The Triangle Inequality versus Projection onto a Dimension in Determining Cosine Similarity Neighborhoods of Non-Negative Vectors*

Marzena Kryszkiewicz

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

Abstract. In many applications, objects are represented by non-negative vectors and cosine similarity is used to measure their similarity. It was shown recently that the determination of the cosine similarity of two vectors can be transformed to the problem of determining the Euclidean distance of normalized forms of these vectors. This equivalence allows applying the triangle inequality to determine cosine similarity neighborhoods efficiently. Alternatively, one may apply the projection onto a dimension to this end. In this paper, we prove that the triangle inequality is guaranteed to be a pruning tool, which is not less efficient than the projection in determining neighborhoods of non-negative vectors.

1 Introduction

In many applications, especially in text mining, biomedical engineering and chemistry, cosine similarity is often used to find objects (nearest neighbors) most similar to a given one. Objects themselves are frequently represented by non-negative vectors. The determination of nearest neighbors is challenging if analyzed vectors are high dimensional. In the case of distance metrics, one may apply the triangle inequality to quickly prune large numbers of objects that certainly are not nearest neighbors of a given vector [1, 3–6]. While the cosine similarity does not preserve the triangle inequality, it was shown recently that the problem of determining the cosine similarity of two vectors can be transformed to the problem of determining the Euclidean distance of normalized forms of these vectors [2]. This equivalence allows applying the triangle inequality to determine cosine similarity neighborhoods efficiently. Alternatively, one may apply the projection onto a dimension to this end. In this paper, we prove that the triangle inequality is guaranteed to be a pruning tool, which is not less efficient than the projection in determining neighborhoods of non-negative vectors.

Our paper has the following layout. In Section 2, we recall basic notions and the equation relating the Euclidean distance and the cosine similarity [2]. In

* 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’.

Section 3, we recall how the problem of determining a cosine similarity neighborhood can be transformed to the problem of determining a neighborhood w.r.t the Euclidean distance [2]. The usage of the triangle inequality for efficient pruning of non-nearest neighbors is recalled in Section 4 [3, 4]. In Section 5, we present an analogous approach to pruning such vectors by using the projection of vectors onto a dimension. Section 6 contains the main contribution of this paper, which consists in showing that for any dimension one may apply pruning of non-nearest neighbors in a set of non-negative normalized vectors by means of the triangle inequality, which is not less efficient than the projection onto this dimension. Section 7 summarizes our work.

2 The Euclidean Distance and the Cosine Similarity

In the paper, we consider vectors of the same dimensionality, say n. A vector u will be also denoted as $[u_1, \ldots, u_n]$, where u_i is the value of the i-th dimension of u, $i = 1..n$. A vector is called *non-negative* if all its dimensions are not negative.

The *Euclidean distance* between vectors u and v is denoted by $Euclidean(u, v)$ and is defined as $\sqrt{\sum_{i=1..n}(u_i - v_i)^2}$. The Euclidean distance preserves *the triangle inequality*; that is, for any vectors u, v, and r, $Euclidean(u, r) \leq Euclidean(u, v) + Euclidean(v, r)$ (or, alternatively $Euclidean(u, v) \geq Euclidean(u, r) - Euclidean(v, r)$).

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

In fact, the cosine similarity and Euclidean distance are related by an equation:

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

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

Clearly, the cosine similarity between any vectors u and v depends solely on the angle between the vectors and does not depend on their lengths, hence the calculation of the $cosSim(u, v)$ may be carried out on their *normalized forms*:

A *normalized form of a vector u* is denoted by $NF(u)$ and is defined as the ratio of u to its length $| u |$. A vector u is defined as a *normalized vector* if $u = NF(u)$. Obviously, the length of a normalized vector equals 1.

