A Similarity Measure for Text Document Using Term Cardinality

S. Sumathi¹, G. Hannah Grace²

¹School of Advanced Sciences, Division of Mathematics, Vellore Institute of Technology, Chennai, Tamilnadu, India.
²E-mail: sumathi.manivannan04@gmail.com

Abstract- With enormous development of digital technology, data is being generated at rapid rate with various application domains. Data has to be extracted or filtered to find useful information. A basic concept for these tasks and applications are the distance measures to effectively determine how similar two objects are. In this paper, a novel similarity measure for clustering text documents is proposed using the cardinality of the terms in the documents. The benchmark algorithm k-medoids is used for clustering task. The results obtained from the proposed distance measure are compared with other standard distance measures like Manhattan, Euclidean distance measure. Dunn Index is used to analyze the cluster validation of the results obtained from the distance measure.

Keywords: Document Clustering, k-Medoids, Similarity/Distance Measure, Dunn Index.

1. Introduction

In past few years there has been a sudden increase in the volume of text documents used on internet. Thus, there is a need for efficient methods for understanding the documents by the machine itself. To address this problem, document similarity plays a vital role for machines to process natural language. A few applications of documents similarity are document clustering, document summarization, document categorization and query-based search [1].

Text documents can be represented as vector space model. In this model, each document is represented as a vector, in which the components are each term of the document. Usually, raw text documents are not preferred for clustering text. Hence, the document has to be pre-processed before clustering task. This preprocessed text document is called a corpus and is used as an input document for clustering. The steps involved in pre-processing are tokenization, stop-word removal, converting to upper case to lower case and stemming [2]. The corpus is now transformed to term document matrix by using the raw count of the terms present in the document. This term document matrix forms the base of the document clustering task. This matrix can be used with several known weighting scheme. The most popular weighting scheme is the term frequency and inverse document frequency (tf-idf) [3]. The flowchart of proposed approach of this paper is shown in figure 1.
2. Distance Measures

Distance measure is a numerical value which describes how dissimilar two objects are. The similarity measure is another notion of distance measure i.e., how similar two objects are. The similarity value ranges from 0 (no similarity) to 1 (most similarity) [4]. The distance is given as 1 minus the similarity value. The most similar objects have the least distance implying that the similarity measure is high and vice-versa.

2.1. Definition

A metric on a set X is a function called the distance function or simply distance, if there exist a map \( d: X \times X \rightarrow \mathbb{R} \) (where \( \mathbb{R} \) is the set of real numbers) such that for all \( x, y, z \) in X, the function satisfy the following axioms:

1. \( d(x, y) \geq 0 \) (non-negativity, or separation axiom)
2. \( d(x, y) = 0 \) if and only if \( x = y \) (coincidence axiom)
3. \( d(x, y) = d(y, x) \) (symmetry axiom)
4. \( d(x, z) \leq d(x, y) + d(y, z) \) (Triangle inequality)

2.2. Euclidean Distance

The Euclidean distance between two data points is defined by the square root of the sum of the squared differences of the all co-ordinates of the data points. For \( x, y \) in \( n \)-dimensional space, the Euclidean distance \( d \) is defined as:

\[
d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

It is the most common metric for numerical data [5].

2.3. Manhattan Distance

The Manhattan distance between two data points is defined by the absolute value of sum of the differences of the all co-ordinates of the data points. For \( x, y \) in \( n \)-dimensional space, the Manhattan distance, \( d \) is defined as:
This is also called taxicab distance \[6\].

3. k-Medoids

Clustering is an unsupervised learning, where the objects of similar patterns are grouped together to form clusters. The clusters obtained can be used to interpret and group similar objects. k-Medoids is a partition approach, in which the objects are partitioned into k clusters. The initial medoid is chosen randomly. The clusters are based on minimum distance. The centroid in k-means is the mean of the objects within the cluster whereas in k-medoids the centroid is taken as an object itself. In general, k-Medoids is used to cluster the types of data in which the mean of objects is not defined or unavailable \[7\]. The flowchart of k-medoids algorithm is shown below in figure 2.

![Flowchart of k-medoids algorithm](image)

4. Dunn Index

The Dunn Index (DI) is a validation index used to analyze the compactness of the cluster obtained from the clustering task. For any cluster partition V of the dataset X with k number of partition (namely \(C_1 \ldots C_k\)), the Dunn index is given by,

\[
D(V) = \min_{1 \leq i \neq j \leq k} \frac{\delta(C_i, C_j)}{\max_{l \in [1,k]} \Delta_l}
\]

where \(\delta(C_i, C_j)\) is the distance between the two different cluster \(C_i, C_j\) (inter-cluster distance) and \(\Delta_l\) is the maximum intra-cluster distance of the k-th cluster \[8\]. In general, a higher value of Dunn index suggests a better cluster quality.

