Comparison of Document Clustering Techniques

Sandip A. Patil[1],Kanchan D. Patil[2] ,Apurva B. Guldagad [3]
Lecturer,Computer Technology Department[1],Assistant Professor of Computer Engineering Department[2],
Student, Computer Technology Department[3]
Sanjivani K.B.P Polytechnic,Kopargaon, Maharashtra, India[1], Sanjivani College of Engineering ,Kopargaon, Maharashtra, India[2],Pirens Institute of Computer Technology,Loni[3]

***Abstract*—**
As web is growing larger day by day and with the growth of social networking, the documents collected from the web are becoming more and more condensed. Such data due to the sparseness imposes new challenges in applying clustering techniques. Clustering such sparse data could lead to new trends in web search and other such applications.

Clustering and classification are the fundamental tasks in Data Mining which is used to retrieve only meaningful data from the available raw data.

Clustering is an unsupervised learning; the goal of clustering is descriptive. Since the goal of clustering is to group the similar objects together, the new groups are of interest in themselves, and their assessment is intrinsic.

Classification is a supervised learning method which is predictive. In classification tasks, the assessment is extrinsic, since the groups must reflect some reference set of classes.

Keywords—Data Mining, Clustering, TFIDF, Classification, Similarity Measure, Distance Matrix

I. Introduction

Document Clustering collects data instances or objects into groups in such a manner that similar instances are grouped together, while different instances belong to different groups.

The instances are thereby organized into an efficient representation that characterizes the population being sampled. Consequently, any instance in a group belongs to exactly one and only one subset.

What will be the natural grouping among these objects?

Clustering is subjective. It may group into family, school group , male, female groups.

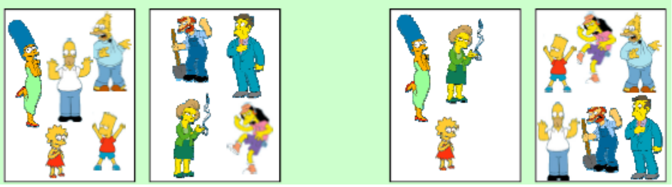


Fig. 1: Clustering

Since clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required.

What is similarity?

Fig.2: Similarity Measure

Two methods are used to estimate this relation:

- Distance Measures
- Similarity Measure

Distance measures are used to determine the similarity or dissimilarity between any pair of objects. It denotes the distance between two objects i and *j* as:

$$d\,(i, j).$$

A symmetric distance measure obtains its minimum value (usually zero) in case of identical objects.

Similarity Measure, an alternative to the distance measure, denoted as:

s (i; j) that compares the two vectors i and j

This function should be symmetrical:

s (i; j) = s (j ; i) and have a large value when i and j are somehow “similar” and constitute the largest value for identical vectors.

The most commonly used similarity measure is Cosine similarity.

II. LITERATURE SURVEY

Document clustering is becoming more and more important with the abundance of text documents available through World Wide Web and corporate document management systems. But there are still some major drawbacks in the existing text clustering techniques that greatly affect their practical applicability.

The drawbacks in the existing clustering approaches are listed below:

- Text clustering that yields a clear cut output has got to be the most favorable. However, documents can be regarded differently by people with different needs vis-à-vis the clustering of texts. For example, a businessman looks at business documents not in the same way as a technologist sees them (Macskassy *et al.* 1998). So clustering tasks depend on intrinsic parameters that make way for a diversity of views.

- Text clustering is a clustering task in a high-dimensional space, where each word is seen as an important attribute for a text. Empirical and mathematical analysis have revealed that clustering in high-dimensional spaces is very complex, as every data point is likely to have the same distance from all the other data points (Beyer *et al.* 1999).
- Text clustering is often useless, unless it is integrated with reason for particular texts are grouped into a particular cluster. It means that one output preferred from clustering in practical settings is the explanation why a particular cluster result was created rather than the result itself. One usual technique for producing explanations is the learning of rules based on the cluster results.

