

Effective Classification and Categorization for Categorical Sets: Distance Similarity Measures

Heba Ayeldeen, Mahmood A. Mahmood and Aboul Ella Hassanien

Abstract Measuring the similarity between objects is considered one of the main hot topics nowadays and the main core requirement for several data mining and knowledge discovery task. For better performance most organizations are in need on semantic similarity and similarity measures. This article presents different distance metrics used for measuring the similarity between qualitative data within a text. The case study represents a qualitative data of Faculty of medicine Cairo University for theses. The dataset is about 5,000 thesis document with 35 departments and about 16,000 keyword. As a result, we are able to better discover the commonalities between theses data and hence, improve the accuracy of the similarity estimation which in return improves the scientific research sector. The experimental results show that Kulczynski distance yields better with a 92.51 % without normalization that correlate more closely with human assessments compared to other distance measures.

Keywords Text clustering · Similarity · Classification

H. Ayeldeen (✉) · M.A. Mahmood · A.E. Hassanien
Scientific Research Group in Egypt (SRGE), Giza, Egypt
e-mail: heba.ayeldeen@gmail.com

M.A. Mahmood
e-mail: dr_mahmoodissr@hotmail.com

A.E. Hassanien
e-mail: aboitcairo@gmail.com

H. Ayeldeen · A.E. Hassanien
Faculty of Computers and Information, Cairo University, Cairo, Egypt

M.A. Mahmood
Institute of Statistical Studies and Research (ISSR), Giza, Egypt

1 Introduction

Text classification is important means and methods of text mining, and also a part of data mining. Text classification is a supervised classification of documents, which divides and categorize a text collection into several subsets called sets, the text of each set has greater similarity than the one in different set. Text classification differs from text clustering where labels within the data sets are well known compared to clustering that deals with unsupervised documents [1].

Recently, semantic similarity has been applied in many different applications including the health and medical domain. For better understanding the textual resources, semantic similarity estimation is used [2].

Properly classifying and clustering texts based on certain criteria and trend improves the understanding and relatedness of data. Decision makers find it easy to extract the knowledge based on good classification [3]. Analyzing the relationships between documents based on concepts and terms is one of the semantic analysis methods. The best way to make use of the information retrieved rather than rebuilding it from scratch [4].

Text categorization has recently become a hot topic in the area of information retrieval. The main objective of text categorization is to assign free text documents to one or more predefined categories based on their content. Traditionally this process is performed manually by domain experts. This process is very time-consuming and costly, thus limiting its applicability [5]. The concepts of text similarity and semantic similarity have been extensively developed to tackle this problem in an automatic way. There are many text similarity methods that can be used to ease text categorize, of these methods are: Semantic sequence kin model (SSK), which extracts semantic sequences from a text as its feature strings, and then takes into account both word and word position when 2 semantic sequences are compared [6]. Common semantic sequence model (CSSM), which is similar to semantic sequence kin model, but uses another formula to calculate similarity of semantic sequences, in which the word position is not considered [7]. Word similarity estimation has many direct applications. In word-sense disambiguation, context terms can be semantically compared to discover the most similar sense. In document categorization or clustering, the semantic similarity between words can be compared to group documents according to their subject. In word spelling correction, semantic similarity can assess which is the most appropriate correction for a potential misspelling according to its similarity against context words [3].

The concepts of similarity and distance are crucial in many scientific applications. Similarity and distance measurements are mostly needed to compute the similarities/distances between different objects, an essential requirement in almost all pattern recognition applications including clustering, classification, feature selection, outlier detection, regression, and search. There exist a large number of similarity measurements in the literature; thus, selecting one most appropriate similarity measurement for a particular task is a crucial issue. The success or failure

of any pattern recognition technique largely depends on the choice of the similarity measurement. The similarity measurements vary depending on the data types used [8].

