basic method of determining ε-neighborhoods based on a distance metric efficiently, as proposed in [5, 6].

Lemma 1.3.1.1 [5, 6]. 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:

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

Lemma 1.3.1.1 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). Hence, $dis(p,q) > \varepsilon$. Thus, the fact that the difference of distances from two vectors p and q to any vector r is greater than ε implies that $q \notin \varepsilon\text{-}NB^D_{dis}(p)$ and $p \notin \varepsilon\text{-}NB^D_{dis}(q)$.

Now, let us consider a vector v such that $dis(v,r) > dis(p,r)$. If we know that $dis(p,r) - dis(q,r) > \varepsilon$, we may conclude that $dis(v,r) - dis(q,r) > \varepsilon$, and thus, $q \notin \varepsilon\text{-}NB^D_{dis}(v)$ and $v \notin \varepsilon\text{-}NB^D_{dis}(q)$ without calculating the real distance between v and q. This observation provides intuition behind Theorem 1.3.1.1, as offered in [5, 6].

Theorem 1.3.1.1 [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 respect 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 $\varepsilon\text{-}NB^D_{dis}(p)$;
b) q_b and all vectors preceding q_b in D do not belong to $\varepsilon\text{-}NB^D_{dis}(p)$.

As follows from Theorem 1.3.1.1, it makes sense to order all vectors in a given vector set D with respect to their distances to a 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.

Example 1.3.1.1. Let us consider a sample set D of two dimensional vectors presented in Figure 1.2. We will illustrate the usefulness of Theorem 1.3.1.1 for determining the ε-neighborhood based on the Euclidean distance for vector p = F in vector set D, given $\varepsilon = 0.5$. As a reference vector r, we will apply $[0, 0]$. Table 1.5 illustrates the considered set D ordered in a non-decreasing way with respect to the distances of its vectors to vector r. We note that $Euclidean(\mathrm{F}, r) = 3.2$, the first vector q_f following vector F in D such that $Euclidean(q_f, r) - Euclidean(\mathrm{F}, r) > \varepsilon$ is vector C ($Euclidean(\mathrm{C}, r) - Euclidean(\mathrm{F}, r) = 4.5 - 3.2 = 1.3 > \varepsilon$), and the first vector q_b preceding vector p in D such that $Euclidean(\mathrm{F}, r) - Euclidean(q_b, r) > \varepsilon$ is G ($Euclidean(\mathrm{F}, r) - Euclidean(\mathrm{G}, r) = 3.2 - 2.4 = 0.8 > \varepsilon$). By Theorem 1.3.1.1, vectors C and G as well as each vector which either follows C or precedes G in D certainly do not belong to ε-$NB^D_{Euclidean}(F)$. As a result, only vector H out of 8 vectors in D has a chance to belong to ε-$NB^D_{Euclidean}(\mathrm{F})$. So, H is the only vector for which it is necessary to calculate its actual distance to F in order to determine ε-$NB^D_{Euclidean}(\mathrm{F})$ properly. □

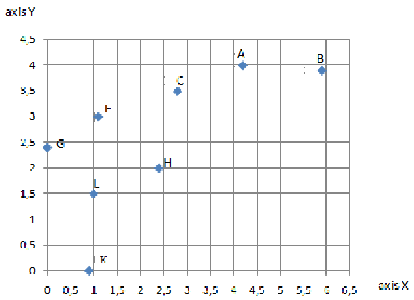

Fig. 1.2 Set of vectors D

The experimental evaluation of the usefulness of Theorem 1.3.1.1 was carried out in [5, 6] by comparing the performance of density based clustering carried out with the *TI-DBSCAN* algorithm and its variants, which used this theorem, and the *DBSCAN* algorithm [2] which used the *R-Tree* index [4]. As follows from the experiments reported in [5, 6], the algorithms using the theorem were always faster, and in almost all cases speeded up the clustering process by at least an order of magnitude, also for high dimensional large vector sets consisting of hundreds of dimensions and tens of thousands of vectors.

Table 1.5 Ordered set of vectors D from Figure 1.2 with their Euclidean distances to reference vector r[0,0]

Q	X	Y	distance(Q,r)
K	0.9	0.0	0.9
L	1.0	1.5	1.8
G	0.0	2.4	2.4
H	2.4	2.0	3.1
F	**1.1**	**3.0**	**3.2**
C	2.8	3.5	4.5
A	4.2	4.0	5.8
B	5.9	3.9	7.1

1.3.2 Efficient Determination of k-Neighborhoods Based on a Distance Metric

In this subsection, we recall the basics of the methods of determining ε-neighborhoods based on a distance metric efficiently, as proposed in [7, 8]. Also in this case, all vectors in a given vector set D are assumed to be ordered with respect to a 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 in a way similar to the one described in Subsection 1.3.1. Clearly, the real distances to p from vectors considered in phase 1, do not need to be 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 [7, 8], 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-neighborhoods of p. Please, see [7, 8] for a more detailed description. The presented approach was justified by Theorem 1.3.2.1, which we offered there.

Theorem 1.3.2.1 [7, 8]. Let dis be a distance metric, r be any vector and D be a set of vectors ordered in a non-decreasing way with respect to their distances to r. Let p be any vector in D and ε be a value such that $|\ \varepsilon\text{-}NB^{D}_{dis}(p)| \geq k, q_f$ 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 $k\text{-}NB^D_{dis}(p)$;
b) q_b and all vectors preceding q_b in D do not belong to $k\text{-}NB^D_{dis}(p)$.