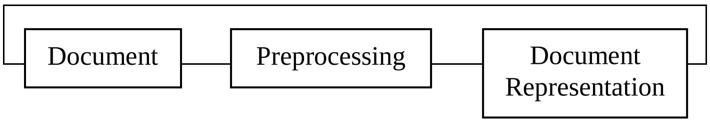
▪ Desirable Properties of a Clustering Algorithm:

- ✓ Scalability (in terms of both time and space)
- ✓ Ability to deal with different data types
- ✓ Minimal requirements for domain knowledge to determine
- ✓ input parameters
- ✓ Able to deal with noise and outliers
- ✓ Insensitive to order of input records
- ✓ Incorporation of user-specified constraints
- ✓ Interpretability and usability

III. EXISTING CLUSTERING TECHNIQUES

Document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters. Clustering of text documents plays a vital role in efficient Document Organization, Summarization, Topic Extraction and Information Retrieval. Initially used for improving the precision or recall in an Information Retrieval System .more recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to user's query or help users quickly identify and focus on the relevant set of results. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting.

Document clustering goes through following phases:



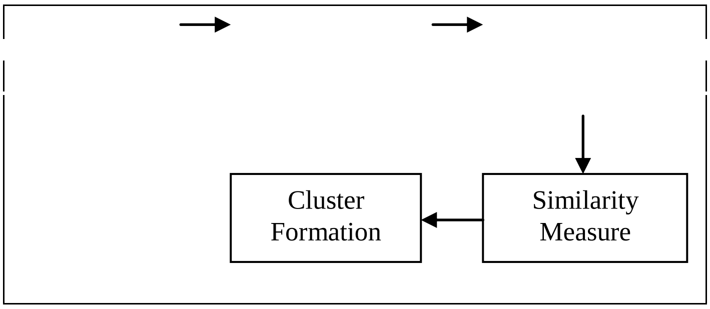


Fig. 3: Basic Steps of Document Clustering

A) Preprocessing:

Document preprocessing is used to clean the outliers from the document. Preprocessing of document include following steps:

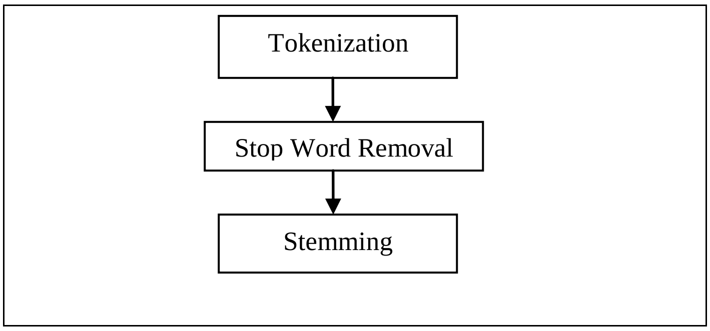


Fig. 4: Tasks in Document Preprocessing

- Tokenization:
  A document or bag of word is partitioned into number of words or tokens.
- Stop Word Removal:
  All common word from document are removed.
  e.g. a, an, the, or, &, etc.
- Stemming:
  This step is the process of conflating tokens to their root form.
  e.g. Words: Working, Works, Worked will be stemmed to word: work.

B) Document Representation:

A bag of word method is used in IR. Typically, document is represented as a feature vector. Frequency (number of times a word has appeared in document) of a term (word in a bag) is used as a weight.
Let D = {d1. . . dn} be a set of documents and T = {t1, . .tm} the set of distinct terms occurring in D. A document is then represented as a m-dimensional vector t_d.
Let tf(d, t) denote the frequency of term t € T in document d € D. Then the vector representation of a document d is

td = (tf(d, t1), . . . , tf(d, tm))

Computing Term weights/TFIDF Analysis:

By taking into account these two factors: term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically. Put another way a search result's score Ranking is the product of TF and IDF:

TFIDF = TF * IDF where:

- TF = C / T where C = number of times a given word appears in a document and T = total number of words in a document.
- Document IDF = D / DF where D = total number of documents in a corpus, and DF = total number of documents containing a given word.

C) Similarity Measure:

A similarity/distance measure must be determined before clustering. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. The nature of similarity measure plays a very important role in the success or failure of a clustering method.

Cosine similarity is one of the most popular similarity measure practical to text documents, such as in various information retrieval applications and clustering too. An important property of the cosine similarity is its independence of document length. For two documents di and dj, the similarity between them can be calculated as:

$$\mathrm{Cos}(d_i, d_j) = \frac{d_i \cdot d_j}{\| d_i \| \, \| d_j \|}$$

Since the document vectors are of unit length, the above equation is simplified to:

$$\mathrm{Cos}(d_i, d_j) = d_i \cdot d_j$$

When the cosine value is 1 the two documents are identical, and 0 if there is nothing in common between them (i.e., their document vectors are orthogonal to each other).