5. Proposed Method

Clustering of text document is a challenging task. Distance measure plays a major role in clustering algorithm. Though, geometric distance measure like Euclidean, Manhattan are handy, the clusters obtained are based on the structure produced by the compactness of the cluster. In this paper, we have proposed a similarity measure based on term cardinality of the document. After pre-processing the documents, the terms of each document is represented as a set and the proposed similarity measure makes use of the cardinality of the document. The cardinality of the document is the number of terms involved in the document.

The similarity measure between each pair of documents A and B is given by,

\[
\text{Sim}(A, B) = \frac{C_A \cdot C_B}{C_{A \cup B}^2}
\]

where \(C_A\) is the cardinality of terms in document A and \(C_{A \cup B}\) is the cardinality of the union of terms in document A and document B.

Using the above similarity measure the distance between two documents A and B is given by
\[ \text{dist}(A, B) = 1 - \text{Sim}(A, B) = 1 - \frac{C_A \cdot C_B}{C_{A \cup B}^2} \]

The proposed distance measure satisfies the three conditions of metric.

1. As \( C_{A \cup B}^2 \geq C_A \cdot C_B \) and \( C_{A \cup B} > 0 \), for all \( A, B \) it follows that \( \text{dist}(A, B) \geq 0 \). Hence axiom 1 is satisfied.

2. Notice that, \( \text{Sim}(A, A) = 1 \), as the cardinality of union of same documents equals to \( A \). Thus, \( \text{dist}(A, B) = 0 \) whenever \( A = B \). Hence axiom 2 is satisfied.

3. It is clear that \( \text{Sim}(A, B) = \text{Sim}(B, A) \). It follows that \( \text{dist}(A, B) = \text{dist}(B, A) \). Hence axiom 3 is satisfied.

4. Let \( A, B, C \) be any three documents. We shall consider the following cases.

\text{Case (i)}

Let, \( A \subseteq C \) and \( B \subseteq C \) then,

\[ \text{dist}(A, C) + \text{dist}(C, B) = \left(1 - \frac{C_A \cdot C_C}{C_{A \cup C}^2}\right) + \left(1 - \frac{C_C \cdot C_B}{C_{C \cup B}^2}\right) \]

\[ \geq 2 - \left(\frac{C_A \cdot C_{A \cup B} + C_{A \cup B} \cdot C_B}{C_{A \cup B}^2}\right) \quad \text{(Since} \ C_{A \cup B} \leq C_C \text{)} \]

\[ \geq 2 - 2 \left(\frac{C_A \cdot C_B}{C_{A \cup B}^2}\right) \]

\[ = 2 \left(1 - \frac{C_A \cdot C_B}{C_{A \cup B}^2}\right) = 2 \cdot \text{dist}(A, B) \geq \text{dist}(A, B) \]

\text{Case (ii)}

Suppose, \( C \subseteq A \) and \( C \subseteq B \) then for any elements in \( A \) or \( B \) but not in \( C \) will only increase \( \text{dist}(A, C) + \text{dist}(C, B) \), but not \( \text{dist}(A, B) \). If \( A = B = C \), then \( 0 + 0 \geq 0 \). Thus, the triangular inequality holds for all cases.

Hence, our proposed measure is a metric.

Thus, the similarity measure between two documents is given by the ratio of product of cardinalities of terms to the square of cardinality of the union of terms.

6. Dataset

The documents used for clustering task are downloaded from the BBC sport website [9]. The number of documents in the BBC sports dataset is 737 which come under five classes namely athletics, cricket, football, rugby, tennis. This document was preprocessed and reduced to a document term matrix which consists of 2042 terms.

7. Results and Discussion

In this paper, we have illustrated the proposed measure with a toy dataset which consist of 6 documents and 8 terms [10]. The raw count is tabulated below in Table 1.

| cluster | cluto | Distance document | evaluate | high measure | use |
|---------|-------|-------------------|----------|--------------|-----|
| D1      | 1     | 0                 | 0        | 0            | 0   |
| D2      | 1     | 0                 | 0        | 0            | 0   |
| D3      | 1     | 0                 | 1        | 0            | 0   |
| D4      | 1     | 0                 | 2        | 0            | 0   |
| D5      | 1     | 0                 | 1        | 0            | 0   |
| D6      | 0     | 0                 | 1        | 0            | 1   |
As an example, \( D_1 = \{ \text{cluster, cluto, high} \} \) and \( D_2 = \{ \text{cluster, cluto} \} \).