Theorem 1 [2]. Let u, v be non-zero vectors. Then:

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

Theorem 1 allows deducing that checking whether the cosine similarity between any two vectors exceeds a threshold ε, where $\varepsilon \in [-1, 1]$, can be carried out as checking if the Euclidean distance between the normalized forms of the vectors is less than the associated threshold $\varepsilon' = \sqrt{2 - 2\varepsilon}$:

Corollary 1 [2]. Let u, v be vectors, $\varepsilon \in [-1, 1]$ and $\varepsilon' = \sqrt{2 - 2\varepsilon}$. Then:

$$cosSim(u, v) \geq \varepsilon \text{ iff } Euclidean(NF(v),\ NF(u)) \leq \varepsilon'.$$

3 Euclidean Distance Neighbourhood and Cosine Similarity Neighbourhood

ε-Euclidean neighborhood of a vector p in D is denoted by $\varepsilon\text{-}NB^{D}_{Euclidean}(p)$ and is defined as the set of all vectors in dataset $D\backslash\{p\}$ that are distant in the Euclidean sense from p by no more than ε. *ε-cosine similarity neighborhood of a vector p in D* is denoted by $\varepsilon\text{-}SNB^{D}_{cosSim}(p)$ and is defined as the set of all vectors in dataset $D\backslash\{p\}$ that are cosine similar to p by no less than ε.

Corollary 1 allows transforming the problem of determining a cosine similarity neighborhood of a given vector u within a set of vectors D to the problem of determining an Euclidean neighborhood of $NF(u)$ within the vector set D' consisting of the normalized forms of the vectors from D.

Theorem 2 [2]. Let D be a set of m vectors $\{p_{(1)}, \ldots, p_{(m)}\}$, D' be the 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, $\varepsilon\text{-}SNB^{D}_{cosSim}(p_{(i)}) = \{p_{(j)} \in D | u_{(j)} \in \varepsilon'\text{-}NB^{D'}_{Euclidean}(u_{(i)})\}$.

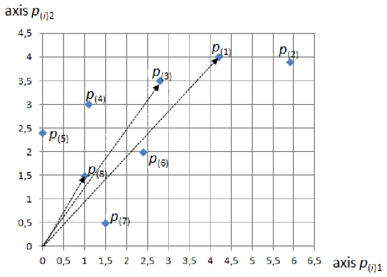

Fig. 1. Sample set D of vectors

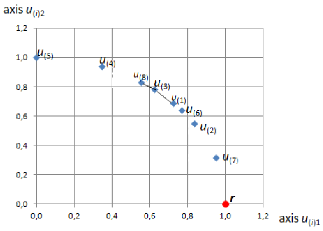

Fig. 2. Set D' containing normalized forms of vectors from D

Example 1. Let us consider the determination of ε-cosine similarity neighborhood of any vector $p_{(i)}$ in dataset $D = \{p_{(1)}, \ldots, p_{(8)}\}$ from Figure 1 for $\varepsilon = 0.9856$ (which roughly corresponds to the angle of 9.74^o). This task can be transformed to the task of determining ε'-Euclidean neighborhood of $u_{(i)} = NF(p_{(i)})$ in the set $D' = \{u_{(1)}, \ldots, u_{(8)}\}$ containing normalized forms of the vectors from D, provided $\varepsilon' = \sqrt{2 - 2\varepsilon} \approx 0.17$. Set D' is presented in Figure 2. □

4 The Triangle Inequality In Determining Euclidean Distance Neighborhoods

We will recall now the method of determining ε-Euclidean neighborhoods as proposed in [3]. We start with Lemma 2, which follows from the triangle inequality.

Lemma 2 [3]. Let D be a set of vectors. For any vectors u, $v \in D$ and any vector r: $Euclidean(u,r) - Euclidean(v,r) > \varepsilon \Rightarrow Euclidean(u,v) > \varepsilon \Rightarrow v \notin \varepsilon$-$NB^D_{Euclidean}(u) \wedge u \notin$ ε-$NB^D_{Euclidean}(v)$.