The rest of this paper is organized as follows. Section 2 states brief explanation of different similarity measures types with respect to the distance. Section 3 focus on optimizing documents/texts as by showing mathematically comparisons between the distance based similarity measures. Finally, Sect. 4 addresses conclusions and future work.

2 Types of Similarity Measures

Distance based similarity measure is considered as an essential parameter for the classification and measuring the similarity or dissimilarity between two or more vectors/objects (Fig. 1).

Following are distance metric families with 34 different distance metrics used for matching the distance and measuring the similarity between any two or more vectors.

2.1 L1 Minkowski Distance Metric Family

In L1 distance metric family all distance metrics are defined for $p = 1$.

1. Sorenson distance [9]

$$d_{Sor}(A, B) = \frac{2c}{a + b} \tag{1}$$

We say two points are adjacent if their L1 distance is defined for $a = 1$. Eq. 1, is known as Sorenson distance sometimes called or Bray Curtis where difference is normalized by the sum of feature vectors at each dimension.

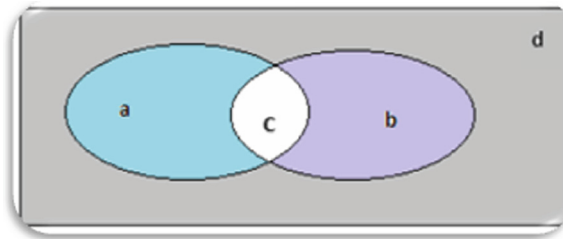


Fig. 1 Variables in similarity equations. **a** Set1. **b** Set2. **c** Intersection/common features. **d** Information system/dataset

2. Kulczynski distance [9, 10]

$$d_{Kul}(P, Q) = \frac{\left(\frac{a}{a+b} + \frac{a}{a+c}\right)}{2} \quad (2)$$

While Eq. 2 shows that the Kulczynski distance is measured as the difference of the two vectors A and B normalized by the minimum value between two of the features in A and B.

3. Gower distance [9, 10]

$$d_{Gow}(A, B) = \frac{a + d}{\sqrt{(a - c)(a + c)(b + d)(c + d)}} \quad (3)$$

In 1985, Gower defined a general similarity coefficient to overcome this as in Eq. 3 where A and B for binary variables stands for the presence/absence of that variable.

2.2 Inner Product Distance Metrics Family

The inner product distance metrics family deals exclusively with similarity measures which incorporate the inner product of two vectors yields a scalar and is sometimes called the scalar product or dot product.

1. Cosine Similarity [9–16]

$$d_{Cos}(A, B) = \frac{a}{\sqrt{(a + b)(a + c)^2}} \quad (4)$$

Ochiai and Carbo are other names to the cosine co-efficient. As given in Eq. 4, the distance is measured based on the angle between two vectors and is thus often called the angular metric.

2. Jaccard Coefficient [9–11, 13, 15, 17–19]

$$d_{Jack}(A, B) = \frac{a}{a + b + c} \quad (5)$$

Similarly Eqs. 4 and 5 measure the similarity between two vectors as it gives the minimum value for more similar feature vectors.

3. Dice Coefficient [9, 10, 17]

$$d_{Dice}(A, B) = \frac{2a}{2a + b + c} \quad (6)$$

Dice distance is very sensitive to small changes as the numerator is multiplied by factor 2.

2.3 Squared Chord Distance Metrics Family

1. Squared Chord [9, 10]

$$d_{Schi}(A, B) = \sqrt{2 \left(1 - \frac{a}{(a+b)(a+c)} \right)} \quad (7)$$

The sum of geometric means \sqrt{AB} , where A referred to as fidelity similarity or squared chord distance metric. Squared chord distance is given by Eq. 7. In squared chord distance, the sum of square of square root difference at each dimension is taken which increases the difference for more dissimilar feature vectors. This distance cannot be used for feature space with negative values.

