Efficient Determination of Binary Non-Negative Vector Neighbors with Regard to Cosine Similarity

Marzena Kryszkiewicz

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

Abstract. The cosine and Tanimoto similarity measures are often and successfully applied in classification, clustering and ranking in chemistry, biology, information retrieval, and text mining. A basic operation in such tasks is identification of neighbors. This operation becomes critical for large high dimensional data. The usage of the triangle inequality property was recently offered to alleviate this problem in the case of applying a distance metric. The triangle inequality holds for the Tanimoto dissimilarity, which functionally determines the Tanimoto similarity, provided the underlying data have a form of vectors with binary non-negative values of attributes. Unfortunately, the triangle inequality holds neither for the cosine similarity measure nor for its corresponding dissimilarity measure. However, in this paper, we propose how to use the triangle inequality property and/or bounds on lengths of neighbor vectors to efficiently determine non-negative binary vectors that are similar with regard to the cosine similarity measure.

Keywords: nearest neighbors, ε-neighborhoods, the cosine similarity, the Tanimoto similarity, data clustering, text clustering

1 Introduction

The cosine and Tanimoto similarity measures are often and successfully applied in classification, clustering and ranking in chemistry, biology, information retrieval, and text mining. A basic operation in such tasks is identification of neighbors. This operation becomes critical for large high dimensional data. The usage of the triangle inequality property was recently offered to alleviate this problem in the case of applying a distance metric [2-6, 8-9]. The triangle inequality holds for the Tanimoto dissimilarity, which functionally determines the Tanimoto similarity, provided the underlying data have a form of vectors with binary non-negative values of attributes [7]. Unfortunately, the triangle inequality holds neither for the cosine similarity measure nor for its corresponding dissimilarity measure. However, in this paper, we propose how to use the triangle inequality property and/or bounds on lengths of neighbor vectors to efficiently determine binary non-negative vectors that are similar with regard to the cosine similarity measure. More specifically, in this paper, we will consider vectors such that a domain of each of their attributes (or dimensions) is binary and may take

adfa, p. 1, 2011.

either value 0 or a positive real value. When the domains of their all dimensions include only 0 and 1, 0 might denote absence of an attribute, while1 its presence. If a positive value different from 1 is allowed for an attribute, it might reflect the importance (weight) of the occurrence of the attribute. The larger positive attribute value, the more important attribute. In the paper, we will call such vectors *binary non-negative vectors*.

Our paper has the following layout. Section 2 provides basic notions and properties used in the paper. In Section 3, we recall the method as offered in [3-4], which applies the triangle inequality property to efficiently calculate neighborhoods using a distance metric also in the case of large high dimensional datasets. In Section 4, we investigate the relationship between the cosine similarity measure and Tanimoto (dis)similarity. In Section 5, we formulate and prove the bounds on the lengths of cosine similar binary non-negative vectors. In Section 6, we investigate the combined usage of the Tanimoto dissimilarity, the triangle inequality and the found bounds on the length of vectors for determining cosine similarity neighborhoods of binary non-negative vectors. Section 7 concludes our work.

2 Basic Notions and Properties

In the chapter, we will consider vectors of the 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 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) $dis(p, q) = 0$ iff $p = q$;
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 *triangle inequality property*. Often, an alternative form of this property is used; namely, $dis(p, q) \geq dis(p, r) - dis(q, r)$.

In order to compare vectors, one may alternatively use *similarity measures*. 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.

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

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

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

Property 1. Let u and v be non-zero vectors. Then, $cosSim(u, v) \in [-1, 1]$.

The *Tanimoto similarity* between vectors u and v is denoted by $T(u, v)$ and is defined as follows,

$$T(u, v) = \frac{u \cdot v}{u \cdot u + v \cdot v - u \cdot v}.$$

In the case of binary vectors with attribute domains restricted to {0, 1}, the Tanimoto similarity between two vectors determines the ratio of the number of attributes (“1s”) shared by both vectors to the number of attributes (“1s”) occurring in either vector.