Let us consider vector q such that $Euclidean(q,r) > Euclidean(u,r)$. If $Euclidean(u,r) - Euclidean(v,r) > \varepsilon$, then $Euclidean(q,r) -$ $Euclidean(v, r) > \varepsilon$, and thus one may conclude that $v \notin \varepsilon$-$NB^D_{Euclidean}(q)$ and $q \notin \varepsilon$-$NB^D_{Euclidean}(v)$ without calculating the real distance between q and v. This observation provides the intuition behind Theorem 3.

Theorem 3 [3]. Let r be any vector and D be a set of vectors ordered in a non-decreasing way w.r.t their distances to r. Let $u \in D, f$ be a vector following vector u in D such that $Euclidean(f,r) - Euclidean(u,r) > \varepsilon$, and p be a vector preceding vector u in D such that $Euclidean(u,r) - Euclidean(p,r) > \varepsilon$. Then:
a) f and all vectors following f in D do not belong to ε-$NB^D_{Euclidean}(u)$;
b) p and all vectors preceding p in D do not belong to ε-$NB^D_{Euclidean}(u)$.

The experiments reported in [3] showed that the determination of ε-Euclidean neighborhoods by means of Theorem 3 was always faster than their determination by means of 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.

5 Vector Projection onto a Dimension in Determining Euclidean Distance Neighborhoods

It is easy to observe that for any dimension $l, l \in [1, \ldots, n]$, and any two vectors u and v, the following holds: $\mid u_l - v_l \mid = \sqrt{(u_l - v_l)^2} \leq \sqrt{\sum_{i=1..n}(u_l - v_l)^2} = Euclidean(u,v)$. Hence, if $\mid u_l - v_l \mid > \varepsilon$, then $Euclidean(u,v) > \varepsilon$; that is, $u \notin \varepsilon$-$NB^D_{Euclidean}(v) \wedge v \notin \varepsilon$-$NB^D_{Euclidean}(u)$. This observation implies Proposition 1.

Proposition 1. Let l be an index of a dimension l, where $l \in [1, \ldots, n]$, and D be a set of vectors ordered in a non-decreasing way w.r.t the values of their l-th dimension. Let $u \in D, f$ be a vector following vector u in D such that $f_l - u_l > \varepsilon$, and p be a vector preceding vector u in D such that $u_l - p_l > \varepsilon$. Then:
a) f and all vectors following f in D do not belong to ε-$NB^D_{Euclidean}(u)$;
b) p and all vectors preceding p in D do not belong to ε-$NB^D_{Euclidean}(u)$.

6 The Triangle Inequality versus Projection onto a Dimension for Non-Negative Normalized Vectors

In this section, we will denote $| \; u_l - v_l |$ by $\Delta_{dim\_l}(u, v)$ and $| \; Euclidean(u, r) - Euclidean(v, r) \; |$ by $\Delta_{ref\_r}(u, v)$. $\Delta_{dim\_l}(u, v)$ can be perceived as a pessimistic estimation of the Euclidean distance between u and v obtained by applying the projection onto l-th dimension, whereas $\Delta_{ref\_r}(u, v)$ can be perceived as a pessimistic estimation of $Euclidean(u, v)$ by means of the triangle inequality applied to reference vector r. In the following, we will focus on reference vectors of a special type. A vector will be called an $l(a)$-*vector* if its l-th coordinate equals a and all remaining coordinates equal 0. If r is an $l(a)$-vector, then $\Delta_{ref\_r}(u, v)$ will be also denoted as $\Delta_{ref\_l(a)}(u, v)$.

Proposition 2. Let $a \in R$, r be an $l(a)$-vector, u and v be non-negative normalized vectors. Then:
a) $Euclidean(u, r) = \sqrt{1 + a^2 - 2au_l}$.
b) If $u_l = v_l$, then $\Delta_{ref\_l(a)}(u, v) = \Delta_{dim\_l}(u, v) = 0$.
c) $Euclidean(u, r) = 0 \Leftrightarrow \sqrt{1 + a^2 - 2au_l} = 0 \Leftrightarrow a = u_l = 1$.

