Document Clustering using Feature Selection Based on Multiviewpoint and Link Similarity Measure

Neelam Singh
Graphic Era University
neelamjain.jain@gmail.com

Neha Garg
Graphic Era University
nehagarg.february@gmail.com

Janmejay Pant
Graphic Era University
tejraj2010@gmail.com

Abstract
Clustering is one of the very powerful and widely used technique in information retrieval. All clustering methods work on finding relationship among data objects. There are various similarity measures used along with criteria functions to find similarity between documents like cosine, jaccard etc. Clustering efficiency and performance is highly dependent on the accuracy of the similarity measure. In this paper we propose a new similarity measure based on link and multiviewpoint similarity to find the closeness of two documents.

Keywords
Clustering, similarity measure, cosine, jaccard, link, multiviewpoint.

1. Introduction
With the increased availability of IOT (Internet of Things) and techniques to capture and analyze data a major question arises as to how data can be explored with much greater efficiency and accuracy.

Various techniques and algorithms have been proposed in order to discover patterns of interests among a vast collection of data sets.

Document clustering or text clustering is one of the most important text mining methods used to help users to effectively search, parse, summarize, and organize text documents. Based on a user query document clustering aims to group documents into meaningful clusters from a given corpus, such that documents belonging to the same group or cluster have common topic.

There are two general categories of clustering methods: agglomerative hierarchical and partitional methods. Agglomerative hierarchical clustering (AHC) algorithms start by assuming each document as a cluster. The algorithm then use different kinds of distance functions to compute the similarity between the pairs of clusters, and then merge the closest pair [1]. This merging step is repeated until the desired number of clusters is created.

Partitional clustering algorithm like k-means works on a one-level partitioning of the documents. The k-means algorithm is based on the idea that a centroid can represent a cluster. These methods require a similarity measure to perform clustering.

A similarity function embedded in a criterion function is to a large extent is responsible to analyze the intrinsic structure of the data. If appropriate similarity measures are used with specific clustering technique the efficiency and accuracy of the technique can be enhanced [2]. As in k-means sum-of-squared error function with Euclidean distance is used while for document clustering cosine similarity measure is used.

Although there are number of similarity measures available yet more robust measures are still required. The work in this paper is based on the fact that the nature and selection of similarity measure is directly proportional to the success and failure of a clustering algorithm. Our work will focus on the use of neighbors and link with the multiviewpoint similarity measure[3], since in cosine function for a given cluster only the pairwise similarity is considered. But it could be possible that some documents in different clusters may have some features in common. Thus it can be considered that if two data points are similar in some respect they can be called neighbors and assuming a similarity threshold a given data point is called the link.

We evaluated the performance of our proposed similarity measure for document clustering with feature selection on various data sets.

The rest of the paper is organized as follows:
In section 2 we will discuss and review the related work. In section 3 document clustering approach with respect to the vector space model, the multiviewpoint similarity measure, the concepts of neighbors and link and feature selection techniques is discussed. In section 4 the proposed similarity measure with experimental results and comparison with other similarity measures has been discussed. In section 5 we will give conclusion and future scope of our work.

2. Related Work
Clustering aims to arrange based on maximum intracluster similarity and intercluster dissimilarity. All the clustering algorithm uses similarity measures (or dissimilarity measure) in order to cluster efficiently.
In most of the literature, Euclidean distance is considered to be most popular and effective.

\[ \text{Eucl}(d_i, d_j) = ||d_i - d_j|| \]

For document clustering cosine similarity measure is the most widely used metric.

\[ \text{Sim}(d_i, d_j) = \cos(d_i, d_j) = d_i^T d_j \]

Duc Thang Nguyen et al. [3] has proposed Clustering with Multiviewpoint-Based similarity measure where they had taken multiple view points to find similarities between documents. According to them the similarity between two points X and Y inside a cluster from a point C outside the cluster is the product of the cosine angle between X and Y from C and the corresponding Euclidean distance.

Congnan Luo et al. has given the concepts of neighbours and links along with cosine similarity measure for clustering [4]. According to them link between two documents shows the number of common neighbours of the two documents, thus will help to achieve stronger clustering with respect to other methods.