Property 2 [10]. Let u and v be vectors. Then, $T(u, v) \in \left[-\frac{1}{3}, 1\right]$.

Both the cosine similarity and the Tanimoto similarity do not preserve the triangle inequality property. Also, $1 - cosSim(u, v)$ does not preserve this property. However, it was proved in [7] that the measure $1 - T(u, v)$, known as *Tanimoto dissimilarity*, preserves the triangle inequality property for each pair (u,v) of binary non-negative vectors.

Below, we provide definitions of neighborhoods in a vector set D with regard to a dissimilarity measure *dis* and, respectively, with regard to a similarity measure *sim*.

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

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

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

$$\varepsilon\text{-}SNB_{sim}^{D}(p) = \{q \in D \setminus \{p\} \mid sim(p, q) \ge \varepsilon\}.$$

Instead of looking for an ε-neighborhood (an ε-similarity neighborhood), one may be interested in determining *k-nearest neighbors* (*k-similarity nearest neighbors*). Let D' be a set containing k vectors from $D \setminus \{p\}$ and $\varepsilon = \max\{dis(p, q) | q \in D'\}$. Then, k-nearest neighbors are guaranteed to be found within ε distance from vector p; that is, are contained in $\varepsilon\text{-}NB_{dis}^{D}(p)$. In practice, the value of ε within which k-nearest neighbors of p are guaranteed to be found is re-estimated (is possibly narrowed) when calculating the distance between p and next vectors from $D \setminus \{p\}$ [5-6]. In an analogous way, *k*-similarity nearest neighbors could be determined. In the following, we will focus on determining ε-(similarity) neighborhoods.

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

In this section, we recall the method of determining ε-neighborhoods based on distance metrics efficiently, as proposed in [3-4].

Lemma 1 [3-4]. Let *dis* be a distance metric *D* be a set of vectors. For any vectors *p*, $q \in D$ and any vector *r*:

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

Now, let us consider vector *q* such that $dis(q, r) > dis(u, r)$. If $dis(u, r) - dis(v, r) > \varepsilon$, then also $dis(q, r) - dis(v, r) > \varepsilon$, and thus, $v \notin \varepsilon\text{-}NB_{dis}{}^{D}(q)$ and $q \notin \varepsilon\text{-}NB_{dis}{}^{D}(v)$ without calculating the real distance between *q* and *v*. This observation provides the intuition behind Theorem 1.

Theorem 1 [3-4]. Let *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 $u \in D$, *f* be a vector following vector *u* in *D* such that $dis(f, r) - dis(u, r) > \varepsilon$, and *p* be a vector preceding vector *u* in *D* such that $dis(u, r) - dis(p, r) > \varepsilon$. Then:

a) *f* and all vectors following *f* in *D* do not belong to $\varepsilon\text{-}NB_{dis}{}^{D}(u)$;
b) *p* and all vectors preceding *p* in *D* do not belong to $\varepsilon\text{-}NB_{dis}{}^{D}(u)$.

As follows from Theorem 1, it makes sense to order all vectors in a given dataset *D* with regard to a reference vector 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] showed that the determination of ε-neighborhoods by means of Theorem 1 was always faster than the determination using the R-Tree index, 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. Analogous results were reported when determining *k*-neighborhoods [5-6].

4 The Cosine Similarity versus the Tanimoto Distance

In this section, we investigate the relation between the cosine similarity and the Tanimoto similarity and dissimilarity measures.

