

**“A STUDY ON SPATIAL OBJECTS ASSOCIATED METRICS AND THEIR
COMPARISON”**

T.V.RAJINI KANTH
ASSISTANT PROFESSOR SELECTION GRADE
DEPT. OF MATHEMATICS
V.N.R.V.J.IE.T
HYDERABAD-72.
E-MAIL: rkanthtv@vnrvjiet.ac.in

&

Dr. C . RAGHAVENDRA .RAO
READER
DEPT. OF MATHEMATICS & STATISTICS
SCHOOL OF M.C.I.S.
UNIVERSITY OF HYDERABAD
HYDERABAD.
E-MAIL: crism@uohyd.ernet.in

ABSTRACT

A Spatial object will be represented by a Boolean vector of size ‘n’ indicating the presence of corresponding spatial primitives and is further used to demonstrate spatial metrics. Spatial characteristics can be measured for characterizing spatial objects. Structural entity relations are developed by using spatial primitives which are used for the purpose of spatial feature extraction & characterization. Comparison among different associated metrics (and also with non metric) in feature extraction of spatial objects has been made for study purpose.

This article aims to bring out the motivation in user community and researchers to develop associated metrics for studying spatial object features and characteristics of a spatial object.

KEYWORDS: Spatial object, Spatial primitives, Spatial characteristics, Spatial metrics, Spatial features, Non-metric, Feature extraction.

Spatial data is normally represented by more Complex Data types like Points, Lines, Polygons, Spatial Data Structures, Spatial computations. The Spatial and related attribute information is stored separately i.e. Attribute information is stored in database where as spatial information in GIS file structure. The four main properties of spatial data are Geometry, Distribution of objects in space, Temporal changes and

Data volume. The Primitives of Spatial Data Mining are Characteristic Rules, Discrimination Rules, Association rules, Thematic Maps and Image Maps.

This paper stresses mainly on the feature extraction based on association rule. Distance Measures is one of the technique and many of the researchers are interested in this.

To discuss whether a set of points is close enough to be considered a cluster, we need a distance measure $D(x,y)$ that tells how far points x and y are. The usual axioms for a distance measure D are:

1. $D(x,x) = 0$. A point is distance 0 from it self
2. $D(x,y) = D(y,x)$. Distance is symmetric.
3. $D(x,y) \leq D(x,z) + D(z,y)$ which is the triangle law of inequality.

Often, our points may be thought to live in a K -dimensional Euclidean space, and the distance between any two points. Say

$x = \{x_1, x_2, \dots, x_k\}$ and $y = \{y_1, y_2, \dots, y_k\}$ is given in one of the usual manners:

1. Common distance ("L2 norm") : $\sqrt{\sum (x_i - y_i)^2}$ (i=1 to k)
2. Manhattan distance ("L1 norm") : $\sum |x_i - y_i|$ (i=1 to k)
3. Max of dimensions (" L00 norm") : $\max |x_i - y_i|$ (i=1 to k)

When there is no Euclidean space in which to place the points, clustering becomes more difficult, Here are some examples where a distance measure without a Euclidean space makes sense.

Distance between Two Spatial Objects can be found, by defining extensions to the traditional methods. The Euclidean and Manhattan measures are often used to measure the distance between two points. The distance between two spatial objects can be defined as extensions to these two traditional definitions:

- Minimum distance between two spatial objects can be defined as

$$\text{dis}(A, B) = \min \text{dis}(\langle x_a, y_a \rangle, \langle x_b, y_b \rangle)$$

$$\langle x_a, y_a \rangle \in A, \langle x_b, y_b \rangle \in B$$
- Maximum distance between two spatial objects can be defined as

$$\text{dis}(A, B) = \max \text{dis}(\langle x_a, y_a \rangle, \langle x_b, y_b \rangle)$$

$$\langle x_a, y_a \rangle \in A, \langle x_b, y_b \rangle \in B$$
- Average distance between two spatial objects can be defined as

$$\text{dis}(A, B) = \text{average}(\langle x_a, y_a \rangle, \langle x_b, y_b \rangle)$$

$$\langle x_a, y_a \rangle \in A, \langle x_b, y_b \rangle \in B$$
- Center distance between two spatial objects can be defined as