Example 1.3.2.1. We will illustrate the usefulness of Theorem 1.3.2.1 for determining the k-neighborhood based on the Euclidean distance for vector p = F in vector set D from Figure 1.2, given $k = 3$. As a reference vector r, we will apply [0, 0]. Table 1.5 illustrates the considered set D ordered in a non-decreasing way with respect to the distance of its vectors to vector r. Let us assume that we have calculated the distances between F, and its two preceding vectors H, G, and one following vector C, respectively, and they are as follows: $Euclidean$(F, H) = 1.64, $Euclidean$(F, G) = 1.25, $Euclidean$(F, C) = 1.77. Now, we set the value of ε to max($Euclidean$(F, H), $Euclidean$(F, G), $Euclidean$(F, C)); that is, $\varepsilon = 1.77$. This means, that vectors H,G,C $\in \varepsilon\text{-}NB^D_{Euclidean}(F)$ and $\mid \varepsilon\text{-}NB^D_{Euclidean}(\mathrm{F}) \mid \geq k$. Thus, $k\text{-}NB^D_{Euclidean}$(F) will be found within the ε radius from F. Now, we note that the first vector q_f following vector F in D such that $Euclidean(q_f, r) - Euclidean(\mathrm{F}, r) > \varepsilon$ is vector A ($Euclidean(\mathrm{A}, r) - Euclidean(\mathrm{F}, r) = 5.8 - 3.2 = 2.6 > \varepsilon$), and the first vector q_b preceding vector F in D such that $Euclidean(\mathrm{F}, r) - Euclidean(q_b, r) > \varepsilon$ is K ($Euclidean(\mathrm{F}, r) - distance(\mathrm{K}, r) = 3.2 - 0.9 = 2.3 > \varepsilon$). By Theorem 1.3.2.1, vectors A, K as well as each vector that either follows A (that is, vector B) or precedes K in D (here, no vector precedes K) do not belong to $k\text{-}NB^D_{Euclidean}(F)$. □

The experimental evaluation of the usefulness of Theorem 1.3.2.1 was carried out in [7, 8] by comparing the performance of the *NBC* density based clustering [13] carried out with differently calculated k-neighborhoods; namely by means of the index created by the *TI-k-Neighboorhood-Index* algorithm and its variants [7, 8], which used this theorem, as well as by means of the *VA-File* index [3] and the *R-Tree* index [4]. As follows from the experiments reported in [7, 8], the algorithms determining k-neighborhoods by means of the theorem were always faster, and in almost all cases speeded up the clustering process by at least an order of magnitude, also for high dimensional large vector sets consisting of hundreds of dimensions and tens of thousands of vectors.

1.3.3 Efficient Determination of k-Nearest Neighbors Based on a Distance Metric

Let $k\text{-}NN^D_{dis}(p)$ be a set of k-similarity nearest neighbors of a vector p in D. Since $k\text{-}NB^D_{dis}(p) \supseteq k\text{-}NN^D_{dis}(p)$, then the vectors that do not belong to $k\text{-}NB^D_{dis}(p)$, do not belong $k\text{-}NN^D_{dis}(p)$ either. This observation allows us to derive Proposition 1.3.3.1 from Theorem 1.3.2.1.

Proposition 1.3.3.1 [9]. Let *dis* be a distance metric, r be any vector and D be a set of vectors ordered in a non-decreasing way with respect to their distances to r. Let p be any vector in D, $k\text{-}NN^D_{dis}(p)$ be a set of k-similarity nearest neighbors of vector p in D and ε be a value such that $\mid \varepsilon\text{-}NB^D_{dis}(p) \mid \geq k$, 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 $k\text{-}NN^D_{dis}(p)$;
b) q_b and all vectors preceding q_b in D do not belong to $k\text{-}NN^D_{dis}(p)$.

Clearly, if $|k\text{-}NB^D_{dis}(p) \mid = k$, then $k\text{-}NN^D_{dis}(p) = k\text{-}NB^D_{dis}(p)$. Otherwise, $k\text{-}NN^D_{dis}(p)$ can be obtained by calculating $k\text{-}NB^D_{dis}(p)$, for instance, as presented in Section 1.3.2 and then removing all but one of the vectors from $k\text{-}NB^D_{dis}(p)$ that are distant from p by the radius of $k\text{-}NB^D_{dis}(p)$.

1.4 The Cosine Similarity Measure and Neighborhoods versus the Euclidean Distance and Neighborhoods

In Section 1.3, 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 1.2.2.1, the cosine similarity measure is not a distance metric as the triangle inequality is not guaranteed to hold for it. However, in this section, we will present that cosine similarity neighborhoods are equivalent to corresponding neighborhoods based on the Euclidean distance.

1.4.1 Relationship between the Cosine Similarity and the Euclidean Distance

We start with formulating and proving two lemmas: the one showing that the cosine similarity between two non-zero vectors can be expressed as a function of their lengths and the Euclidean distance between them (Lemma 1.4.1.1), and the next one showing that the cosine similarity between two normalized (forms of) non-zero vectors can be expressed as a function of solely their Euclidean distance (Lemma 1.4.1.2).