2. Hellinger distance [9]

$$d_{Heling}(P, Q) = 2 \sqrt{\left(1 - \frac{a}{(a+b)(a+c)} \right)} \quad (8)$$

The Hellinger distance is another way to measure the distance to check the similarity of even the dissimilarity of two or more points. In 1909 introduced by Ernst Hellinger is the Hellinger distance which is used to quantify the similarity between two probability distributions. Talking about the discrete distribution, for two discrete probability distributions A and B their Hellinger distance is defined as in Eq. 8 which is directly related to the Euclidean norm of the difference of the square root vectors.

2.4 Other Distance Metrics

1. Average distance [9, 13]

$$d_{Avg}(A, B) = \frac{\sum |a - b| + \max |a - b|}{2} \quad (9)$$

Distance given by Eq. 9 is just the average of city block and Chebyshev distance.

2. Sokal and Sneath II [10, 17]

$$S_{SOKAL\&Sneath-II} = \frac{a}{a + 2b + 2c} \quad (10)$$

For the similarity measure as in Eq. 10, this measure has a minimum value of 0, has no upper limit, and is undefined when there are no non matches ($b = 0$ and $c = 0$).

3 Experimental Results

The main target of the study is to increase the scientific research field in the different faculties of Cairo University. To do so, we focused in the theses mining concept for instance for the Faculty of Medicine, Cairo University. Data was collected from the digital library of the Faculty of Medicine and then cleansed for further process. The data collected was theses documents including the title of the theses and the abstract with keywords. The theses documents are classified into Master and doctorate theses. Documents are tracked within the last 10 years separated and categorized based on the departments within the Medicine school. Data collected was 4,878 theses with 15,808 keyword across all departments.

Faculty of Medicine in Cairo University is classified into 35 departments. The aim of the work is to find out the departments that can work together easily to increase the research within the faculty. As well as making it easy for departments to find a way to increase the research within each department.

Different distance measures were applied to get the nearest departments that can work together. Results includes measures of the equations Euclidean distance, City block distance, Sorenson distance, Kulczynski distance, Cosine Similarity, Jaccard Coefficient, Dice Coefficient, Hellinger distance, Sokal and Sneath and Johnson similarity equation.

As previously mentioned, the main objective is to find out the best of departments' combinations that can collaborate easily together. After applying the Cosine similarity on four randomly selected departments (Anatomy, Andrology, Anesthesia and Cardiology), the results shows for instance that the department of Anatomy can work easily with the department of Cardiology with a percentage of 60 %. While the department of Anesthesia with Cardiology by 90 %. While after applying the Dice distance equation the combination of Anatomy with the Cardiology decreased to 40 %. Then the results decreased again with the Jaccard Coefficient, Sokal and Sneath as well as Sorenson; and increased again with the Kulczynski distance equation Tables 1, 2, 3, 4, 5, 6.

After making comparisons between the different distance equations (Cosine similarity distance, Dice Coefficient, Jaccard Coefficient, Sokal and Sneath distance, Kulczynski distance and Sorenson distance), we find out that the Kulczynski

Table 1 Applying cosine similarity

Departments	Anatomy	Andrology	Anesthesia	Cardiology
Anatomy	1	0.6	0.2	0.5
Andrology	0.6	1	0.3	0.6
Anesthesia	0.2	0.3	1	0.9
Cardiology	0.5	0.6	0.9	1

Table 2 Applying Dice coefficient

Departments	Anatomy	Andrology	Anesthesia	Cardiology
Anatomy	1	0.6	0.1	0.4
Andrology	0.6	1	0.2	0.5
Anesthesia	0.1	0.2	1	0.9
Cardiology	0.4	0.5	0.9	1

Table 3 Applying Jaccard coefficient

Departments	Anatomy	Andrology	Anesthesia	Cardiology
Anatomy	1	0.4	0.1	0.3
Andrology	0.4	1	0.1	0.3
Anesthesia	0.1	0.1	1	0.8
Cardiology	0.3	0.3	0.8	1