$$\text{dis}(A, B) = \text{dis}(\langle x_{ca}, y_{ca} \rangle, \langle x_{cb}, y_{cb} \rangle)$$

$$\langle x_a, y_a \rangle \in A, \langle x_b, y_b \rangle \in B$$

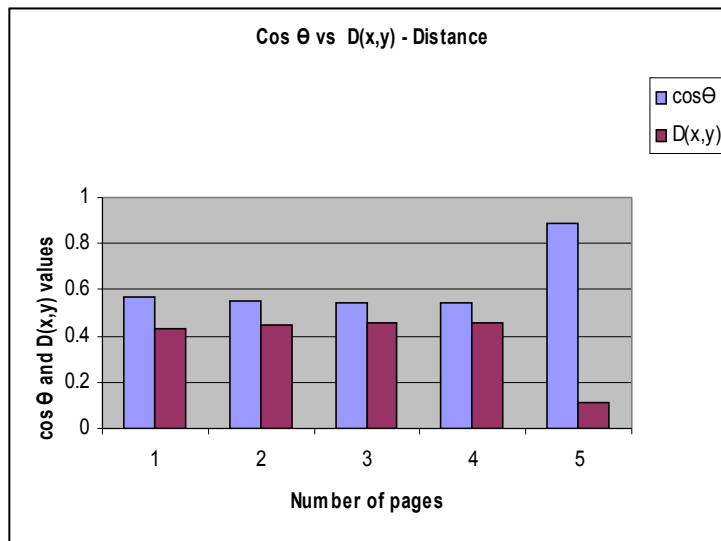
$$\langle x_{ca}, y_{ca} \rangle \text{ is a center point of } A \quad \langle x_{cb}, y_{cb} \rangle \text{ is a center point of } B.$$

The following are the Conclusions based on **D(X,Y)** values by which one can say whether the given two documents are related or not etc.

1. **$D(x,y)=0.0$** "The Documents Are Perfectly Matching and one and the same"
2. **$D(x,y) > 0.0 \ \&\& \ D(x,y) \leq 0.4$**
 "Both the Documents are related to same topic of particular area..."
3. **$D(x,y) > 0.4 \ \&\& \ D(x,y) \leq 0.7$** "The Documents Are Related To Same Area"
4. **$D(x,y) > 0.7 \ \&\& \ D(x,y) \leq 0.9$** "The Documents are not completely related."
5. **$D(x,y) > 0.9$** "The Documents Are Related To Different Areas."

The following are the Table and Graph drawn for Page wise **Cos Θ** values and **D(x, y)** values using **Distance** Formula for two documents.

S.NO.	Cos θ	D(x,y)
1	0.566255	0.433745
2	0.550515	0.449485
3	0.540302	0.459698
4	0.540302	0.459698
5	0.889594	0.110406
Total	3.086968	1.913032
Average	0.617394	0.382606



The D(x,y) value we got by Distance formula = **0.264393**

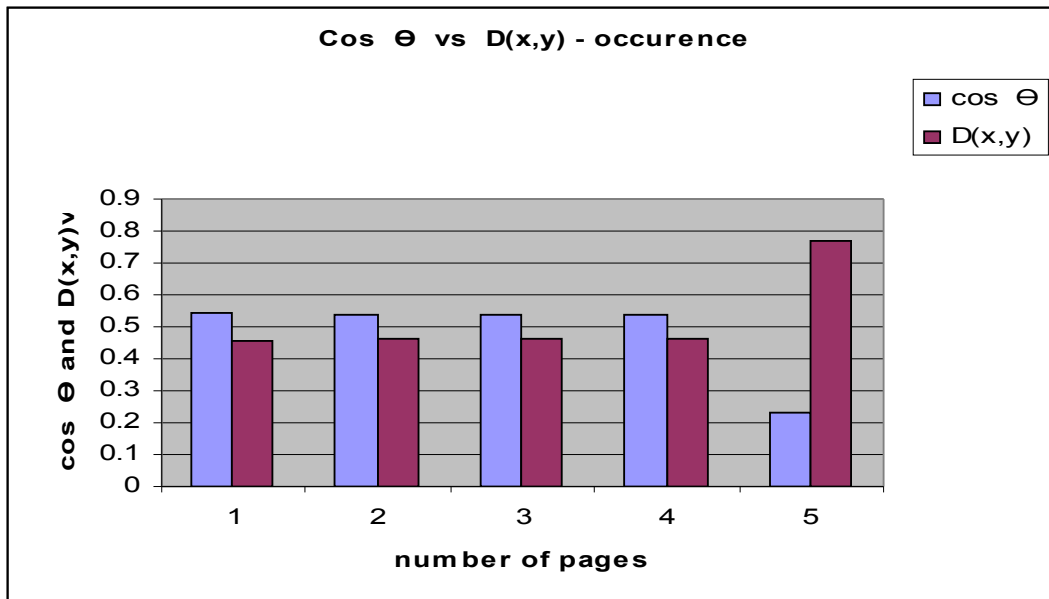
The $D(x,y)$ value we got by average distance formula = **0.382606**