Lemma 1.4.1.1 [9]. Let u and v be non-zero vectors. Then:

$$\text{a) } cosSim(u, v) = \frac{(u \cdot u) + (v \cdot v) - (u - v) \cdot (u - v)}{2 \mid u \parallel v \mid};$$

$$\text{b) } cosSim(u,v) = \frac{\mid u \mid^2 + \mid v \mid^2 - Euclidean^2(u,v)}{2 \mid u \parallel v \mid}.$$

Proof. Ad a) Since $(u-v)\cdot(u-v) = (u\cdot u)+(v\cdot v)-2(u\cdot v)$ and $cosSim(u,v) = \frac{u\cdot v}{|u\|v|}$, then $(u-v)\cdot(u-v) = (u\cdot u)+(v\cdot v)-2(cosSim(u,v) \mid u \parallel v \mid)$. Hence, $cosSim(u,v) = \frac{(u\cdot u)+(v\cdot v)-(u-v)\cdot(u-v)}{2|u\|v|}$.
Ad b) Follows immediately from Lemma 1.4.1.1a, the fact that $u \cdot u = \mid u \mid^2$, $v \cdot v = \mid v \mid^2$ and $Euclidean(u,v) = \sqrt{(u-v)\cdot(u-v)}$. □

Lemma 1.4.1.2 [9]. Let u, v be non-zero vectors. Then:

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

Proof. $cosSim(NF(u), NF(v)) = \frac{|NF(u)|^2+|NF(v)|^2-Euclidean^2(NF(u),NF(v))}{2|NF(u)\|NF(v)|}$ by Lemma 1.4.1.1b. Hence, since $\mid NF(u) \mid = \mid NF(v) \mid = 1$, we conclude that $cosSim(NF(u), NF(v)) = \frac{2-Euclidean^2(NF(u),NF(v))}{2}$. □

Now, as the cosine similarity of two non-zero vectors is equal to the cosine similarity of their normalized forms, we conclude from Lemma 1.4.1.2 that the cosine similarity between two non-zero vectors can be expressed as a function of solely the Euclidean distance of their normalized forms (Theorem 1.4.1.1).

Theorem 1.4.1.1 [9]. Let *u,v* be non-zero vectors. Then:

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

Also, based on Theorem 1.4.1.1, we may conclude further that the cosine similarity between two non-zero vectors can be expressed as a function of solely α and the Euclidean distance of their $\alpha-$normalized forms (Proposition 1.4.1.1).

Proposition 1.4.1.1. Let $\alpha \neq 0$ and *u,v* be non-zero vectors. Then:

$$cosSim(u,v) = \frac{2 - \frac{1}{\alpha^2}Euclidean^2(\alpha NF(u), \alpha NF(v))}{2}.$$

Proof. By Theorem 1.4.1.1, $cosSim(u,v) = \frac{2-Euclidean^2(NF(u),NF(v))}{2} = \frac{2-Euclidean^2(\frac{\alpha NF(u)}{\alpha},\frac{\alpha NF(v)}{\alpha})}{2} = \frac{2-\frac{1}{\alpha^2}Euclidean^2(\alpha NF(u),\alpha NF(v))}{2}$. □

Corollary 1.4.1.1. Let $\alpha \neq 0$ and *u,v* be α-normalized non-zero vectors. Then:

$$cosSim(u,v) = \frac{2 - \frac{1}{\alpha^2}Euclidean^2((u),(v))}{2}.$$

1.4.2 Vector Cosine Similarity Neighborhoods and Normalized Vector Neighborhoods based on the Euclidean Distance

In this subsection, we will use Theorem 1.4.1.1 to derive relationships between vector cosine similarity neighborhoods and corresponding normalized vector neighborhoods based on the Euclidean distance [9]. First, we start with Lemma 1.4.2.1a, in which we formulate and prove that a comparison of the cosine similarity between two non-zero vectors with an ε threshold is equivalent to a comparison of the Euclidean distance between their normalized forms with an ε' threshold being a function of ε. In Lemma 1.4.2.1b, we formulate and prove that a comparison of the cosine similarity between any two non-zero vectors s and p with the cosine similarity between vector p and any vector q is equivalent to a comparison of the Euclidean distances between their normalized forms.

Lemma 1.4.2.1 [9]. Let p, q and s be non-zero vectors, $\varepsilon \in [-1, 1]$ and $\varepsilon' = \sqrt{2 - 2\varepsilon}$. Then:

a) $cosSim(p, q) \geq \varepsilon$ iff $Euclidean(NF(p), NF(q)) \leq \varepsilon'$;
b) $cosSim(s, p) > cosSim(q, p)$ iff $Euclidean(NF(s), NF(p)) < Euclidean(NF(q), NF(p))$.