There are also text based similarities like phrase based and content related similarity measure which are also gaining popularity. Lakkaraju et al. used tree based similarity with reference to concepts in text [5]. Chim and Deng used suffix tree model and vector space model for finding phrase based similarity [6].

3. Background

3.1 Document Clustering

The problem of document clustering is generally defined as follows: given a set of documents, they are clustered in such a way that the documents in one cluster share the same feature and the documents in different clusters represent different features. This process relies on similarity measure to find clusters [7].

3.2 Vector Space Model

This is an algebraic model in which documents are considered as vectors and are represented by term frequency such that –

\[ \text{dtf} = [t_{f_1}, t_{f_2}, \ldots, t_{f_D}] \]

where

- \( t_{fi} \) – frequency of ith term in document
- \( D \) – total number of unique terms in text database.

3.3 Multiviewpoint Similarity Measure

According to Nguyen et al. similarity of two documents \( d_i \) and \( d_j \) belonging to same cluster is the average of similarity measured with respect to all other documents existing outside the cluster.

\[ MVS(d_i, d_j) = \frac{1}{n} \sum_{d_n \in S_r} (d_i - d_n)^T (d_j - d_n) \]

\( d_i, d_j \) represent vectors of the documents in the given corpus.
- \( n \) number of documents
- \( m \) number of terms
- \( S \) set of all documents
- \( S_r \) A document from the given corpus

The similarity between two points \( d_i \) and \( d_j \) inside cluster \( S_r \), viewed from a point \( d_n \) outside the cluster is the product of the cosine angle between \( d_i \) and \( d_j \) and Euclidean distances from \( d_n \) to these two points.

3.4 Link and Neighbors

In a data set the document similar to document \( d_i \) are said to be its neighbor. For finding neighbor of every document in a given data set neighbor matrix, \( M(nxn) \) is created as an adjacency matrix with respect to similarity such that:

\[ \text{sim}(d_i, d_j) \geq \mu \]

where

\[ 0 \leq \mu \leq 1 \]

and \( n \) is the number of documents

The value of link function \( \text{link}(d_i, d_j) \) is the number of common neighbors between \( d_i \) and \( d_j \) and can be calculated as-

\[ \text{link}(d_i, c_j) = \sum_{m=1}^{n} M[m, i] \cdot M[m, n+j] \]

where

- \( d_i \) represents document
- \( c_j \) represents cluster
- \( M[m,n] \) represents a neighbor matrix
- \( i \) represents rows
- \( j \) represents columns

3.5 Feature Selection

It is also known as variable selection where from a given collection of documents subset of relevant attributes is selected. Various algorithms for feature selection are often categorized as wrapper, filter or embedded approach [8]. We will be deploying sequential feature selection method in order to find relevant features required for document clustering.

4. Document Clustering with link based multiviewpoint similarity measure (MVS)

4.1 Link based MVS measure

For document or text clustering cosine similarity still remains as the most popular measure because of its simple mechanism and easy computation, though its effectiveness lessens when synonyms are used across documents, also the number of unique terms might vary
in different documents making vocabulary size quite large. Although MVS finds similarity between documents based on multiple view with respect to documents outside cluster but using link function the knowledge of neighbor documents can be used that can more precisely will find the closeness of documents as the cluster will not split because of large vocabulary. So the proposed similarity measure for link based MVS measure can be defined as

\[ f(d_i,d_j) = \alpha \cdot \frac{\text{link}(d_i,d_j)}{\max_{d_k,d_l \in S_i} \text{MVS}(d_k,d_l)} + (1 - \alpha) \cdot \text{MVS}(d_i,d_j) \]

Where \( \alpha \) is a threshold value such that \( 0 \leq \alpha \leq 1 \)

### 4.2 Algorithm

Given a set of N documents to be clustered, the process of clustering involves the following steps-

1. **STEP 1** – Create an NXN neighbor matrix based on similarity.
2. **STEP 2** – Find the link similarity measure using the neighbor matrix and also find the maximum link value.
3. **STEP 3** – Using the values finally calculates the similarity based on the multiviewpoint and link similarity measure.
4. **STEP 4** – Start by assigning each document to a cluster, such that for N items will always be N clusters based on similarities.
5. **STEP 5** – With the help of tfidf reassign cluster by finding the most similar pair of clusters and merge the similar cluster.
6. **STEP 6** – Compute similarities between the new cluster and each of the old clusters.
7. **STEP 7** – Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

### 4.3 Results and Comparative analysis

Both multiviewpoint and link similarity measure are used to calculate the similarities between documents, we integrated both of them to evaluate the similarities in order to form clusters.