The difference value we got by distance formula from the average(page wise) = **0.118213**

The conclusion for average distance and distance formula is same i.e. topics are related to same area.

The following are the Table and Graph drawn for Page wise $\text{Cos } \Theta$ values and $D(x,y)$ values using **Position based occurrence** Formula

S.NO.	Cos θ	D(x,y)	
1	0.543591	0.456409	
2	0.54046	0.45954	
3	0.540327	0.459673	
4	0.540309	0.459691	
5	0.230461	0.769539	
Total		2.395147	2.604853
Average		0.479029	0.520971



The $D(x,y)$ value we got by occurrence formula = **0.216266**

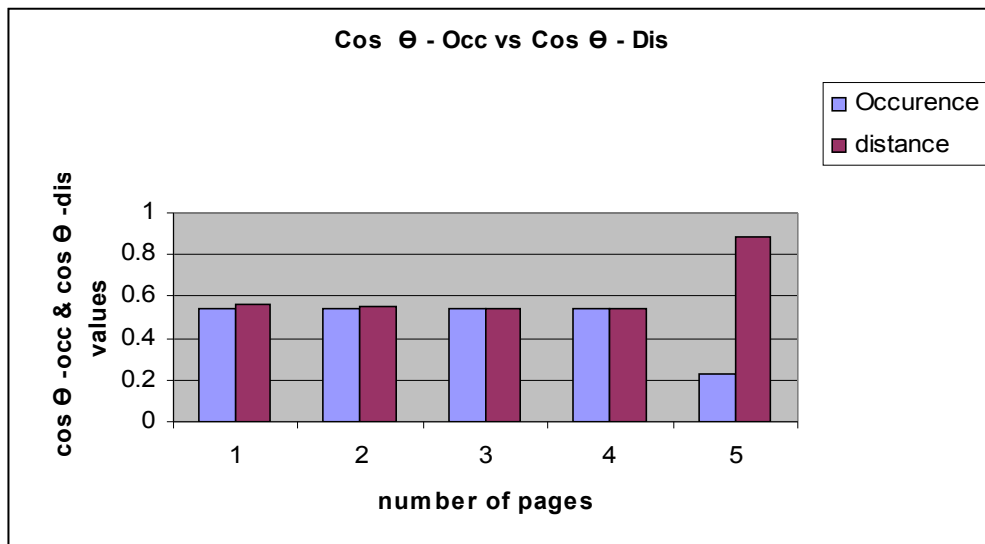
The $D(x,y)$ value we got by average occurrence formula = **0.520971**

The conclusion we got by average occurrence and occurrence formula are not one and

S. No	Cos θ -Occ(p)	Cos θ - dis
1	0.54359103	0.566255
2	0.54045989	0.550515
3	0.5403269	0.540302
4	0.54030876	0.540302
5	0.23046077	0.889594
Total	2.39514736	3.086968
Average	0.47902947	0.6173936