Proof. Ad a) $cosSim(p, q) \geq \varepsilon$ iff /* by Theorem 1.4.1.1 */ $\frac{2 - Euclidean^2(NF(p), NF(q))}{2} \geq \varepsilon$ iff $Euclidean(NF(p), NF(q)) \leq \sqrt{2 - 2\varepsilon} = \varepsilon'$;
Ad b) $cosSim(s, p) > cosSim(q, p)$ iff /* by Theorem 1.4.1.1 */ $\frac{2 - Euclidean^2(NF(s), NF(p))}{2} > \frac{2 - Euclidean^2(NF(q), NF(p))}{2}$ iff $Euclidean(NF(s), NF(p)) < Euclidean(NF(q), NF(p))$. □

Lemma 1.4.2.1 enables us to formulate and prove Lemma 1.4.2.2, in which we show that the problem of determining vectors in D that are more cosine similar to a given vector $p_{(i)}$ than another vector $p_{(j)}$ can be treated as the problem of determining the normalized forms of vectors from D that are less distant in the Euclidean sense from $NF(p_{(i)})$ than $NF(p_{(j)})$.

Lemma 1.4.2.2 [9]. Let D be an ordered set of m non-zero 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 for any vectors $p_{(i)}, p_{(j)}$ in D:

a) $MoreSimilar^{D}_{cosSim}(p_{(i)}, p_{(j)}) = \{p_{(l)} \in D \backslash \{p_{(i)}\} \mid u_{(l)} \in LessDissimilar^{D'}_{Euclidean}(u_{(i)}, u_{(j)})\}$;
b) $\mid MoreSimilar^{D}_{cosSim}(p_{(i)}, p_{(j)}) \mid < k$ iff $\mid LessDissimilar^{D'}_{Euclidean}(u_{(i)}, u_{(j)}) \mid < k$.

Proof. Ad a) $MoreSimilar^{D}_{cosSim}(p_{(i)}, p_{(j)}) = \{p_{(l)} \in D \backslash \{p_{(i)}\} \mid cosSim(p_{(l)}, p_{(i)}) > cosSim(p_{(j)}, p_{(i)})\} =$

/* by Lemma 1.4.2.1b */
$\{p_{(l)} \in D\backslash\{p_{(i)}\} \mid u_{(l)} \in D'\backslash\{u_{(i)}\} \wedge Euclidean(u_{(l)}, u_{(i)})) <$
$Euclidean(u_{(j)}, u_{(i)})\} =$
$\{p_{(l)} \in D\backslash\{p_{(i)}\} \mid u_{(l)} \in LessDissimilar^{D'}_{Euclidean}(u_{(i)}, u_{(j)})\}$.
Ad b) Follows immediately from Lemma 1.4.2.2a. □

Now, we are ready to formulate and prove the equivalence of cosine similarity neighborhoods and corresponding neighborhoods based on the Euclidean distance.

Theorem 1.4.2.1 [9]. Let D be an ordered set of m non-zero 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) $\varepsilon\text{-}SNB^{D}_{cosSim}(p_{(i)}) = \{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in \varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$;
b) $k\text{-}SNB^{D}_{cosSim}(p_{(i)}) = \{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$;
c) If $k\text{-}NN^{D'}_{Euclidean}(u_{(i)})$ is a set of k-nearest neighbours of $u_{(i)}$ in D', then $\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\}$ is a set of k-similarity nearest neighbors of $p_{(i)}$ in D.

Proof. Ad a) $\varepsilon\text{-}SNB^{D}_{cosSim}(p_{(i)}) =$
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid cosSim(p_{(i)}, p_{(j)}) \geq \varepsilon\} =$ /* by Lemma 1.4.2.1a */ $=$
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in D'\backslash\{u_{(i)}\} \wedge Euclidean(u_{(i)}, u_{(j)}) \leq \varepsilon'\} =$
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in \varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$.
Ad b) $k\text{-}SNB^{D}_{cosSim}(p_{(i)}) =$
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid\mid MoreSimilar^{D}_{cosSim}(p_{(i)}, p_{(j)}) \mid < k\} =$
/* by Lemma 1.4.2.2b */ $= \{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in D'\backslash\{u_{(i)}\} \wedge$
$\mid LessDissimilar^{D'}_{Euclidean}(u_{(i)}, u_{(j)}) \mid < k\} =$
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$.
Ad c) Let $k\text{-}NN^{D'}_{Euclidean}(u_{(i)})$ be a set of k-nearest neighbours of $u_{(i)}$ in D'.
Then, $k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\}$ contains k vectors and thus
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\}$ also contains k vectors.
In addition, the fact that $\{k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\} \subseteq \{k\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$
implies that $\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\} \subseteq$
$\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NB^{D'}_{Euclidean}(u_{(i)})\} =$
/* by Theorem 1.4.2.1b */ $= k\text{-}SNB^{D}_{cosSim}(p_{(i)})$.
Therefore, $\{p_{(j)} \in D\backslash\{p_{(i)}\} \mid u_{(j)} \in k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\}$ is a set of k-cosine similarity nearest neighbors of $p_{(i)}$ in D. □

One may easily observe that the equivalence of cosine similarity neighborhoods and neighborhoods based on the Euclidean distance that was stated in Theorem 1.4.2.1 becomes the equality in the case of normalized vectors.

Corollary 1.4.2.1. Let D be a set of normalized non-zero vectors, $p \in D$, $\varepsilon \in$[-1,1] and $\varepsilon' = \sqrt{2 - 2\varepsilon}$. Then:

a) $\varepsilon\text{-}SNB^{D}_{cosSim}(p) = \varepsilon'\text{-}NB^{D}_{Euclidean}(p)$;

b) $k\text{-}SNB^{D}_{cosSim}(p) = k\text{-}NB^{D}_{Euclidean}(p)$;
c) NN is a set of k-cosine similarity nearest neighbours of p in D iff NN is a set of k-nearest neighbours of p in D with regard to the Euclidean distance.