Table 4 Applying Kulczynski distance

Departments	Anatomy	Andrology	Anesthesia	Cardiology
Anatomy	1	0.7	0.5	0.6
Andrology	0.7	1	0.5	0.6
Anesthesia	0.5	0.5	1	0.9
Cardiology	0.6	0.6	0.9	1

Table 5 Applying Sokal and Sneath distance

Departments	Anatomy	Andrology	Anesthesia	Cardiology
Anatomy	1	0.3	0	0.2
Andrology	0.3	1	0	0.2
Anesthesia	0	0	1	0.7
Cardiology	0.2	0.2	0.7	1

Table 6 Applying Sorensen distance

Departments	Anatomy	Andrology	Anesthesia	Cardiology
Anatomy	1	0.1	0	0.1
Andrology	0.1	1	0	0.1
Anesthesia	0	0	1	0
Cardiology	0.1	0.1	0	1

Fig. 2 Anatomy with others

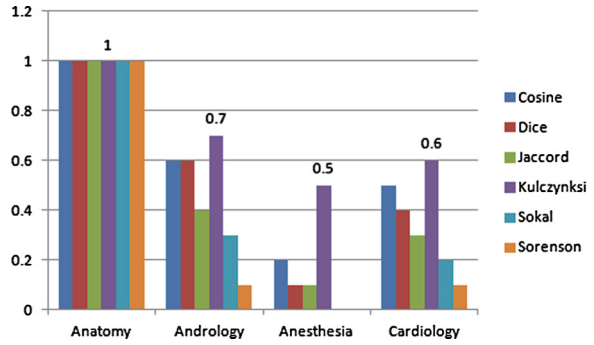


Fig. 3 Andrology with others

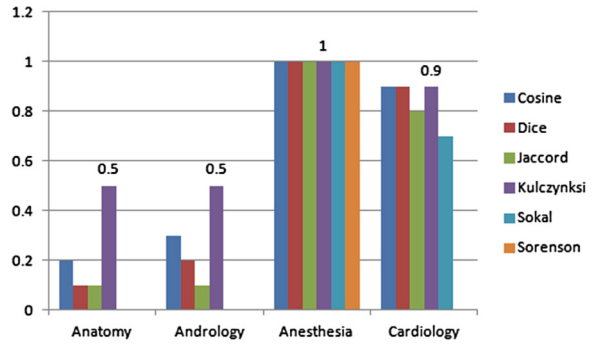


Fig. 4 Anesthesia with others

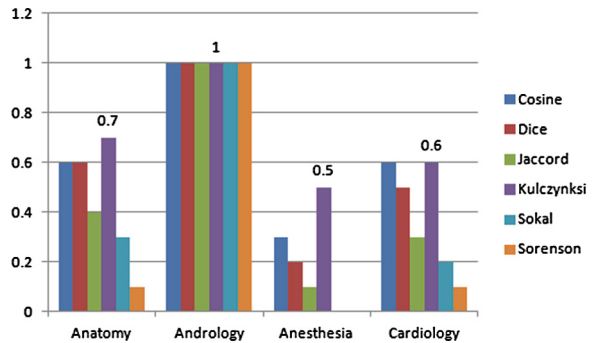
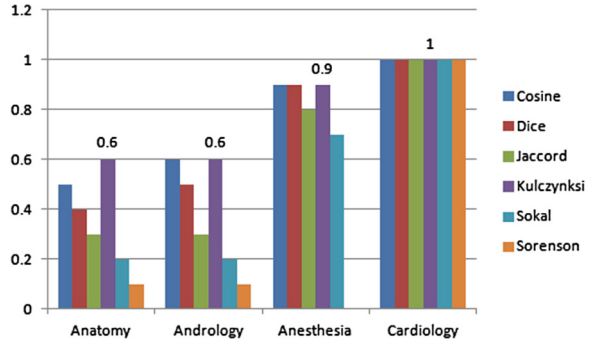


Fig. 5 Cardiology with others



distance is better than any other equations used as shown in the figures below which is more relevant to human assessments (Figs. 2, 3, 4 and 5).

