COMPARISON OF SIMILARITY AND DISSIMILARITY FOR INFORMATION RETRIEVAL

[1]Khin Lay Myint, [2]Khin Shin Thant, [3]Moe Thidar Naing,[4]Hlaing Htake Khaung Tin
Faculty of Information Science, University of Computer Studies, Hinthada, Myanmar[1, 2], Faculty of Computer Science, University of Computer Studies, Pakokku, Myanmar[3,4]
khinlaymyint.cu@gmail.com[1], shinthantkhin10@gmail.com, moethidarnaig2018@gmail.com[3], hlainghtakekhaungtin@gmail.com[4]

ABSTRACT

This paper existing a model approach for the simulation of different distance for doing approximations. A brief existence for the first time to the most common approximation metrics for distance computations and their applications within the context of simulation is already arranged. A discrete-event simulation model present using a practical example for the calculation of distances. Comparing different distance metrics with the distances on an exemplary can determine choice a satisfactory distance metric exists for calculating distances in data analysis, not having the incorporate of an external route planning web service.

Index Terms—Management Information System, proximity measures, similarity distance, dissimilarity distance and informational retrieval.

INTRODUCTION

"Information use" is to involve understanding what information sources people choose and the way in which people use information to make sense of their lives and situations. Data define information that is used by people to make decision.

Information systems are a new institution for conducting business, health, education, social, sports, defense, government, construction and etc. Information use is personal that if people have a way of entering the information, they will use it. People's use of many types of information sources varies according to individual factors. Information use can be making important as in organizational decision-making, as when information is happened continuously motivate.

Management information systems the most exciting topic is not interruptions change in technology, management applies the technology, and the impact on obtain success. New information show, old ones decline, and complete firms are those who learn how to apply the new technology.

Similarity and dissimilarity are very important, so, they are applied by a number of data mining techniques, common people require ways to assess how similar otherwise dissimilar objects are in judgment to one another. The information is a group of data objects that the objects in the correct data are similar and dissimilar to the objects in other incorrect data. Information analysis as well people information foundation methods to recognize probable information because highly dissimilar is been by objects to others. The classification can use knowledge of object similarities where a class label assigns a given object based on other objects in the model.

BACKGROUND THEORY

The kinds of proximity measures, those are Nominal attributes and Binary attributes and Ordinal attributes. The typical of Proximity measures, the similarity or dissimilarity to be real between the objects, items, stimuli, or persons that underlie an empirical study.

Between the two objects of similarity is a numeral measure of the degree, the two objects are alike. As a result, similarities are not lower for pairs of objects that are more alike. Similarities are usually positive and are often between 0 and 1 (no similarity and complete similarity).

Between the two objects of dissimilarity is the numerical measure of the degree to which the two objects are not the same. Dissimilarity is not higher for more similar pairs of objects.

The distance is applied as a synonym for dissimilarity. Sometimes, dissimilarities fall in the interval (0, 1), but its' range is from 0 to ∞. The distance measures are applied for calculate the dissimilarity of objects express by numeric attributes. These measure contain the Euclidean, Manhattan, and Minkowski distances.

This research compare these measure of distance because we have more reliable for computing. Euclidean distance is the first distance. The two objects is: $i = (x_{i1}, x_{i2}, ..., x_{ip})$

$j = (x_{j1}, x_{j2}, ..., x_{jp})$

Described by p = numeric attributes. The two objects of Euclidean distance defined follow:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

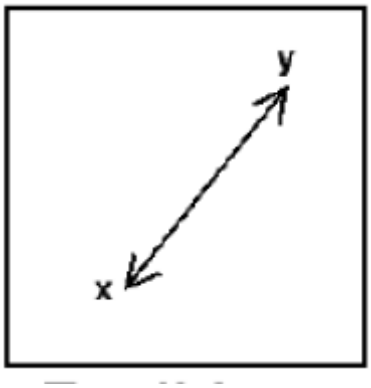

Figure 1: Euclidean distance

The second distance measure is Manhattan distance, if an attribute is numeric, then the local distance function can be defined as the absolute of values, local distances are often normalized so that they lie in the range 0..... 1, defined follow:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