1.4.3 Vector Cosine Similarity Neighborhoods and α-Normalized Vector Neighborhoods based on the Euclidean Distance

In this subsection, we will generalize the results from Section 1.4.2 in that we will derive relationships between vector cosine similarity neighborhoods and corresponding α-normalized vector neighborhoods based on the Euclidean distance.

Lemma 1.4.3.1. Let $\alpha \neq 0$, p, q and s be non-zero vectors, $\varepsilon \in [-1, 1]$ and $\varepsilon' = \mid \alpha \mid \sqrt{2 - 2\varepsilon}$. Then:

a) $cosSim(p, q) \geq \varepsilon$ iff $Euclidean(\alpha NF(p), \alpha NF(q)) \leq \varepsilon'$;
b) $cosSim(s, p) > cosSim(q, p)$ iff $Euclidean(\alpha NF(s), \alpha NF(p)) < Euclidean(\alpha NF(q), \alpha NF(p))$.

Proof. Follows from Proposition 1.4.1.1. Can be proved in analogous way as Lemma 1.4.2.1. □

Lemma 1.4.3.2. Let $\alpha \neq 0$, D be an ordered set of m non-zero vectors $(p_{(1)}, \ldots, p_{(m)})$, D' be the ordered set of m vectors $(u_{(1)}, \ldots, u_{(m)})$ such that $u_{(i)} = \alpha NF(p_{(i)})$, $i = 1..m$, $\varepsilon \in$ [-1, 1] and $\varepsilon' = \mid \alpha \mid \sqrt{2 - 2\varepsilon}$. Then for any vectors $p_{(i)}, p_{(j)}$ in D:

a) $MoreSimilar^{D}_{cosSim}(p_{(i)}, p_{(j)}) = \{p_{(l)} \in D \backslash \{p_{(i)}\} \mid u_{(l)} \in LessDissimilar^{D'}_{Euclidean}(u_{(i)}, u_{(j)})\}$;
b) $\mid MoreSimilar^{D}_{cosSim}(p_{(i)}, p_{(j)}) \mid < k$ iff $\mid LessDissimilar^{D'}_{Euclidean}(u_{(i)}, u_{(j)}) \mid < k$.

Proof. Ad a) Follows from Lemma 1.4.3.1b.
Ad b) Follows immediately from Lemma 1.4.3.2a. □

Theorem 1.4.3.1. Let $\alpha \neq 0$, D be an ordered set of m non-zero vectors $(p_{(1)}, \ldots, p_{(m)})$, D' be the ordered set of m vectors $(u_{(1)}, \ldots, u_{(m)})$ such that $u_{(i)} = \alpha NF(p_{(i)})$, $i = 1..m$, $\varepsilon \in$[-1,1] and $\varepsilon' = \mid \alpha \mid \sqrt{2 - 2\varepsilon}$. Then:

a) $\varepsilon\text{-}SNB^{D}_{cosSim}(p_{(i)}) = \{p_{(j)} \in D \backslash \{p_{(i)}\} \mid u_{(j)} \in \varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$;
b) $k\text{-}SNB^{D}_{cosSim}(p_{(i)}) = \{p_{(j)} \in D \backslash \{p_{(i)}\} \mid u_{(j)} \in k\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$;
c) If $k\text{-}NN^{D'}_{Euclidean}(u_{(i)})$ is a set of k-nearest neighbours of $u_{(i)}$ in D', then $\{p_{(j)} \in D \backslash \{p_{(i)}\} \mid u_{(j)} \in k\text{-}NN^{D'}_{Euclidean}(u_{(i)})\}$ is a set of k-similarity nearest neighbors of $p_{(i)}$ in D.

Proof. Analogous to the proof of Theorem 1.4.2.1.
Ad a) Follows from Lemma 1.4.3.1a.

Ad b) Follows from Lemma 1.4.3.2b.
Ad c) Follows from Theorem 1.4.3.1b. □

Corollary 1.4.3.1. Let D be a set of α-normalized non-zero vectors, $p \in D$, $\varepsilon \in$[-1,1] and $\varepsilon' = \mid \alpha \mid \sqrt{2-2\varepsilon}$. Then:

a) ε-$SNB^D_{cosSim}(p) = \varepsilon'$-$NB^D_{Euclidean}(p)$;
b) k-$SNB^D_{cosSim}(p) = k$-$NB^D_{Euclidean}(p)$;
c) NN is a set of k-cosine similarity nearest neighbours of p in D iff NN is a set of k-nearest neighbours of p in D with regard to the Euclidean distance.

1.5 Determination of Cosine Similarity Neighborhoods as Determination of Neighborhoods Based on the Euclidean Distance

Theorem 1.4.3.1 (Theorem 1.4.2.1) tells us that cosine similarity neighborhoods in vector set D can be determined as respective neighborhoods based on the Euclidean distance in vector set D' consisting of $\alpha-$normalized (normalized) forms of the vectors from D. Thus, we propose the following approach to determination of cosine similarity neighborhoods:

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

Examples 1.5.1 and 1.5.2 illustrate this approach to determining cosine similarity neigboorhoods.