As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. These existing approaches form the basis for the various innovations in the field of clustering. From the existing clustering techniques, it is clearly observed that the clustering techniques provide significant results and performance.

Common clustering techniques come under two broad categories:

1. Partitional:

Construct various partitions and then evaluate them by some criterion.

Fig. 5: Partitional Clustering Technique

Nonhierarchical, each instance is placed in exactly one of K non-overlapping clusters. Since only one set of clusters is output, the user normally has to input the desired number of Clusters K.

i) Decide K, and initialize K centers (randomly)

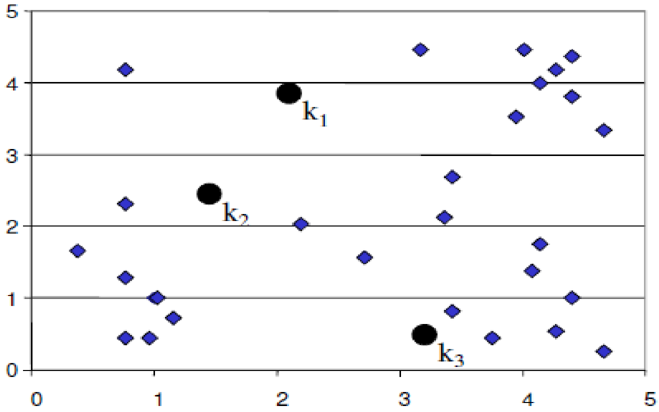


Fig. 6: Decide K and centers

ii) Assign all objects to the nearest center and move a center to the mean of its members.

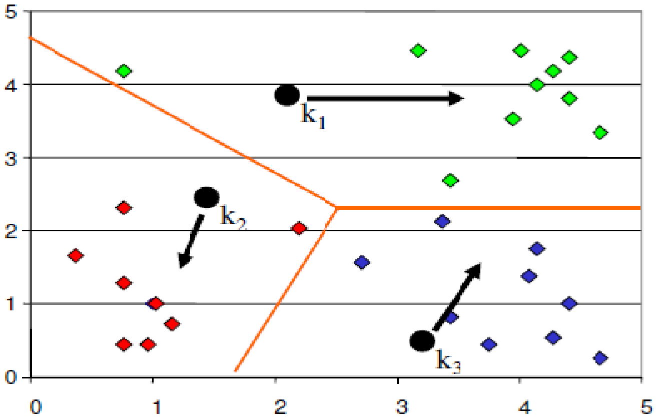


Fig.7: Assigning objects

iii) After moving centers, re-assign the objects…

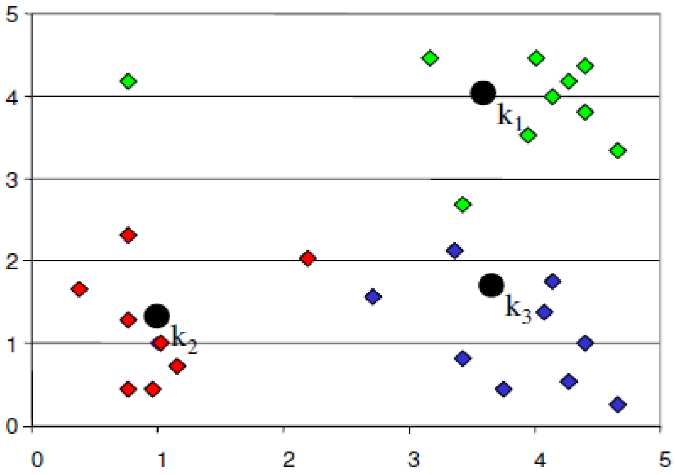


Fig.8: Reassigning objects

iv) After moving centers, re-assign the objects to nearest centers. Move a center to the mean of its new members.

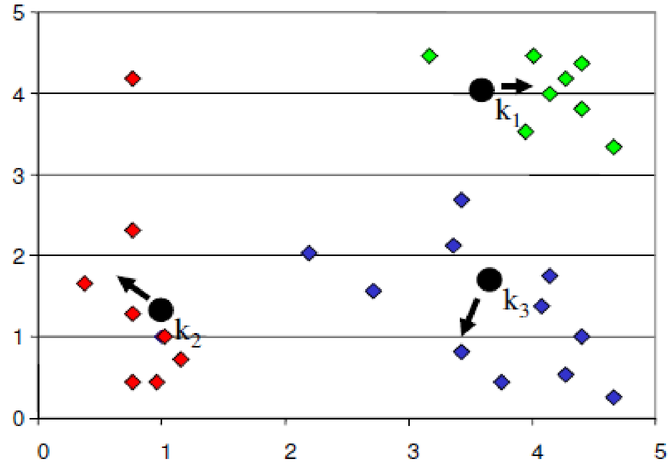


Fig.9: Move center

v) Re-assign and move centers, until no objects changed membership.