the same and the difference value is **0.304705**

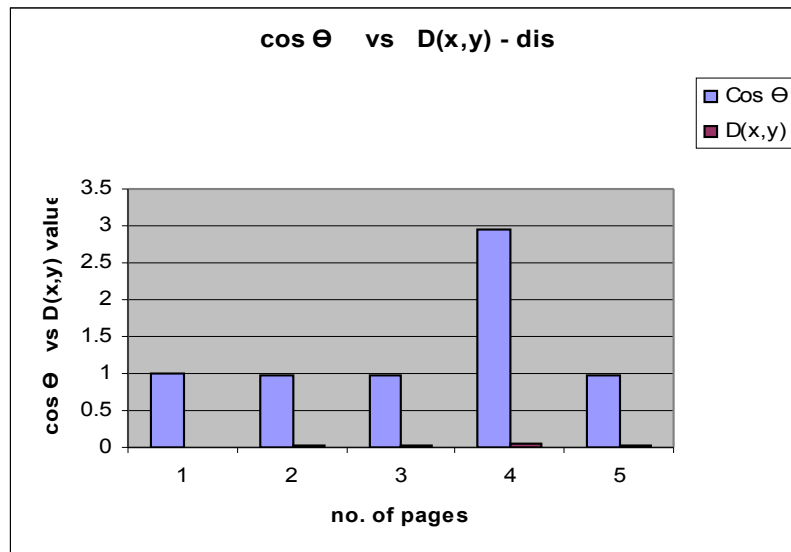
The following table gives the comparison values of cos θ from Position based occurrence formula to Distance formula.



The difference value from occurrence to distance is **0.138364**

The following are the Table and Graph drawn for Page wise **Cos Θ** values and **D(x, y)** values using **Distance** Formula for another two documents.

SNO	Cos θ	D(x,y)
1	0.993091	0.006909
4	0.970994	0.029006
5	0.978466	0.021534
Total	2.942551	0.057449
Average	0.98085	0.01915



The D(x,y) value we got by Distance formula = **0.009938**

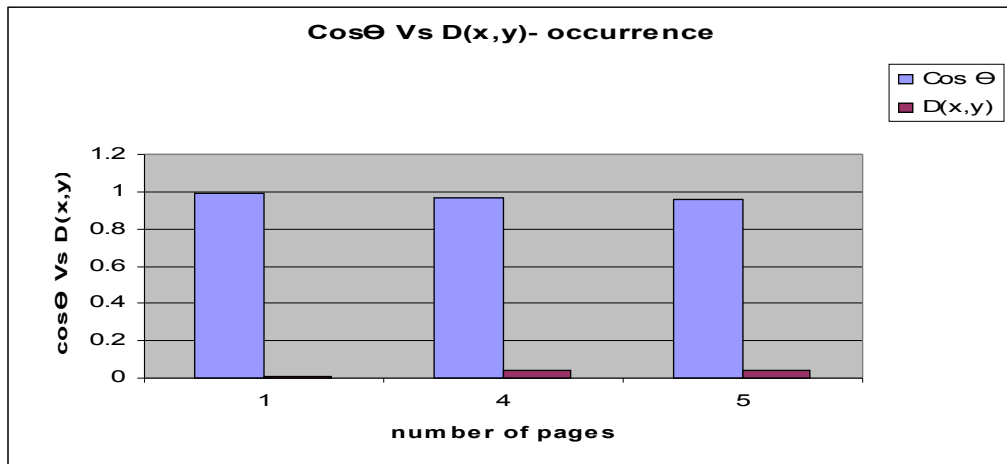
The D(x,y) value we got by average distance formula = **0.01915**

The difference value we got by distance formula from the average (page wise) = **0.009212**

The conclusion for average distance and distance formula is same i.e. topics are related to same area and almost one and the same.

The following are the Table and Graph drawn for Page wise **Cos Θ** values and **D(x, y)** values using Position Based Occurrence Formula for another two documents.

SNO	cos θ	D(x,y)
1	0.994097	0.005903
4	0.962874	0.037126
5	0.95755	0.04245
Total	2.914522	0.085478
Average	0.971507	0.028493



The D(x,y) value we got by occurrence formula = **0.000281**

The D(x,y) value we got by average occurrence formula = **0.028493**

The conclusion we got by average occurrence and occurrence formula are one and the same in conclusion and the difference value is 0.028212.

The following table gives the comparison values of cos θ from Position based occurrence formula to Distance formula.