Proof. Ad a) $Euclidean(u, r) = \sqrt{\sum_{i=1..n, i \neq l} (u_i - 0)^2 + (u_l - a)^2} = \sqrt{\sum_{i=1..n} u_i^2 + a^2 - 2au_l} = \sqrt{1 + a^2 - 2au_l}$.
Ad b, c) Follow from Proposition 2a. □

Lemma 3. Let $a = 1$, r be an $l(a)$-vector, u and v be non-negative normalized vectors such that $0 \leq u_l \leq v_l = 1$. Then, $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v)$.

Proof. $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v) \Leftrightarrow \sqrt{1 + a^2 - 2au_l} - \sqrt{1 + a^2 - 2av_l} \geq v_l - u_l \Leftrightarrow \sqrt{2 - 2u_l} \geq 1 - u_l \Leftrightarrow \sqrt{2(1 - u_l)} \geq 1 - u_l$, which is fulfilled as $1 - u_l \in [0, 1]$ and in such a case $\sqrt{1 - u_l} \geq 1 - u_l$. □

Lemma 4. Let $a > 0$, r be an $l(a)$-vector, u and v be non-negative normalized vectors such that $0 < u_l \leq v_l$ and either $v_l \neq 1$ or $a \neq 1$. Then, $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v)$ for $a \geq 1/(2u_l)$.

Proof. $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v) \Leftrightarrow \sqrt{1 + a^2 - 2au_l} - \sqrt{1 + a^2 - 2av_l} \geq v_l - u_l \Leftrightarrow \sqrt{1 + a^2 - 2au_l} + u_l \geq \sqrt{1 + a^2 - 2av_l} + v_l$. Let $x \in (0, 1]$ and $f(x) = \sqrt{1 + a^2 - 2ax} + x$. Then, $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v) \Leftrightarrow f(u_l) \geq f(v_l)$. We will prove now that $f(u_l) \geq f(v_l)$ if $a > 1/(2u_l)$, by investigating the monotonicity of $f(x)$. $f'(x) = \frac{-a}{\sqrt{1+a^2-2ax}} + 1$ (by Proposition 2c, $\sqrt{1 + a^2 - 2ax} \neq 0$). Hence, $f'(x) \leq 0$ (and so, $f(x)$ is non-increasing) for $x \geq 1/(2a)$. Thus, $f(u_l) \geq f(v_l)$ if $\min\{u_l, v_l\} \geq 1/(2a)$. Therefore, $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v)$ if $a \geq 1/(2u_l)$. □

Lemma 5. Let $a > 0$, r be an $l(a)$-vector, u, v be non-negative normalized vectors and $0 < u_l \leq v_l$. Then, $\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v)$ for $a \geq 1/(2u_l)$.

Proof. Follows from Lemma 3 and Lemma 4. □

Theorem 4. Let $a > 0$, r be an $l(a)$-vector, D be a set of non-negative normalized vectors none of which has dimension l equal to 0. Then for any $u, v \in D$:

$$\Delta_{ref\_l(a)}(u, v) \geq \Delta_{dim\_l}(u, v) \text{ for } a \geq \frac{1}{2\mu}, \text{ where } \mu = \min\{u_l | u \in D\}.$$

Proof. Follows from Lemma 5 and the fact that $\Delta_{ref\_l(a)}(v,u) = \Delta_{ref\_l(a)}(u,v)$ and $\Delta_{dim\_l}(v,u) = \Delta_{dim\_l}(u,v)$. □

Theorem 4 tells us that for $a \geq 1/(2\mu)$, where $\mu = \min\{u_l | u \in D\}$, the pessimistic estimation of the Euclidean distance between two vectors by means of the triangle inequality applied to $l(a)$-reference vector is not less accurate than the pessimistic estimation of their Euclidean distance by means of the projection onto dimension l, provided the distance is calculated only among non-negative normalized vectors none of which has l-th dimension equal to 0.