Example 1.5.1 (Determination of an ε-cosine similarity neighborhood). In this example, we will consider determination of ε-cosine similarity neighborhood of vector $p_{(3)}$ in vector set $D = (p_{(1)}, \ldots, p_{(8)})$ from Figure 1.3 (and Table 1.6) applying, for instance, $\alpha = 1$. We assume that the cosine similarity threshold $\varepsilon = 0.9856$, which roughly corresponds to the angle 9.74°. Figure 1.4 shows set $D' = (u_{(1)}, \ldots, u_{(8)})$ that contains $\alpha-$normalized forms of vectors from D. Clearly, the lengths of all of them are equal to $\mid \alpha \mid$; that is, 1. Now, we will determine the corresponding Euclidean distance threshold ε' as $\mid \alpha \mid \sqrt{2-2\varepsilon}$ (according to Theorem 1.4.3.1a). Hence, $\varepsilon' \approx 0.1697 \leq 0.17$. At this moment, we may start the procedure of determining ε-cosine similarity neighborhood for vector $p_{(3)}$ in vector set D as the procedure of determining ε'-neighborhood for vector $u_{(3)}$ in vector set D' of $\alpha-$normalized vectors with regard to the 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 1.7 shows set D' ordered in a non-decreasing way with regard to the Euclidean distances of its vectors to vector r.

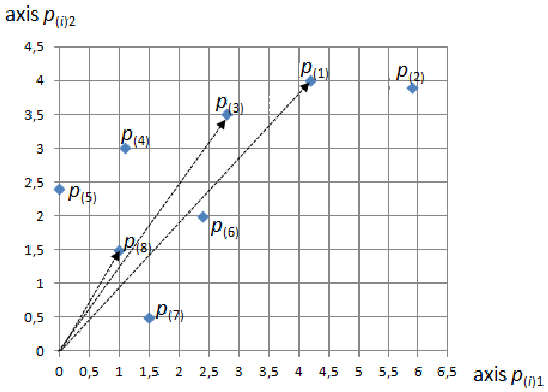

Fig. 1.3 Sample set D of vectors

Table 1.6 Sample set D

Vector $p_{(i)}$	$p_{(i)1}$	$p_{(i)2}$
$p_{(1)}$	4.20	4.00
$p_{(2)}$	5.90	3.90
$\boldsymbol{p_{(3)}}$	**2.80**	**3.50**
$p_{(4)}$	1.10	3.00
$p_{(5)}$	0.00	2.40
$p_{(6)}$	2.40	2.00
$p_{(7)}$	1.50	0.50
$p_{(8)}$	1.00	1.50

As follows from Table 1.7, the first vector q_f following vector $u_{(3)}$ in D' for which $Euclidean(q_f, r) - Euclidean(u_{(3)}, r) > \varepsilon'$ is vector $u_{(4)}$ ($Euclidean(u_{(4)}, r) - Euclidean(u_{(3)}, r) = 1.15 - 0.87 = 0.28 > \varepsilon'$), and the first vector q_b preceding vector $u_{(3)}$ in D', for which $Euclidean(u_{(3)}, r) - Euclidean(q_b, r) > \varepsilon'$ is vector $u_{(6)}(Euclidean(u_{(3)}, r) - Euclidean(u_{(6)}, r) = 0.87 - 0.68 = 0.19 > \varepsilon'$). Thus, by Theorem 1.3.1.1, neither vector $u_{(4)}$ nor the vectors following $u_{(4)}$ in D' as well as neither vector $u_{(6)}$ nor the vectors

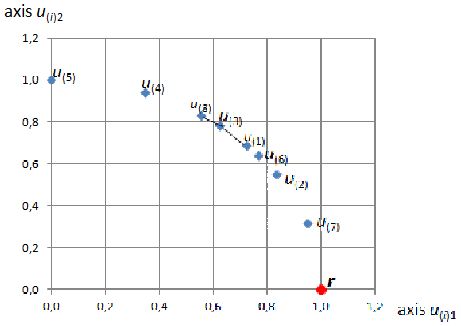

Fig. 1.4 Set D' containing normalized forms of vectors from D

Table 1.7 Normalized vectors in set D' sorted w.r.t. their distances to vector $r = [1, 0]$

Vector $u_{(i)}$	$u_{(i)1}$	$u_{(i)2}$	$Euclidean(u_{(i)}, r)$
$u_{(7)}$	0.95	0.32	0.32
$u_{(2)}$	0.83	0.55	0.58
$u_{(6)}$	0.77	0.64	0.68
$u_{(1)}$	0.72	0.69	0.74
$\boldsymbol{u_{(3)}}$	**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

preceding $u_{(6)}$ in D' belong to ε'-$NB^{D'}_{Euclidean}(u_{(3)})$. Hence, only vectors $u_{(1)}$ and $u_{(8)}$ may belong to ε'-$NB^{D'}_{Euclidean}(u_{(3)})$ and only for these vectors it is necessary to calculate their real Euclidean distances to $u_{(3)}$. These distances are as follows, $Euclidean(u_{(1)}, u_{(3)}) = 0.13$ and $Euclidean(u_{(8)}, u_{(3)}) = 0.07$. Since both values are less than ε', then ε'-$NB^{D'}_{Euclidean}(u_{(3)}) = \{u_{(1)}, u_{(8)}\}$, and by Theorem 1.4.3.1a, ε-$SNB^{D}_{cosSim}(p_{(3)}) = \{p_{(1)}, p_{(8)}\}$. Similarly, one may determine ε-cosine similarity neighborhood for the remaining vectors in D using already sorted set D'. □