Lemma 2 [1]. Let *u* and *v* be non-zero vectors. Then:

$$T(u, v) = \frac{cosSim(u,v)}{A - cosSim(u,v)}, \text{ where } A = \frac{|u|}{|v|} + \frac{|v|}{|u|}.$$

Proof. $T(u, v) = \dfrac{u \cdot v}{u \cdot u + v \cdot v - u \cdot v} = \dfrac{cosSim(u,v)\,|u||v|}{|u|^2 + |v|^2 - cosSim(u,v)\,|u||v|} = \dfrac{cosSim(u,v)}{\dfrac{|u|^2 + |v|^2}{|u||v|} - cosSim(u,v)} = \dfrac{cosSim(u,v)}{\dfrac{|u|}{|v|} + \dfrac{|v|}{|u|} - cosSim(u,v)} = \dfrac{cosSim(u,v)}{A - cosSim(u,v)}$. □

Proposition 1. Let u and v be non-zero vectors. Then:

$$cosSim(u, v) = \frac{T(u,v)A}{1 + T(u,v)}, \text{ where } A = \frac{|u|}{|v|} + \frac{|v|}{|u|}.$$

Proof. Follows from Lemma 2. □

Lemma 3. Let a and b be real numbers such that $ab > 0$. Then $\dfrac{a}{b} + \dfrac{b}{a} \geq 2$.

Proof. Follows from $(a - b)^2 \geq 0$. □

Theorem 2. Let u and v be non-zero vectors, $A = \dfrac{|u|}{|v|} + \dfrac{|v|}{|u|}$, $\varepsilon \in [-1,1]$ and $\varepsilon' = \dfrac{\varepsilon}{A - \varepsilon}$. Then $\varepsilon' \leq \varepsilon$ for $\varepsilon \in (0,1]$, $\varepsilon' \geq \varepsilon$ for $\varepsilon \in [-1,0)$, $\varepsilon' = \varepsilon$ for $\varepsilon = 0$ and

$$cosSim(u, v) \geq \varepsilon \Leftrightarrow T(u, v) \geq \varepsilon' \Leftrightarrow 1 - T(u, v) \leq 1 - \varepsilon'.$$

Proof. By Lemma 3, $A \geq 2$. Hence, $\varepsilon' \leq \varepsilon$ for $\varepsilon \in (0,1]$, $\varepsilon' \geq \varepsilon$ for $\varepsilon \in [-1,0)$ and $\varepsilon' = \varepsilon$ for $\varepsilon = 0$. By Proposition 1, $cosSim(u, v) \geq \varepsilon \Leftrightarrow \dfrac{T(u,v)A}{1 + T(u,v)} \geq \varepsilon$ and by Property 2, $1 + T(u, v) > 0$. Hence, $cosSim(u, v) \geq \varepsilon \Leftrightarrow T(u, v) \geq \dfrac{\varepsilon}{A - \varepsilon} = \varepsilon'$. □

The corollary beneath follows immediately from Theorem 2.

Corollary 1. Let $D \cup \{u\}$ be a set of non-zero vectors, $\varepsilon \in [-1,1]$ and $\varepsilon'(v, w) = \dfrac{\varepsilon}{\dfrac{|v|}{|w|} + \dfrac{|w|}{|v|} - \varepsilon}$ for any vectors v, w in $D \cup \{u\}$. Then:

$$\varepsilon\text{-}SNB_{cosSim}{}^{D}(u) = \{v \in D \setminus \{u\} \mid T(u,v) \geq \varepsilon'(u,v)\} =$$

$$\{v \in D \setminus \{u\} \mid 1 - T(u,v) \leq 1 - \varepsilon'(u,v)\}.$$

Corollary 2. Let $D \cup \{u\}$ be a set of binary non-negative vectors and $\varepsilon \in [-1,1]$. Then, $\varepsilon\text{-}SNB_{cosSim}{}^{D}(u)$ can be determined by means of the Tanimoto dissimilarity and supported by the usage of the triangle inequality as specified in Lemma 1.

5 Bounds on Lengths of Binary Non-Negative Cosine Similar Vectors

In this section, we consider the possibility of reducing the search space of candidate neighbor vectors by taking into account the lengths of vectors in a given dataset *D*.

Theorem 3. Let *u* and *v* be binary non-negative non-zero vectors such that *cosSim*(*u*, *v*) $\geq \varepsilon$ and $\varepsilon \in (0, 1]$. Then:

a) $|v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right]$;

b) $|v|^2 \in \left[\varepsilon^2|u|^2, \frac{|u|^2}{\varepsilon^2}\right]$.