Let us consider a data set S containing 6 documents \{d1, d2, d3, d4, d5, d6\}. The neighbor matrix for the given documents thus created is-

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>d6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Based on this matrix link matrix is formed which tells the number of neighbors a given document can have.

All these values are then assigned to the Link based MVS measure to calculate document similarities. The following table shows a comparison of Link based MVS measure with some other similarity measure by selecting four documents with highest link values.

<table>
<thead>
<tr>
<th>di,dj</th>
<th>Cosine Multiviewpoint (MVS)</th>
<th>Cosine Link</th>
<th>MVS Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1,d2</td>
<td>.25</td>
<td>.16</td>
<td>.18</td>
</tr>
<tr>
<td>d1,d3</td>
<td>.23</td>
<td>.07</td>
<td>.16</td>
</tr>
<tr>
<td>d1,d4</td>
<td>.2</td>
<td>.11</td>
<td>.14</td>
</tr>
<tr>
<td>d2,d3</td>
<td>.75</td>
<td>.51</td>
<td>.82</td>
</tr>
<tr>
<td>d2,d4</td>
<td>.2</td>
<td>.11</td>
<td>.15</td>
</tr>
</tbody>
</table>

Using the similarity measure clustering is performed and the steps are repeated till the process cannot be repeated further i.e. all the documents are assigned to a respective cluster.

### 4.4 Evaluation and Experimental setup

It is very difficult comparing the impact of similarity metrics on cluster quality, because evaluating cluster quality is difficult in itself. We had used the coephenetic correlation coefficient for finding how well our proposed similarity measure helps to evaluate the similarities within a data set. The coephenetic correlation coefficient can be defined as the linear correlation coefficient between the cophenetic distances obtained from the tree, and the original distances (or dissimilarities) used to construct the tree. Thus, it is a measure of how precisely the tree represents the dissimilarities among observations [9].

The following dendrograms show the formation of clusters using different similarity measure using hierarchical clustering. We had used the hierarchical clustering [10] along with various similarity measures like Euclidean Distance, Cosine similarity measure, Multiviewpoint similarity measure and link based multiviewpoint similarity measure to create tree and the respective dendrograms and clusters.
The table shows the Coephenetic Correlation Coefficient of various similarity measures with respect to dendrogram formation. The cophenetic distance between two observations is represented in a dendrogram by the height of the link at which those two observations are first joined. That height is the distance between the two sub clusters that are merged by that link. The magnitude of this value should be very close to 1 for a high-quality solution.

Table 3 Coephenetic Correlation Coefficient

<table>
<thead>
<tr>
<th></th>
<th>Euclidean</th>
<th>Cosine</th>
<th>Multiviewpoint</th>
<th>MVSlink</th>
</tr>
</thead>
<tbody>
<tr>
<td>di,dj</td>
<td>0.833</td>
<td>0.1224</td>
<td>0.8</td>
<td>0.88</td>
</tr>
</tbody>
</table>

So based on this we came to the finding that MVSlink shows a high solution with respect to other similarity measure for finding similarities between document pair using a clustering algorithm.

5. Conclusion

In this paper, we propose a Link based Multiviewpoint similarity (MVSlink) measuring method for document clustering. Theoretical analysis and experimental evaluation shows that MVSlink is potentially more suitable for text documents as compared to other methods like cosine similarity measure.

The key contribution of the paper is to calculate similarity based on the concept of neighbors along with multiviewpoint. Future scope of this method will be on how the measure will work on sparse and high-dimensional data.

6. References