Lemma 6. Let $a > 0$, r be an $l(a)$-vector, u, v be non-negative normalized vectors and $0 = u_l < v_l$. Then, $\Delta_{ref\_l(a)}(u,v) \geq \Delta_{dim\_l}(u,v)$ for $a \geq 1/v_l - v_l/4$.
Proof. $\Delta_{ref\_l(a)}(u,v) \geq \Delta_{dim\_l}(u,v) \Leftrightarrow \sqrt{1+a^2-2au_l} - \sqrt{1+a^2-2av_l} \geq v_l - u_l \Leftrightarrow \sqrt{1+a^2} - v_l \geq \sqrt{1+a^2-2av_l} \Leftrightarrow 1+a^2+v_l^2-2v_l\sqrt{1+a^2} \geq 1+a^2-2av_l \Leftrightarrow \sqrt{1+a^2} \leq v_l/2+a \Leftrightarrow 1+a^2 \leq v_l^2/4+a^2+av_l \Leftrightarrow a \geq 1/v_l - v_l/4$. □

Lemma 7. Let $\mu \in [0,1]$. Then $1/\mu - \mu/4 \geq 1/(2\mu)$.
Proof. $4 \geq 2\mu^2$. Hence, $1/(2\mu) \geq \mu/4$. Thus, $1/\mu - \mu/4 \geq 1/(2\mu)$. □

Theorem 5. Let $a > 0$, r be an $l(a)$-vector, D be a set of non-negative normalized vectors. Then for any vectors u, v in D:

$$\Delta_{ref\_l(a)}(u,v) \geq \Delta_{dim\_l}(u,v) \text{ for } a \geq \frac{1}{\mu} - \frac{\mu}{4}, \text{ where } \mu = \min\{u_l | u \in D \wedge u_l \neq 0\}.$$

Proof. Follows from Theorem 4, Lemma 6, Lemma 7, Proposition 2b and the fact that $\Delta_{ref\_l(a)}(v,u) = \Delta_{ref\_l(a)}(u,v)$ and $\Delta_{dim\_l}(v,u) = \Delta_{dim\_l}(u,v)$. □

Theorem 5 tells us that for $a \geq 1/\mu - \mu/4$, where $\mu = \min\{u_l \mid u \in D \wedge u_l \neq 0\}$, the pessimistic estimation of the Euclidean distance between two vectors by means of the triangle inequality applied to $l(a)$-reference vector is not less accurate than the pessimistic estimation of this distance by means of the projection onto dimension l, provided the distance is calculated only among non-negative normalized vectors.

Example 2. Let us consider the set D of vectors from Example 1 and the set $D' = \{u_{(1)}, \ldots, u_{(8)}\}$ of their normalized forms. In Table 1, we present the values of the vectors in D'. Let us compare the results of using the projection of vectors onto dimension 1 and of using reference vector $r_0 = [1, 0]$ (i.e., 1(1.00)-reference vector). We can see that $\Delta_{ref\_r0}(u_{(i)}, u_{(3)})$ is greater than $\Delta_{dim\_1}(u_{(i)}, u_{(3)})$ in the case of vectors $u_{(i)} = u_{(7)}$, $u_{(2)}$, $u_{(6)}$, $u_{(1)}$, identical for vectors $u_{(i)} = u_{(3)}$, $u_{(8)}$, $u_{(4)}$, and less for vector $u_{(i)} = u_{(5)}$ (please see Table 1).