Proof. Ad a) Since *u* and *v* are binary non-negative vectors, then for any dimension *i*: $u_i v_i$ equals either $u_i u_i$ or $u_i 0$. Hence, $u_i v_i \leq u_i u_i$. Analogously, $u_i v_i \leq v_i v_i$.

Now, $\frac{|u|^2}{|u||v|} = \frac{\sum_{i=1..n} u_i u_i}{|u||v|} \geq \frac{\sum_{i=1..n} u_i v_i}{|u||v|} = \frac{u \cdot v}{|u||v|} = cosSim(u, v) \geq \varepsilon$.

Hence, $\frac{|u|^2}{|u||v|} \geq \varepsilon$. Thus, $|v| \leq \frac{|u|}{\varepsilon}$.

In addition, $\frac{|v|^2}{|u||v|} = \frac{\sum_{i=1..n} v_i v_i}{|u||v|} \geq \frac{\sum_{i=1..n} u_i v_i}{|u||v|} = \frac{u \cdot v}{|u||v|} = cosSim(u, v) \geq \varepsilon$.

Hence, $\frac{|v|^2}{|u||v|} \geq \varepsilon$. Thus, $|v| \geq \varepsilon|u|$.

Ad b) Follows immediately from Theorem 3a. □

Let *u* be a vector for which we wish to find its ε-cosine similarity neighborhood, where $\varepsilon \in (0, 1]$. Theorem 3 tells us that among vectors shorter than *u* only those not shorter than $\varepsilon|u|$ may belong to the ε-cosine similarity neighborhood of *u*, while among vectors longer than *u* only those not longer than $\frac{|u|}{\varepsilon}$ may belong to this neighborhood.

Corollary 3. Let $D \cup \{u\}$ be a set of binary non-negative non-zero vectors, $\varepsilon \in (0,1]$. Then:

$$\varepsilon\text{-}SNB_{cosSim}{}^{D}(u) = \left\{ v \in D \setminus \{u\} \,\middle|\, |v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right] \wedge cosSim(u, v) \geq \varepsilon \right\}.$$

6 The Tanimoto Distance and Lengths of Vectors as a Means to Restrict a Set of Candidates for Members of Cosine Similarity Neighborhoods of Binary Non-Negative Vectors

In this section, we investigate the usage of both the Tanimoto dissimilarity, the triangle inequality and the found bounds on the length of vectors for determining cosine similarity neighborhoods of binary non-negative vectors.

Lemma 4. Let $a,c,d > 0$, $c \le d$, $b \in [c, d]$. Let $A = \max\left\{\frac{a}{c}+\frac{c}{a}, \frac{a}{d}+\frac{d}{a}\right\}$ Then:

a) $\frac{a}{b}+\frac{b}{a} \le A$;

b) If $\varepsilon \in [0,1]$, then $1-\frac{\varepsilon}{\frac{a}{b}+\frac{b}{a}-\varepsilon} \le 1-\frac{\varepsilon}{A-\varepsilon}$.

Proof. Ad a) Let $f(b)$: $[c, d] \rightarrow \mathrm{R}^+$ be a function such that $f(b) = \frac{a}{b}+\frac{b}{a}$. Then, $f'(b) = \frac{-a}{b^2}+\frac{1}{a}$.

Case $a \in [c, d]$. $f'(b) = 0$ iff $b = a$; $f'(b) < 0$ iff $b \in [c, a)$; $f'(b) > 0$ iff $b \in (a, d]$. Hence, $f(b)$ has minimal value for $b = a$; $f(b)$ is non-increasing in $[c, a)$ and $f(b)$ is non-decreasing in $(a, d]$. Thus, the greatest value of $f(b) = \max\left\{\frac{a}{c}+\frac{c}{a}, \frac{a}{d}+\frac{d}{a}\right\}$.