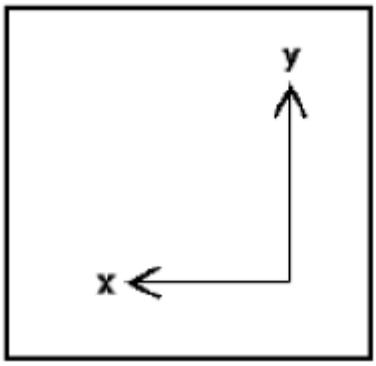

Figure 2: Manhattan distance

Euclidean and Manhattan distance gratify the mathematical properties as following:

Positivity: $d(i, j) \geq 0$ (positive number)

Identity of indiscernible: $d(i, i) = 0$ (an object to itself)

Symmetry: $d(i, j) = d(j, i)$ (functions)

Triangle inequality= $d(i, j) \leq d(i, k) + d(k, j)$ (directly from object *i* to *j* into space is less than making a detour over any other object *k*)

The third distance measure is Minkowski distance. That is a generalization of the Euclidean and Manhattan distances. The following defined:

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}.$$

h = real number

Lp norm calls a distance in some literature, where our notation of *h* prefers from the symbol *p*. *p* have been kept, the amount of attributes to be reliable.

Last distance measure is supremum distance. Minkowski distance's generalization for $h \to \infty$. To pull out the result it, the attribute *f* discover the maximum difference in values, those are given between the two objects. The supremum distance is following defined:

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

The L^{∞} norm = *uniform norm.*

A measure of similarity is been Cosine similarity that can be applied to contrast documents, appear documents' ranking with admiration to a given vector of query words. Different between two vector comparison, x and y. A similarity function is measured as following:

$$sim(x, y) = \frac{x \cdot y}{||x||||y||},$$

||x || = vector x's Euclidean norm

$x = (x_1, x_2, ..., x_{p)}$

The solution: $\sqrt{x1^2 + x2^2 + x2^2 \ldots\ldots\ xp^2}$

The measure calculates the angle of cosine between vector x and y.

- cosine value 0 = no match (two vectors)

- cosine value = the two vectors, small angle and big match

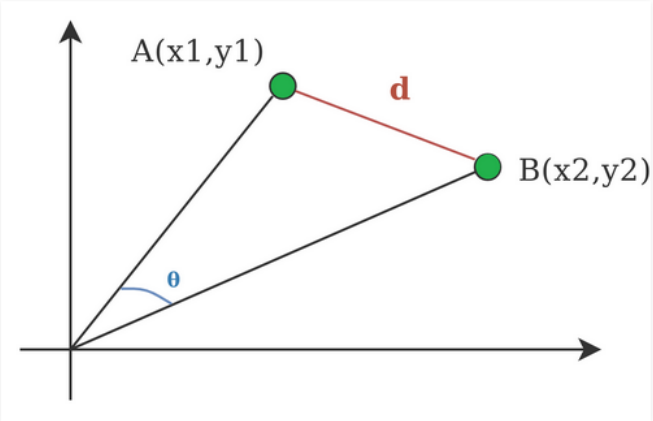

Figure 3: cosine similarity

This table is having a great effect to determine the similarity measures in data analysis. There is no common accept subject similarity measure. The depending on the similarity measure can use result. Dissimilarity measure may be alike after some transformation.