4 Conclusion and Future Work

In this paper, we showed mathematically how texts can be clustered by using different similarity distance equations on documents. By taking an example where these documents were clustered and classified better with the Kulczynski distance by 92.51 % directly without applying normalization than other distance measures. Other algorithms can be considered as well for future work, like applying the genetic programming; neural networks and comparing the results simultaneously.

References

1. Dumais, S., Meek, D., Metzler, D.: Similarity measures for short segments of text. In: Proceeding ECIR'07 Proceedings of the 29th European Conference on IR Research, pp. 16–27 (2007)
2. Hassanien, A.E., Fasmly, A.A., Ayeldeen, H.: Evaluation of semantic similarity across MeSH ontology: A cairo university thesis mining case study. In: 12th Mexican International Conference on Artificial Intelligence, pp. 139–144. Mexico City (2013)
3. Batet, DSaM: Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *J. Biomed. Inf. Arch.* **44**, 749–759 (2011)
4. Kitasuki, T., Aritsugi, M., Rahutomo, F.: Test collection recycling for semantic text similarity. In: Proceeding IIWAS12 Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services, pp. 286–289 (2012)
5. Liu, T., Guo, J.: Text similarity computing based on standard deviation. In: *Advances in Intelligent Computing*, vol. 1, pp. 23–26. Springer, Berlin (2005)
6. Shen, J.Y., Bao, J.P., Liu, X.D., Liu, H.Y., and Zhang, X.D.: Finding plagiarism based on common semantic sequence model. In: *Proceedings of the 5th International Conference on Advances in Web-Age Information Management*, pp. 640–645 (2004)

7. Lyon, C.M., Bao, J.P., Lane, P.C.R., Ji, W., Malcolm, J.A.: Copy detection in chinese documents using ferret. *Lang. Resour. Eval.* 1–10 (2006, in press)
8. Bandyopadhyay, S., Saha, S.: Unsupervised classification. In: *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*, pp. 59–73. Springer, Berlin (2013)
9. Bharkad, S.D., Kokare, M.: Performance evaluation of distance metrics: Application to fingerprint recognition. *Int. J. Pattern Recognit. Artif. Intell.* **25** (2011)
10. Choi, S.H., Choi, S.S., Tappert, C.C.: A survey of binary similarity and distance measures. *J. Syst. Cybern. Inf.* **8**(1), 43–48 (2010, Key: citeulike:7358808)
11. Huang, A.: Similarity measures for text document clustering. In: *New Zealand Computer Science Research Student Conference*, pp. 49–56 (2008)
12. McGill, M.J., Salton, G.: *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983)
13. Leydesdorff, L.: Similarity measures, author cocitation analysis, and information theory. *J. Am. Soc. Inform. Sci. Technol.* **56**, 769–772 (2005)
14. S. B. a. S. Saha, *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications: Springer Berlin Heidelberg*, 2013
15. Lalitha, Y.S., Sandhya, N., Govardhan, A., Anuradha, K.: Analysis of similarity measures for text clustering. In: *International Conference on Information Systems Design and Intelligent Applications*, p. 976, Vishakhapatnam (2012)
16. Leydesdorff, L., Zaal, R.: Co-words and citations. Relations between document sets and environments. *Informetrics*, vol. 87, pp. 05–119. Elsevier, Amsterdam (1988)
17. De Baets, S.J.B., De Meyer, H.: On the transitivity of a parametric family of cardinality-based similarity measures. *Int. J. Approx. Reason.* **50**, 104–116 (2009)
18. Tang, C., Zhang, A., Jiang, D.: Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* **16**, 1370–1386 (2004)
19. Faria, D., Pesquita, C., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. *PLOS: Comput. Biol.* **5** (2009)