Thus \( D_1 \cup D_2 = \{ \text{cluster, cluto, high} \} \).

Here, \( C_{D_1} = 3 \), \( C_{D_2} = 2 \) and \( C_{D_1 \cup D_2} = 3 \)

Thus, \( \text{Sim}(D_1, D_2) = \frac{3 \times 2}{3^2} = \frac{2}{3} = 0.667 \).

The similarity matrix obtained for the above dataset is given in Table 2.

|     | D1    | D2    | D3    | D4    | D5    | D6    |
|-----|-------|-------|-------|-------|-------|-------|
| D1  | 1     | 0.667 | 0.562 | 0.36  | 0.306 | 0.234 |
| D2  | 0.667 | 1     | 0.375 | 0.5   | 0.278 | 0.204 |
| D3  | 0.562 | 0.375 | 1     | 0.5625| 0.4167| 0.306 |
| D4  | 0.36  | 0.5   | 0.5625| 1     | 0.4167| 0.4167|
| D5  | 0.306 | 0.278 | 0.4167| 0.4167| 1     | 0.694 |
| D6  | 0.234 | 0.2041| 0.306 | 0.4167| 0.694 | 1     |

The k-medoids algorithm (for \( k=2, 3, 4 \)) employed with Euclidean, Manhattan and our proposed measure are used for clustering task. The clusters obtained are validated using Dunn Index. The Dunn Index obtained for Euclidean, Manhattan and our proposed measure is evaluated and it is shown in Table 3.

| Number of Cluster | Euclidean | Manhattan | Proposed Metric |
|------------------|-----------|-----------|-----------------|
| \( k=2 \)        | 0.6123    | 0.7       | 1               |
| \( k=3 \)        | 0.7071    | 0.8       | 1.0011          |

Further, the proposed measure is analyzed using documents downloaded from the BBC sport website. As the original document consist of 5 class labels, we have analyzed the results for \( k = 5 \) using k-Medoids algorithm. The clustering results are validated using Dunn Index. The value obtained using Dunn Index is given in Table 4.

| Number of Cluster | Euclidean | Manhattan | Proposed Metric |
|------------------|-----------|-----------|-----------------|
| \( k=5 \)        | 0.07952198| 0.04494382| \textbf{0.207456} |

From the above table, we can see that the cluster obtained from the proposed measure is of good quality when compared to both Euclidean and Manhattan.

8. Conclusion

In this paper, a novel distance measure for text document clustering is proposed. The BBC Sports dataset used is pre-processed and a document term matrix is obtained. This matrix is employed with benchmark algorithm k-medoids for clustering task. The results obtained from the proposed distance measure are compared with other standard distance measures like Manhattan, Euclidean distance measure. Dunn Index is used to analyze the cluster validation of the results obtained from the distance measure. We find the proposed measure perform well for sparse matrix. Our future work is
to extend this work for weighted terms and also to employ other clustering algorithm with the proposed measure.

References

[1] Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, John Wiley & Sons. Ltd, (2010).
[2] Ronen Feldman and James Sanger, The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, (2007).
[3] Charu Aggarwal, ChengXiang Zhai, Mining Text Data (Springer) Ed. (2012).
[4] Santini, S., Jain, R.: Similarity Measures. IEEE Trans. Pattern Analysis and Machine Intelligence 21(9), 871–883 (1999)
[5] Wang H, Wang H, Wang W, Wang W, Yang H, Yang H, et al. Clustering by pattern similarity in large data sets. 2002 ACM SIGMOD international conference on Management of Data. New York, New York, USA: ACM Press; 2002. p. 394.
[6] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.
[7] Saxena, Amit, et al. "A review of clustering techniques and developments." Neurocomputing 267 (2017): 664-681.
[8] Legány, Csaba, Sándor Juhász, and Attila Babos. "Cluster validity measurement techniques." Proceedings of the 5th WSEAS international conference on artificial intelligence, knowledge engineering and data bases. World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA, 2006.
[9] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.
[10] Grace, G. Hannah, and Kalyani Desikan. "Document clustering using a new similarity measure based on energy of a bipartite graph." Indian Journal of Science and Technology 9 (2010): 40.