Let $\mu = \min\{u_{(i)1} | u_{(i)} \in D \wedge u_{(i)1} \neq 0\} = 0.34$. Then $1/(2\mu) \approx 1.470588 < 1.48$. Let r_1 be 1(1.48)-reference vector. By Theorem 4, $\Delta_{ref\_r1}(u_{(i)}, u_{(3)}) \geq \Delta_{dim\_1}(u_{(i)}, u_{(3)})$ for all vectors in D' that have non-zero dimension 1.

Now, $1/\mu - \mu/4 \approx 2.856176 < 2.86$. Let r_2 be 1(2.86)-reference vector. By Theorem 5, $\Delta_{ref\_r2}(u_{(i)}, u_{(3)}) \geq \Delta_{dim\_1}(u_{(i)}, u_{(3)})$ for all vectors $u_{(i)}$ in D'. □

Table 1. Normalized vectors in set $D' = \{u_{(1)}, \ldots, u_{(8)}\}$ and pessimistic estimations of distances between $u_{(3)}$ and vectors in D' by means of the projection onto dimension 1 and the triangle inequality w.r.t reference vectors: $r_0 = [1,\ 0]$, $r_1 = [1.48,\ 0]$, $r_2 = [2.86,\ 0]$

Vector			*Euclidean*	Δ_{ref_r0}	Δ_{ref_r1}	Δ_{ref_r2}	Δ_{dim_1}
$u_{(i)}$	$u_{(i)1}$	$u_{(i)2}$	$(u_{(i)}, r_0)$	$(u_{(i)}, u_{(3)})$	$(u_{(i)}, u_{(3)})$	$(u_{(i)}, u_{(3)})$	$(u_{(i)}, u_{(3)})$
$u_{(7)}$	0.95	0.32	0.32	**0.55**	**0.54**	**0.44**	0.33
$u_{(2)}$	0.83	0.55	0.58	**0.29**	**0.31**	**0.27**	0.21
$u_{(6)}$	0.77	0.64	0.68	**0.19**	**0.21**	**0.19**	0.15
$u_{(1)}$	0.72	0.69	0.74	**0.13**	**0.13**	**0.12**	0.01
$\boldsymbol{u_{(3)}}$	**0.62**	**0.78**	0.87	**0.00**	**0.00**	**0.00**	0.00
$u_{(8)}$	0.55	0.83	0.94	**0.07**	**0.09**	**0.08**	0.07
$u_{(4)}$	0.34	0.94	1.15	0.28	**0.32**	**0.32**	0.28
$u_{(5)}$	0.00	1.00	1.41	0.54	0.63	**0.66**	0.62

7 Summary

The problem of determining a cosine similarity neighborhood of a given vector u within a set of vectors D can be transformed to an equivalent problem of determining an Euclidean neighborhood of the normalized form of u within the vector set D' consisting of vectors of length 1 being the normalized forms of the vectors from D [2]. The triangle inequality applied to an arbitrary (reference) vector can be used for pruning when looking for an Euclidean neighborhood [3, 4]. The projection onto an arbitrary dimension can be also used to this end. We proved that for any dimension l and for any set of non-negative normalized vectors, one may always determine a reference vector that guarantees not worse pruning efficiency when looking for an Euclidean neighborhood by means of the triangle inequality than the efficiency achievable by using the projection onto l.

References

1. Elkan, C: Using the Triangle Inequality to Accelerate k-Means. In. Proc. of ICML03, Washington (2003) 147–153
2. 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)
3. Kryszkiewicz M., Lasek P.: TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality. RSCTC 2010 (2010) 60–69
4. Kryszkiewicz M., Lasek P.: A Neighborhood-Based Clustering by Means of the Triangle Inequality. IDEAL 2010 (2010) 284–291
5. Moore, A. W., The Anchors Hierarchy: Using the Triangle Inequality to Survive High Dimensional Data. In: Proc. of UAI, Stanford (2000) 397-405
6. Patra B.K., Hubballi N., Biswas S., Nandi S.: Distance Based Fast Hierarchical Clustering Method for Large Datasets. RSCTC 2010 (2010) 50–59