Hierarchical Filter based Document Clustering Algorithm

Mulluri Raghupathi, R. Lakshmi Tulasi
M.Tech Student, 2Professor & Head, Dept. of CSE
QIS Institute of Technology, Ongole

ABSTRACT:
Clustering is one of the major important tasks in data mining. The task of clustering is to find the fundamental structures in data and categorize them into meaningful subgroups for supplementary study and examination. Existing K-Means clustering with MVS measure it doesn’t best position to cluster the data points. This problem will lead to gain less optimal solution for clustering method. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal. In this proposed approach, multiview clustering is applied on different applications namely on text documents and real-time document clustering on local disks. Proposed approach gives better clustering accuracy in terms of different sizes of data.

1. INTRODUCTION
Data mining is that the method of extracting or mining information from great deal of information. It’s Associate in analytic method designed to explore giant amounts of information in search of consistent patterns and systematic relationships between variables and to validate the findings by the detected patterns to new subsets of information. It is often viewed as a result of natural evolution in development of Functionalities like data assortment, information creation, information management, information analysis. It is the process where intelligent methods are applied in order to extract data patterns from databases, data warehouses, or other information repositories. Clusters are often thought of the foremost necessary unsupervised learning problem, thus as Each different drawback of this sort, it deals with finding a structure in an exceedingly assortment of unlabelled information. A cluster is a set of objects that are coherent internally, however clearly dissimilar to the objects to different clusters. A loose definition of cluster could also be “the methodology of organizing objects into groups whose members’ are similar in some way”.

Fig1 Document Clustering Representation

In the existing system, clustering is one among that interesting and very important topic in data mining. The aim of clustering is to hunt out intrinsic structures in data, and organize them into meaningful subgroups for more study and analysis. There are clustering algorithms published every year. They can be proposed for very distinct analysis fields, and developed using totally different techniques and approaches. This paper proposed a Multiviewpoint-based Similarity measuring method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. Based on MVS, two criterion functions, IR and IV, and their respective clustering algorithms, MVSCIR and MVSC-IV, have been introduced [1]. A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity (or distance) among data. Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. Hence, effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to the data at hand. For instance, the original k-means has sum-of-squared-error objective function that uses Euclidean distance. [8] In a very sparse and high-dimensional
domain like text documents, spherical k-means, that uses cosine similarity (CS) rather than Euclidean distance as the measure, is deemed to be more suitable.

2. Related Work

The main distinctness of our concept with a traditional dissimilarity/similarity measure is that the aforementioned dissimilarity/similarity exercises only a single view point for which it is the base and where as the mentioned Clustering with Multiviewpoint-Based Similarity Measure uses many different viewpoints that are objects and are assumed to not be in the same cluster with two objects being measured. By utilizing multiple viewpoints, countless descriptive evaluation could be accomplished. In order to assist this declaration, the theoretical analysis and empirical study are carried. Depending on this new measure two criterion functions are proposed for document clustering.

Document clustering is one of the text mining techniques. It has been around since the inception of text mining domain. It is a process of grouping objects into some categories or groups in such a way that there is maximization of intra-cluster object similarity and inter-cluster dissimilarity. Here an object does mean a document and term refers to a word in the document. Each document considered for clustering is represented as an \( m \) – dimensional vector \( d \). The \( m \) represents the total number of terms present in the given document. Document vectors are the result of some sort of weighting schemes like TF-IDF (Term Frequency – Inverse Document Frequency). Many approaches came into existence for document clustering. They include information theoretic co-clustering, non – negative matrix factorization, and probabilistic model based method and so on. However, these approaches did not use specific measure in finding document similarity. In this paper we consider methods that specifically use certain measurement. From the literature it is found that one of the popular measures is Euclidean distance.

3. PROPOSED MODEL

Document clustering can loosely be defined as “clustering of documents”. Clustering is a process of understanding the similarity and/or dissimilarity between the given objects and thus, dividing them into meaningful subgroups sharing common characteristics. Good clusters are those in which the members inside the cluster have quite a deal of similar characteristics.

**Input:** Friends, Romans, Countrymen, lend me your ears;  
**Output:** Friends, Romans, Countrymen, lend me your ears

**Tokenization** is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.
by constructing a feature vector for each new phrase then a similarity measure is recursively calculated between each new phrase and the phrases that already have been added to the feature vector and when the similarity exceeds a threshold (assumed value); then one of them will be discarded.