2-D dataset for Minkowski distance the following:

- x = new data point
- x = (1.4, 1.6)
- h = 3

Table 1: dataset table

	a_1	a_2	a_3	a_4	a_5
p_1	1.5	2	1.6	1.2	1.5
p_2	1.7	1.9	1.8	1.5	1.0

Table 2: Cosine Similarity distance result.

	a_1	a_2	a_3	a_4	a_5
Cosine similarity	0.9999913914	0.9957522613	0.9999694838	0.999028253	0.9653633931

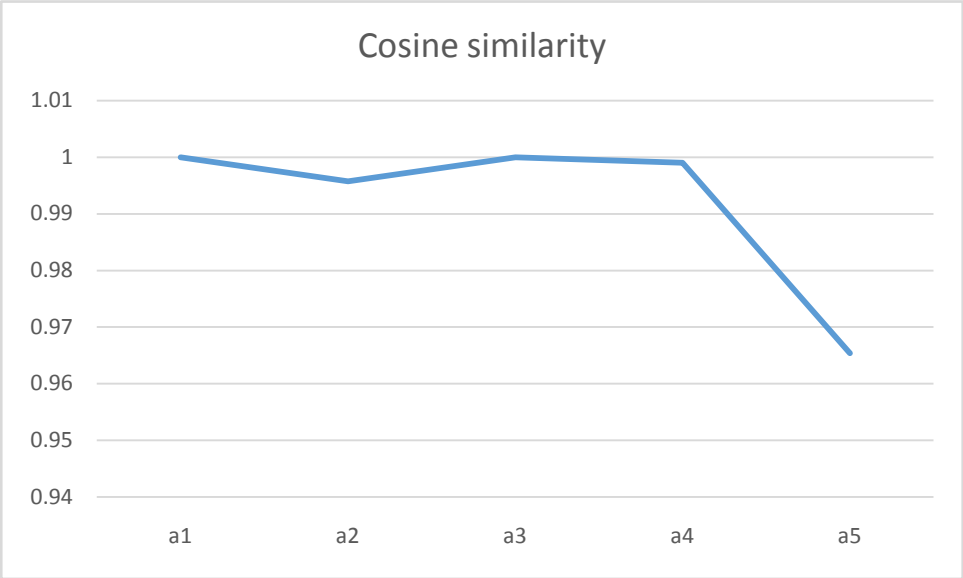

Figure 4: cosine similarity

Table 3: Dissimilarity distance result.

	a_1	a_2	a_3	a_4	a_5
Euclidean distance	0.1414213562	0.6708203932	0.2828427125	0.2236067977	0.608276253
Manhattan distance	0.2	0.9	0.4	0.3	0.7
Minkowski distance	0.125992105	0.6240251469	0.25198421	0.2080083823	0.5990726415
Supremum distance	0.1	0.4	0.2	0.1	0.1

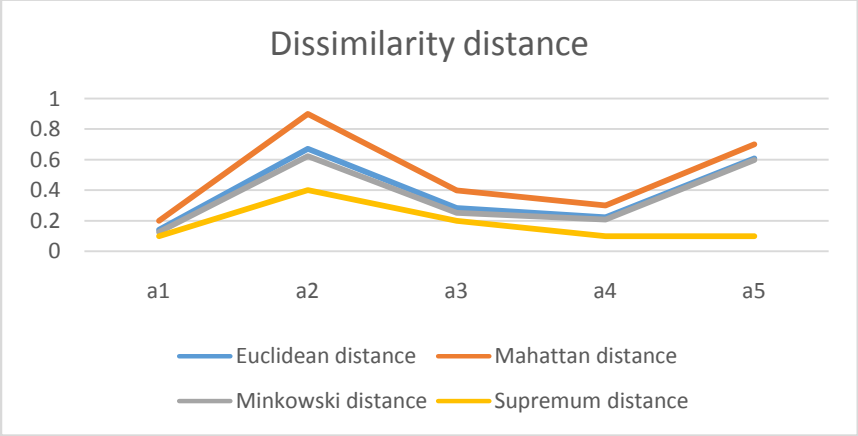

Figure 5: Dissimilarity distance

Table 3: similarity and dissimilarity distance result

	a_1	a_2	a_3	a_4	a_5
Euclidean distance	0.1414213562	0.6708203932	0.2828427125	0.2236067977	0.608276253
Manhattan distance	0.2	0.9	0.4	0.3	0.7
Minkowski distance	0.125992105	0.6240251469	0.25198421	0.2080083823	0.5990726415
Supremum distance	0.1	0.4	0.2	0.1	0.1
Cosine similarity	0.9999913914	0.9957522613	0.9999694838	0.999028253	0.9653633931