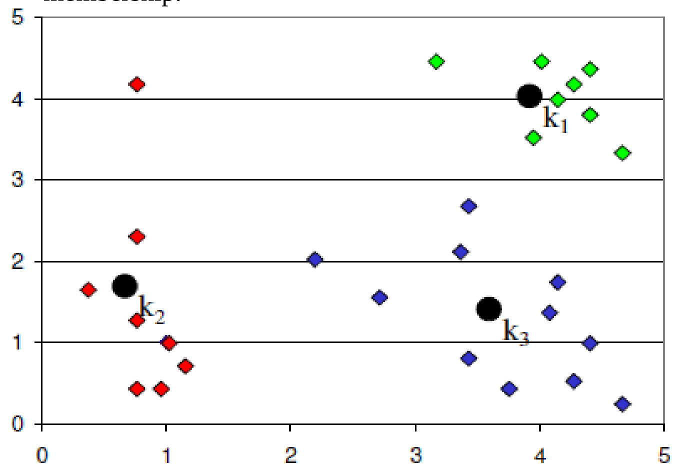


Fig.10: Move center

- Algorithm k-means

  1. Decide on a value for K, the number of clusters.
  2. Initialize the K cluster centers (randomly, if necessary).
  3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.
  4. Re-estimate the K cluster centers, by assuming the memberships found above are correct.
  5. Repeat 3 and 4 until none of the N objects changed membership in the last iteration.

- Comments on K-Means:

Strength

✓ Simple, easy to implement and debug

✓ Intuitive objective function: optimizes intra-cluster similarity

✓ Relatively efficient: O(tkn), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.

Weakness

✓ Applicable only when mean is defined, then what about categorical data?

✓ Often terminates at a local optimum. Initialization is important.

✓ Need to specify K, the number of clusters, in advance

✓ Unable to handle noisy data and outliers

✓ Not suitable to discover clusters with non-convex shapes.

2. Hierarchical:

Create a hierarchical decomposition of the set of objects using some criterion.

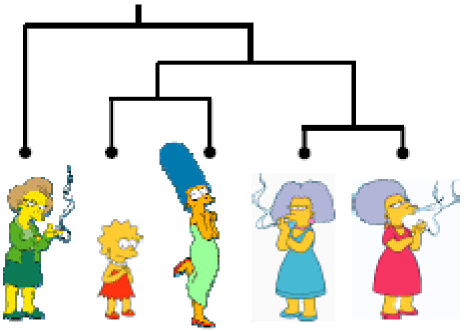
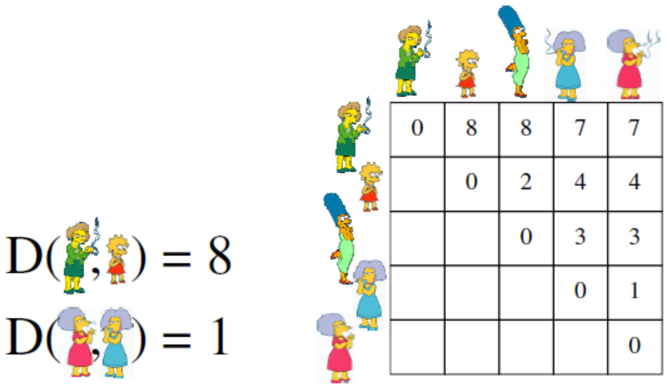


Fig. 11: Hierarchical Clustering Technique

- Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.
- Top-Down (divisive): Starting with all the data in a single cluster, consider every possible way to divide the cluster into two. Choose the best division and recursively operate on both sides.