SNO	cos θ -occ	cos θ -dis
1	0.994097	0.993091
2	0.962874	0.970994

Map World Forum

Hyderabad, India

3	0.95755	0.978466
Total	2.914522	2.942551
Average	0.971507	0.98085



The difference value from occurrence (pos) to distance is **0.009343**

Comparison of Methods based on D (x, y) values is given for few documents as an example

1 (5:8)	0.382606 (Dis.)	Same Topics & related to same area
	0.520971 (Pos-occ.)	Related to same area but not same topics
	0.291717531 (occ.)	Same Topics & related to same area
2 (5:8)	0.01915 (Dis.)	Documents are almost one and the same
	0.028493 (Pos-occ.)	Documents are almost one and the same
	0.148760571 (occ.)	Same Topics & related to same area
3 (1:1)	0.459697694 (Dis.)	Related to same area but not same topics
	0.459692297 (Pos-occ.)	Related to same area but not same topics
	0.459697694 (occ.)	Related to same area but not same topics
4 (1:1)	0.362368602 (Dis.)	Same Topics & related to same area
	0.144483258 (pos-occ)	Same Topics & related to same area
	0.459697694 (Occ.)	Related to same area but not same topics
5 (3:1)	0.206375408 (Dis.)	Same Topics & related to same area
	0.094210045 (Pos-occ.)	Same Topics & related to same area
	0.0664157 (occ.)	Documents are almost one and the same

OCCURRENCE FORMULA:

Advantages:

- Whether there will be an occurrence of particular word/ words exists, if so with what occurrence it exists in two documents either page wise or document wise.

- Able to know whether can we come up with a rough idea that they are related to same area or not.
- If we make it as preliminary test it will be advantageous.
- Time taken will be small (i.e. time complexity is not much or less time consuming).
- Even space complexity doesn't arise (i.e. It won't require much space also).

Disadvantages:

- The accuracy in conclusion may not be that much advantageous to arrive at conclusion.
- It won't capture the **Local information**.
- It is based on only **statistical information**, which is not sufficient to arrive at conclusion perfectly.
- It is not independent of nature (i.e. format style, font size, bold, etc.) of the document by which the occurrence number of the key word will change accordingly. By which, it gives a different result, although documents are one and the same.
- Suppose only one page is available in each document and the occurrences are for example 2 and 5 then
$$\cos\theta = (2 \times 5) / (\sqrt{2^2} \times \sqrt{5^2}) = 1$$
 i.e. Independent of occurrence number which gives wrong conclusion.

DISTANCE FORMULA

Advantages:

- The differences in occurrence numbers (i.e. first, second, third, etc. occurrences numbers) will be taken in to consideration for arriving to the result.
- It uses distance as the factor in concluding the result.
- The results are more appropriate compare to the other methods.
- It is independent of nature (i.e. format style, font size, bold, etc.) of the document.
- It can be treated as final test.
- It captures local information of the document. So it is more relevant approach.
- Even if any text is added to it in the beginning or end of any paragraph or to the document this won't get effected.
- If some text is added in middle the occurrence number and by which distance will also change. With this little appropriateness will change and doesn't affect much.

Disadvantages:

- The time complexity will be more compare to occurrence formula method.
- Even the space complexities are more, compare to occurrence formula method.

POSITION BASED OCCURRENCE:

Advantages:

- The occurrences number (position) of a keyword either page wise or document wise will be taken into consideration.
- In most of the cases it is as good as Distance (metric) method
- It is less time consuming compare to Distance method.
- It is better method compare to occurrence method.

Disadvantages:

- Time complexity is more comparing to occurrence method.
- This method may not give exact conclusion in some of the cases as Distance method gives.

CONCLUSIONS:

It is found that from the above experiments and statistical information among the three methods the Distance Metric is far good in arriving to the exact conclusion that, whether the two given documents are one and the same, related to same area, or related to different areas etc.

REFERENCES:

- <http://www.eng.tau.ac.il/~maimon/ifn-kdd/>
- maimon@eng.tau.ac.il
- Mlast@csee.usf.edu
- <http://www.digimine.com/usama/datamine>
- <http://allanon.gmd.de/and/descartes.html>
- <http://www.dbminer.com>
- xkuba@fi.muni.cz
- wuw@cs.umn.edu
- bapics@uohyd.ernet.in
- <http://datamining.csiro.au/adm02>