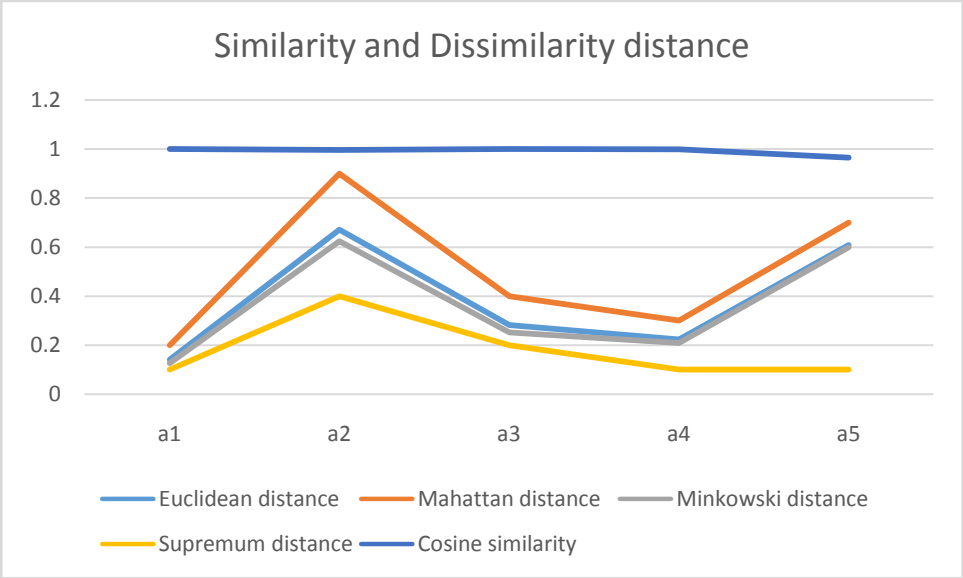

Figure 6: similarity and dissimilarity distance

EXPERIMENTAL RESULTS

It is result that references to all data are available into acknowledgment section. A set of similarity and dissimilarity compute for above showed 5 dataset for clarify and compare the distance measure. The results of dataset are represented in Table 3 or Figure 6. The experiments were conducted using similarity (cosine similarity) and dissimilarity (Euclidian distance, Manhattan distance, Minkowski distance and Supremum distance, which are distance-based. As it is showed in index hand out to evaluate and compare the results.

The similarity and dissimilarity were used in this experiment. Due to the fact similarity and dissimilarity results are support for the best of decision making about the distance. Minkowski distance is a generalization of

the Euclidean and Manhattan distance. Supremum distance is the generalization of the Minkowski distance. Therefore, according to experiment result, Euclidean distance is the best distance more than Minkowski distance.

CONCLUSION

Selecting the distance measure is one of the confrontations come upon when trying to people a distance-based to a dataset. The check of similarity measures reason confusion and complexities in choosing a suitable measure. Similarity calculates may execute differently for datasets with similarity and dissimilarity. This paper's aim was to make clear and more reliable for information in the experiments. In this result, similarity and dissimilarity were compared using simple 5 dataset. Overall, the results point out is among the top most reliable measure for all people. Moreover, this measure is one of the more reliable distance. However, this measure is mostly recommended for all technicians.

REFERENCES

1. Apparicio P, Abdelmajid M, Riva M, Shearmur R. Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. International Journal of Health Geographics.. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.
2. https://slidewiki.org/print/1265/data-mining/1280-2/2.
3. Baroni-Urbani, C., Buser, M.W., (1976), "Similarity of Binary Data"
4. Shirkhorshidi AS, Aghabozorgi S, Wah TY, Herawan T. Big Data Clustering: A Review Computational Science and Its Applications.
5. Aczél J, Saaty TL Procedures for synthesizing ratio judgements. J Math Psychol.
6. Balakrishnan V, Sanghvi LD Distance between populations on the basis of attribute data. Biometrics.
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3835347/