After obtaining the term weights of all topic phrases, it is easy to apply the cosine similarity to compute the similarity of any two documents. Let vectors \( d_x = \{x_1, x_2, \ldots, x_M\} \) and \( d_y = \{y_1, y_2, \ldots, y_M\} \) denote two documents \( d_x \) and \( d_y \), where \( x_i \) and \( y_i \) are the weights of corresponding topic phrase term. Then the similarity of two documents in (1) is calculated by the following formula:

\[
\text{Sim}(x, y) = 1 - \frac{\text{vector1}.\text{innerProduct}(\text{vector2})}{\text{vector1}.\text{norm}() / \text{vector2}.\text{norm}()};
\]

PROBABILITY CALCULATION:

Hierarchical Document CLUSTERING ALGORITHM:

1. Start
2. Assign \( \sqrt{n} \) objects as initial for next step
3. Get initial clusters using HAC algorithm
4. Make these clusters as initial \( k \) document clustering
5. Assign Cluster centers randomly
6. For each object in document collection
   a. If new object
      then check the nearest mean,
      join the cluster with nearest
      mean,
      calculate the new mean for the
      cluster
      end if
   end for

Figure 3: Document representation

Figure 4: Hierarchical clustering algorithm

PSEUDO CODE:

Input: Documents
Procedure:
  int numDocuments = st.countTokens() - 1;
  String record = st.nextToken
  for each document \( i \) in numDocuments
    then record = st.nextToken();
    Document createDocument(record);
    if (document != null)
      then documents.add(document);
end if
end for

4. EXPERIMENTAL RESULTS

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operation system platform is Microsoft Windows XP Professional (SP2).

Experimental results:

![Input Clustering Threshold](image)

size of docs :6
Distance Similarity Measure : : NaN

====================

Sum of Squared Distances ==> ( Document: 0, Title: Google's proposed acquisition of Motorola Mobility getting closer look in European review ) 0.9421438700044162
Document: 1, Title: Google, Motorola merger review in Europe delayed
Document: 2, Title: Google launches political hub for 2012 elections
Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses
Document: 4, Title: Nasa's gravity twins now circling Moon
Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core

Distance Similarity Measure : : 0.9421438700044162

====================

Sum of Squared Distances ==> ( Document: 4, Title: Nasa's gravity twins now circling Moon ) 1.9170459194988676
Document: 0, Title: Google's proposed acquisition of Motorola Mobility getting closer look in European review
Document: 1, Title: Google, Motorola merger review in Europe delayed
Document: 2, Title: Google launches political hub for 2012 elections
Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses
Document: 4, Title: Nasa's gravity twins now circling Moon
Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core

Distance Similarity Measure : : 0.9421438700044162

====================

Sum of Squared Distances ==> ( Document: 2, Title: Google launches political hub for 2012 elections ) 3.821521252059784
Document: 0, Title: Google's proposed acquisition of Motorola Mobility getting closer look in European review
Document: 1, Title: Google, Motorola merger review in Europe delayed
Document: 2, Title: Google launches political hub for 2012 elections
Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses
Document: 4, Title: Nasa's gravity twins now circling Moon
Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core

Distance Similarity Measure : : 0.9421438700044162

====================

Sum of Squared Distances ==> ( Document: 0, Title: Google's proposed acquisition of Motorola Mobility getting closer look in European review ) 5.738367219261234
Document: 1, Title: Google, Motorola merger review in Europe delayed
Document: 2, Title: Google launches political hub for 2012 elections
Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses
Document: 4, Title: Nasa's gravity twins now circling Moon
Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core

Distance Similarity Measure : : 0.9421438700044162

Distance Similarity Measure : : 0.9421438700044162

Distance Similarity Measure : : 0.9421438700044162

Distance Similarity Measure : : 0.9421438700044162

Distance Similarity Measure : : 0.9421438700044162

Distance Similarity Measure : : 0.9421438700044162
Motorola Mobility getting closer look in European review

Document: 1, Title: Google, Motorola merger review in Europe delayed

Sum of Squared Distances ==> ( Document: 2, Title: Google launches political hub for 2012 elections

Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses

Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core

Sum of Squared Distances ==> ( Document: 2, Title: Google launches political hub for 2012 elections

Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses

Distance Similarity Measure : : 0.9374284435896368

Cluster 0

Document: 0, Title: Google's proposed acquisition of Motorola Mobility getting closer look in European review

Cluster 1

Document: 2, Title: Google launches political hub for 2012 elections

Document: 3, Title: Google launches U.S. election site in time for Iowa caucuses

Cluster 2

Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core

Distance Similarity Measure : : 0.9374284435896368

Cluster 3

Document: 5, Title: Twin NASA Probes Circling Moon, Hoping to Answer Questions About Core
5. CONCLUSION

The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints. Theoretical analysis show that Multi-viewpoint based similarity measure (MVS) is potentially more suitable for text documents than the popular cosine similarity measure. The future methods could make use of the same principle, but define alternative forms for the relative similarity in or do not use average but have other methods to combine the relative similarities according to the different viewpoints. In future, it would also be possible to apply the proposed criterion functions for hierarchical clustering algorithms. It would be interesting to explore how they work types of sparse and high-dimensional data.

6. REFERENCES

Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229, Sept. 2008.


AUTHORS

Raghupathi.M is a student of Computer Science Engineering from QIS Institute of Technology, Presently pursuing M.Tech (CSE) from this college. She received B.Tech from JNTUK in the year of 2012.

Lakshmi Tulasi.R is a Professor, H.O.D of QIS Institute of Technology, Ongole. She received M.Tech from JNTUCEA. she is pursuing Ph.D at JNTUH. She is a good Researcher in semantic web, Computer Networks. She attended Various National and International Workshops and Conferences.