Example 1.5.2 (Determination of a k-cosine similarity neighborhood and k-cosine similarity nearest neighbors). In this example, we will first consider determination of a k-cosine similarity neighborhood, where $k = 2$, of vector $p_{(3)}$ in the vector set $D = (p_{(1)}, \ldots, p_{(8)})$ from Figure 1.3 (and Table 1.6). Then we will determine k-cosine similarity nearest neigbors k-$SNN^{D}_{cosSim}(p_{(3)})$. In the example, we will apply $\alpha = 1$.

We start with the calculation of set $D' = (u_{(1)}, \ldots, u_{(8)})$ of $\alpha-$normalized forms of vectors from D. Figure 1.4 presents D'. Now, the determination of k-cosine similarity neighborhood for vector $p_{(3)}$ in vector set D can be per-

formed as the determination of k-neighborhood $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ of vector $u_{(3)}$ in set D' of $\alpha-$normalized vectors with regard to the Euclidean distance. This procedure starts with ordering D' with regard to the Euclidean distances of its vectors to a same reference vector r. In the example, we assume $r = [1, 0]$. Table 1.7 shows the result of this sorting.

Now, we need to estimate the radius within which k-nearest neighbors of $u_{(3)}$ occur. 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: $Euclidean(u_{(1)}, u_{(3)}) = 0.13$ and $Euclidean(u_{(8)}, u_{(3)}) = 0.07$. Let $\varepsilon' = \max(Euclidean(u_{(1)}, u_{(3)}), Euclidean(u_{(8)}, u_{(3)}))$; that is, $\varepsilon' = 0.13$. Please, note that $u_{(1)}, u_{(8)} \in \varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ and $\mid \varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(3)}) \mid \geq k$. The latter fact implies that $\varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ contains k-neighborhood of $u_{(3)}$ in D' (by Corollary 1.2.3.1). Nevertheless, it is yet not certain if $u_{(1)}$ and/or $u_{(8)}$ belong to this neigborhood of $u_{(3)}$ in D'.

As follows from Table 1.7, the first vector q_f following vector $u_{(3)}$ in D' for which $Euclidean(q_f, r) - Euclidean(u_{(3)}, r) > \varepsilon'$ is vector $u_{(4)}$ ($Euclidean(u_{(4)}, r) - Euclidean(u_{(3)}, r) = 1.15 - 0.87 = 0.28 > \varepsilon'$), and the first vector q_b preceding vector $u_{(3)}$ in D', for which $Euclidean(u_{(3)}, r) - Euclidean(q_b, r) > \varepsilon'$ is vector $u_{(6)}$ ($Euclidean(u_{(3)}, r) - Euclidean(u_{(6)}, r) = 0.87 - 0.68 = 0.19 > \varepsilon'$). Thus, by Theorem 1.3.2.1, 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 $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$. Hence, only vectors $u_{(1)}$ and $u_{(8)}$ may belong to $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ and only for these vectors it is necessary to calculate their real Euclidean distances to $u_{(3)}$. As $k = 2$ and only two vectors: $u_{(1)}$ and $u_{(8)}$ were not eliminated, they constitute $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$. Thus, by Theorem 1.4.3.1b, $k\text{-}SNB^{D}_{cosSim}(p_{(3)}) = \{p_{(1)}, p_{(8)}\}$.

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'.

Now, let us consider the determination of $k\text{-}SNN^{D}_{cosSim}(p_{(3)})$. In our example, $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$, and consequently $k\text{-}SNB^{D}_{cosSim}(p_{(3)})$, have exactly k vectors each. Hence, $k\text{-}SNN^{D}_{cosSim}(p_{(3)}) = k\text{-}SNB^{D}_{cosSim}(p_{(3)}) = \{p_{(1)}, p_{(8)}\}$. Beneath, however, we present how $k\text{-}SNN^{D}_{cosSim}(p_{(3)})$ could be determined if we ignored the fact that the cardinality of $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ (and the cardinality of $k\text{-}SNB^{D}_{cosSim}(p_{(3)})\}$) is k. First, one could determine k-nearest neighbors of $(u_{(3)})$ in D' based on $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$, the radius of which is known and equals the maximum of the Euclidean distances between $u_{(3)}$ and the vectors in $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$; that is, 0.13. Thus, k-nearest neighbors of $u_{(3)}$ in D' would contain all vectors from $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ that are less distant from $u_{(3)}$ than 0.13 (here: only vector $u_{(8)}$) and exactly one arbitrary vector in $k\text{-}NB^{D'}_{Euclidean}(u_{(3)})$ that is distant from $u_{(3)}$ by 0.13 (here:

vector ($u_{(1)}$). Next, by Theorem 1.4.3.1c, $k\text{-}SNN^{D}_{cosSim}(p_{(3)})$ would be found as $\{p_{(1)}, p_{(8)}\}$. □