Case $a < c$. Then, $f'(b) > 0$ iff $b \in [c, d]$. Thus, $f(b)$ is non-decreasing in $[c, d]$. So, the greatest value of $f(b) = f(d) = \frac{a}{d}+\frac{d}{a}$.

Case $a > d$. Then, $f'(b) < 0$ iff $b \in [c, d]$. Thus, $f(b)$ is non-increasing in $[c, d]$. So, the greatest value of $f(b) = f(c) = \frac{a}{c}+\frac{c}{a}$.

So, we have proved that in all three cases $\frac{a}{b}+\frac{b}{a} \le \max\left\{\frac{a}{c}+\frac{c}{a}, \frac{a}{d}+\frac{d}{a}\right\} = A$.

Ad b) Follows immediately, from Lemma 4a. □

Lemma 5. Let u and v be non-zero vectors, $\varepsilon \in (0,1]$ and $|v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right]$. Then

$$1-\frac{\varepsilon}{\frac{|u|}{|v|}+\frac{|v|}{|u|}-\varepsilon} \le 1-\varepsilon^2.$$

Proof. Let $a = |u|$, $b = |v|$, $c = \varepsilon|u|$, $d = \frac{|u|}{\varepsilon}$ and $|v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right]$. Then, $b \in [c, d]$. Let $A = \max\left\{\frac{a}{c} + \frac{c}{a}, \frac{a}{d} + \frac{d}{a}\right\}$. Since, $\frac{a}{c} + \frac{c}{a}$ = $\frac{1}{\varepsilon} + \varepsilon$ and $\frac{a}{d} + \frac{d}{a}$ = $\frac{1}{\varepsilon} + \varepsilon$, then A = $\frac{1}{\varepsilon} + \varepsilon$. Hence, and by Lemma 4b, $1 - \frac{\varepsilon}{\frac{|u|}{|v|} + \frac{|v|}{|u|} - \varepsilon}$ = $1 - \frac{\varepsilon}{\frac{a}{b} + \frac{b}{a} - \varepsilon}$ $\leq 1 - \frac{\varepsilon}{A - \varepsilon}$ = $1 - \varepsilon^2$. □

Theorem 4. Let $D \cup \{u\}$ be a set of binary non-negative non-zero vectors, $\varepsilon \in (0,1]$ and $\varepsilon'(v, w) = \frac{\varepsilon}{\frac{|v|}{|w|} + \frac{|w|}{|v|} - \varepsilon}$ for any vectors v, w in $D \cup \{u\}$. Then:

$$\varepsilon\text{-}SNB_{cosSim}{}^{D}(u) = \left\{v \in D \setminus \{u\} \,\middle|\, |v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right] \wedge 1 - T(u,v) \leq 1 - \varepsilon'(u,v)\right\} \subseteq$$

$$\left\{v \in D \setminus \{u\} \,\middle|\, |v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right] \wedge 1 - T(u,v) \leq 1 - \varepsilon^2\right\} \subseteq (1-\varepsilon^2)\text{-}NB_{1-T}{}^{D}(u).$$

Proof. $\varepsilon\text{-}SNB_{cosSim}{}^{D}(u)$ = (by Corollary 3 and Theorem 2)

$$\left\{v \in D \setminus \{u\} \,\middle|\, |v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right] \wedge 1 - T(u,v) \leq 1 - \varepsilon'(u,v)\right\} \subseteq \text{(by Lemma 5)}$$

$$\left\{v \in D \setminus \{u\} \,\middle|\, |v| \in \left[\varepsilon|u|, \frac{|u|}{\varepsilon}\right] \wedge 1 - T(u,v) \leq 1 - \varepsilon^2\right\} \subseteq$$

$$\left\{v \in D \setminus \{u\} \,\middle|\, 1 - T(u,v) \leq 1 - \varepsilon^2\right\} = (1-\varepsilon^2)\text{-}NB_{1-T}{}^{D}(u). \; □$$

Corollary 4. Let $D \cup \{u\}$ be a set of binary non-negative non-zero vectors and $\varepsilon \in (0,1]$. Then, $\varepsilon\text{-}SNB_{cosSim}{}^{D}(u)$ can be determined by means of the Tanimoto distance and supported by the usage of the triangle inequality as specified in Lemma 1 and Theorem 1.