Consider e.g. Bottom-Up (agglomerative) Clustering:

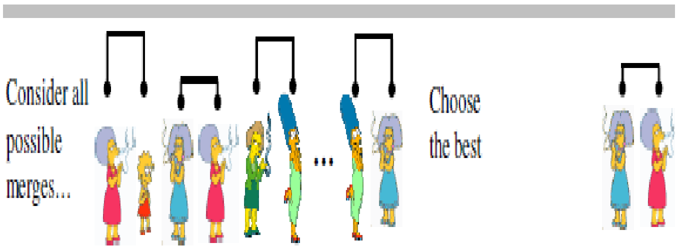
i) Begin with a distance matrix which contains the distances between every pair of objects in our database.



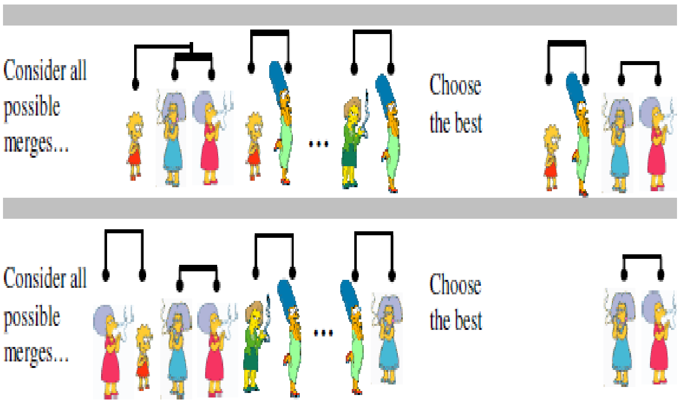
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

Fig. 12: Distance Matrix

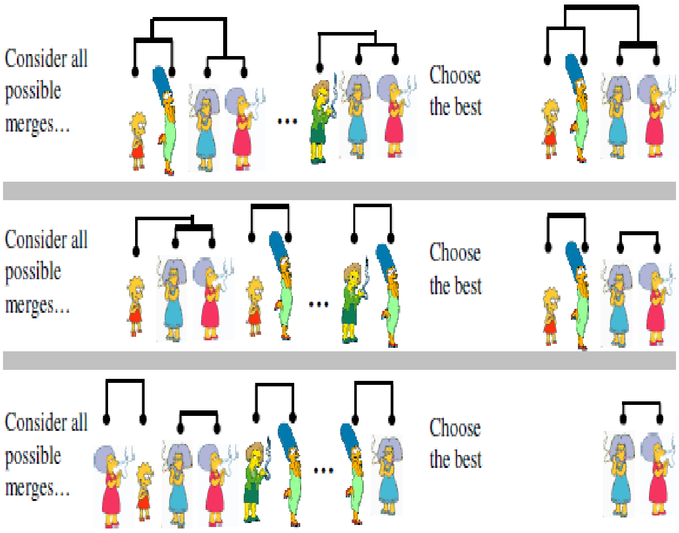
ii) Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



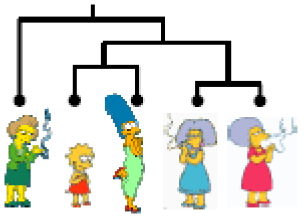
ii) Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.



iii) Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.


Consider all possible merges…
…
Choose the best
Consider all possible merges…
…
Choose the best
Consider all possible merges…
…
Choose the best

iv) Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together



- Summary of Hierarchal Clustering Methods:
  - ✓No need to specify the number of clusters in advance.
  - ✓Hierarchical structure maps nicely onto human intuition for some domains.
  - ✓They do not scale well: time complexity of at least O(n2), where n is the number of total objects.
  - ✓Like any heuristic search algorithms, local optima are a problem.
  - ✓Interpretation of results is (very) subjective.

REFERENCES

[1] "A Comparison of Document Clustering Techniques" by Michael Steinbach George Karypis Vipin Kumar , Department of Computer Science / Army HPC, Research Center, University of Minnesota.

[2] "Document Clustering on Various Similarity Measures" by Ms.K.Sruthi Mr.B.Venkateshwar Reddyin 2013, IJARCSSE.

[3] "A Survey of Document Clustering Techniques & Comparison of LDA and moVMF" by Yu Xiao December 10, 2010.

[4] "Similarity Measures for Text Document Clustering" by Anna Huang in NZCSRSC 2008, April 2008, Christchurch, New Zealand.

[5] "Comparative Study of Clustering Techniques for Short Text Documents" by Aniket Rangrej ,Sayali Kulkarni,Ashish V. Tendulkar in WWW 2011, March 28–April 1, 2011, Hyderabad, India.ACM 978-1-4503-0637-9/11/03.