Please note that it is not necessary to verify if an $\alpha-$normalized vector $u_{(l)}$ is a true neighbor of an analyzed $\alpha-$normalized vector $u_{(i)}$ in D' by calculating the Euclidean distance between them. The verification can be carried out by calculating $cosSim(u_{(l)}, u_{(i)})$ as $\frac{1}{\alpha^2}(u_{(l)} \cdot u_{(i)})$ (or simply $u_{(l)} \cdot u_{(i)}$ if vectors $u_{(l)}, u_{(i)}$ are normalized) and, eventually, employing the cosine threshold ε provided $\varepsilon' = \mid \alpha \mid \sqrt{2 - 2\varepsilon}$. In fact, it is not even necessary to store $\alpha-$normalized forms of vectors from D. It is sufficient to store only a sorted index containing the Euclidean distances of $\alpha-$normalized forms of vectors from D to a reference vector. When the $\alpha-$normalized forms are not stored, the verification for a (non-stored) $\alpha-$normalized vector $u_{(l)}$ can be carried on the corresponding original vectors in D as follows: $cosSim(u_{(l)}, u_{(i)}) = cosSim(p_{(l)}, p_{(i)}) = \frac{p_{(l)} \cdot p_{(i)}}{|p_{(l)}||p_{(i)}|}$.

1.6 Conclusions

In this chapter, we have offered a new solution to determining vector cosine similarity neighborhoods that consists in transforming the original problem into the problem of determining neighborhoods of ($\alpha-$)normalized forms of the original vectors with regard to the Euclidean distance. We have discussed possible variants of the approach to calculating cosine similar neighborhoods that was proposed in [9] and was based on applying normalized forms of vectors. The fact that the the problem of determining cosine similarity neighborhoods is transformable to the problem of determining neighborhoods based on the Euclidean distance allows applying the triangle inequality, which was proved in [1, 5, 6, 7, 8, 10, 11] to be a powerful tool for making the neighboorhod determination efficient even in the case of high dimensional large vector sets consisting of hundreds of dimensions and tens of thousands of vectors. As a consequence, our solution helps to surpass the curse of dimensionality in the case of determining cosine similarity neighborhoods.

Acknowledgements

This work was supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 devoted to 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: Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), August 21-24, Washington, DC, USA, pp. 147–153. AAAI Press (2003)
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial database with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), August 2-4, Portland, Oregon, USA, pp. 226–231. AAAI Press (1996)
3. Blott, S., Weber, R.: A simple vector approximation file for similarity search in high-dimensional vector spaces. Technical Report 19, ESPRIT project HERMES, vol. 9141 (1997)
4. Guttman, A.: R-Trees: A dynamic index structure for spatial searching. In: SIGMOD'84, Proceedings of Annual Meeting, June 18-21, Boston, Massachusetts, USA pp. 4757. ACM SIGMOD (1984)
5. Kryszkiewicz, M., Lasek, P.: TI-DBSCAN: Clustering with DBSCAN by means of the triangle inequality. ICS Research Report 3, Institute of Computer Science, Warsaw University of Technology, Warsaw (2010)
6. Kryszkiewicz M., Lasek P.: TI-DBSCAN: Clustering with DBSCAN by means of the triangle inequality. In: Proceedings of 7th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2010), June 28-30, Warsaw, Poland, *Lecture Notes in Computer Science*, vol. 6086, pp. 60–69. Springer-Verlag, Berlin, Heildelberg, Germany (2010)
7. Kryszkiewicz M., Lasek P.: A neighborhood-based clustering by means of the triangle inequality. In: Proceedings of 11th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2010), September 1-3, Paisley, UK, *Lecture Notes in Computer Science*, vol. 6283, pp. 284–291. Springer-Verlag, Berlin, Heildelberg, Germany (2010)
8. Kryszkiewicz, M., Lasek, P.: A neighborhood-based clustering by means of the triangle inequality and reference points. ICS Research Report 3, Institute of Computer Science, Warsaw University of Technology, Warsaw (2011)
9. Kryszkiewicz, M.: Efficient determination of neighborhoods defined in terms of cosine similarity measure. ICS Research Report 4, Institute of Computer Science, Warsaw University of Technology, Warsaw (2011)
10. Moore, A. W.: The anchors hierarchy: Using the triangle inequality to survive high dimensional data. In: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence (UAI '00), June 30 - July 3, Stanford, California, USA, pp. 397-405. Morgan Kaufmann, San Francisco, CA (2000)
11. Patra, B.K., Hubballi, N., Biswas, S., Nandi, S.: Distance based fast hierarchical clustering method for large datasets. In: Proceedings of 7th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2010), June 28-30, Warsaw, Poland, *Lecture Notes in Computer Science*, vol. 6086, pp. 50–59. Springer-Verlag, Berlin, Heildelberg, Germany (2010)
12. Salton, G., Wong, A., Yang, C. S.: A vector space model for automatic indexing. Communications of the ACM, **18**(11), 613–620 (1975)
13. Zhou, S., Zhao, Y., Guan, J., Huang, J.Z.: A neighborhood-based clustering algorithm. In: Advances in Knowledge Discovery and Data Mining, 9th Pacific-Asia Conference, May 18-20, Hanoi, Vietnam, *Lecture Notes in Computer Science*, vol. 3518, pp. 361–371. Springer-Verlag, Berlin, Heildelberg, Germany (2005)