Theorem 4 suggests a number of ways in which $\varepsilon\text{-}SNB_{cosSim}{}^{D}(u)$, where $\varepsilon \in (0, 1]$, can be determined. First of all, $\varepsilon\text{-}SNB_{cosSim}{}^{D}(u)$ is a subset of $\varepsilon'\text{-}NB_{1-T}{}^{D}(u)$, $\varepsilon' = 1 - \varepsilon^2$, which can be determined by means of the triangle inequality (e.g. as proposed in [3-4] and recalled in Section 3), provided dataset D contains only binary non-negative non-zero vectors. In addition, only those vectors v in $\varepsilon'\text{-}NB_{1-T}{}^{D}(u)$ the length of which belongs to the interval $\left[\varepsilon\,|u|, \frac{|u|}{\varepsilon}\right]$ and that fulfill the condition $1-T(u,v) \le 1-\varepsilon'(u,v)$, where $\varepsilon'(u, v)$ = $\frac{\varepsilon}{\frac{|u|}{|v|}+\frac{|v|}{|u|}-\varepsilon}$, have a chance to belong to ε-$SNB_{cosSim}{}^{D}(u)$. In the sequel, the former condition is denoted by LC, the latter by TC and their conjunction by LTC. In Table 1, we present the results of evaluating the selectiveness of these conditions which we carried out on the dataset with 11 binary attributes and 2047 different vectors.

Table 1. Average numbers of evaluated vectors for a vector *u*, where LC(*u*) – the percentage of vectors in the dataset fulfilling the condition $|v| \in \left[\varepsilon\,|u|, \frac{|u|}{\varepsilon}\right]$, *r* = [10101010101] – used reference vector, TC(*u*) - the percentage of vectors *v* that were not eliminated by applying the triangle inequality condition (1–*T*(*u*, *r*) – (1–*T*(*u*, *v*)) $\le 1-\frac{\varepsilon}{\frac{|u|}{|v|}+\frac{|v|}{|u|}-\varepsilon}$), LTC(*u*) – the percentage of vectors fulfilling both TL(*u*) and TC(*u*)

ε	$\frac{\lvert \varepsilon\text{-}SNB_{cosSim}{}^{D}(u)\rvert}{\lvert D\rvert}$	LC(*u*)	TC(*u*)	LTC(*u*)
0.9855	0,05%	16,44%	16,35%	4,85%
0.9200	0,25%	27,24%	53,21%	12,87%

As follows from these results, average selectiveness of LC is similar to that of TC for very high value of ε (0.9855) and greater than TC for lower value of ε (0.9200). In both cases, it is most beneficial to apply both conditions (LTC).

7 Conclusions

In the paper, we have proposed a new solution to determining neighborhoods defined in terms of the cosine similarity measure for binary non-negative vectors. We have proposed and proved that this problem can be transformed to the problem of determining neighborhoods defined in terms of the Tanimoto dissimilarity. This equivalence allows us to apply solutions based on using the triangle inequality that were proposed recently in the literature. In addition, we showed that in the case of binary non-negative vectors, one may restrict the cosine similarity neighbor search area by applying our proposed bounds on the lengths of candidate vectors.

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". I would like to thank an anonymous reviewer for suggestions on how to simplify statements and proofs.

References

1. Leo, E.: New relations between similarity measures for vectors based on vector norms, ASIS&T Journal, 60 (2), pp. 232-239 (2009)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. Lipkus, A.H.: A proof of the triangle inequality for the Tanimoto dissimilarity. Journal of Mathematical Chemistry, 26 (1-3), pp. 263-265 (1999)
8. 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)
9. 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)
10. Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. J. Chem. Inf. Comput. Sci., 38 (6), pp. 983–996 (1998)