Clustering Software Methods and Comparison

Rachana Kamble*
CSE Dept.,
TIT, Bhopal
rachanakamble@gmail.com

Mehajabi Sayeeda
CSE Dept.,
TIT, Bhopal
Mejhabi786@gmail.com

Abstract

Document clustering as associate not supervised approach extensively won’t to navigate, filter, summarize and manage huge group of document repositories just like the World Wide Web (WWW). Recently, Document clustering is that the method of segmenting a selected group of texts into subgroups as well as content based mostly similar ones. the aim of document clustering is to fulfill human interests in info looking and understanding. element based mostly software system development has gained lots of sensible importance within the field of system engineering from educational researchers and additionally from business perspective. Finding parts for economical code utilize is one among the necessary issues aimed by researchers. Clump reduces the search area of parts by grouping similar entities along so guaranteeing reduced time complexity because it reduces the search time for part retrieval. This work can study the key challenges of the clustering drawback, because it applies to the text domain. Additionally can discuss the key ways used for text clustering, and their relative benefits.

1. Introduction

As growth of internet application makes today’s life very easy. Everyone likes to search information using internet. To provides user a good navigation and browsing facilities. For this reason document clustering is good technique to provide searching document easy access. Hence clustering schemes are widely studied in last decade. In clustering methods all groups of document are assigned to a different topic. Normally clustering is schemes which are evolved basic concepts like data illustration model, similarity live, cluster model, and cluster algorithmic program. Most of this documents cluster strategies area applied the Vector Space Document (VSD) schemes. The basic phenomena are structure of this data representation for any text file whether it’s a webpage or document file. This file is converted as a feature array of the deferent terms presented in document of a data set. All term of words within document normally smallest term of document hence used as feature of document in VSD schemes. This is important that words are necessary elements for any language documents such as natural languages (including hindi or english) to represent to extract pattern. Mostly, the calculation of weights (normally indicates tf-idf, term-frequencies and inverse document-frequencies) of the words are as well as have within feature array. The similarity membership value within any two documents can be derived from calculating any one of the numerous similarity measures using the two corresponding feature values, such as cosine similarity or Jaccard relationship function measure or Euclidean similarity measure.

Clustering is the process of organizing data objects into a set of disjoint classes known as cluster. Purpose of clustering is that finding cluster comparable between themselves as well as different toward the terms fit in to other clusters. Document clustering is the task of automatically organizing text documents into meaningful clusters or group. In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics.

Clustering is an example of unsupervised categorization. Classification means a method with the purpose of allocate data terms headed for a locate of classes. Unsupervised represent that clustering not needed mention groups and training models though categorizes the data terms. Clustering is a crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image process, marketing, data processing, economics, etc. Cluster analysis is one among the first knowledge analysis tools in data processing. Most of clustering techniques belong only two basic type’s Hierarchical scheme and Partition methods. For Hierarchical method partition given document collection within smaller sets of documents sets in hierarchy manner. A partition method breaks dataset within predefined numbers of documents set in a step.

The input to clustering algorithm usually may be any set of entities or patterns or text files or images or software components. The output of clustering algorithm will be a partition of cohesive groups. The descriptions or representations of clusters may be used for decision making in selecting a software component or pattern of interest. One interesting property of clustering is all the patterns within a cluster share common properties in some sense and patterns in different clusters are dissimilar. From perspective of software engineering, all the components within same cluster have high cohesion.
and low coupling. Software component clusters can be treated as highly cohesive groups with low coupling which is the desired feature. One disadvantage of existing data clustering methods is that they do not adequately address the problem of processing large datasets with a limited amount of resources. Using these limitations as our motivation, so if we can try to reduce the dataset for training process it can help in reducing the cost of training which in turn improves efficiency of clustering. If done so, clustering takes less amount of space and hence forms a compact storage of patterns. Clustering is not any one specific algorithm that we can stick firm to, but it must be viewed as the general task to be solved

A divisive clustering starts with one cluster of all data points and recursively splits into the most appropriate clusters. The process continues until a stopping criterion is achieved. There are two main issues in clustering techniques. At first, finding the optimal number of clusters in a given dataset and secondly, given two sets of clusters, computing a relative measure of goodness between them. For both these purposes, a criterion function or a validation function is usually applied. In conventional clustering objects that are similar are allocated to the same cluster while objects differ are put in different clusters. These clusters are hard clusters. In soft clustering an object may be in more than two or more clusters

![Fig 1. The Stages of the Process of Clustering](image)

1. Collection of Data includes the processes like crawling, indexing, filtering etc which are utilize to accumulate the documents so as to require be clustered, index them to save and access in appropriate manner and select them to eliminate unusual data and information such as stemming and stop words need to remove.
2. Preprocessing consists of steps that take as input a plain text document and output a set of tokens (which can be single terms or n-grams) need to added in vector. This method follow these steps-
   2.1 Filtering is method of eliminate special characters, full-stop and other punctuations from documents because they don’t useful for information extraction. Discriminative power under the vector model. This is more critical in the case of formatted documents, such as web pages, where formatting tags can either be discarded or identified and their constituent terms attributed different weights [1].
   2.2 Stemming the method of dropping words to their support structure, or stem. in support of instance the words “connected”, “connection”, “connections” are all reduced to the stem “connect.” Porter's algorithm is the de facto standard stemming algorithm.
2.3 Stopword elimination: A stopword is distinct term, which is not consideration to express any logical information as a dimension in the vector space (i.e. without context). A usual technique to eliminate stopwords is toward evaluate every word with a collection of recognized stopwords. Another approach is to first apply a part-of-speech tagger and then reject all tokens that are not nouns, verbs, or adjectives.
2.4 Frequent word removal: trimming removal terms that come into sight with extremely small occurrence during the document body. The fundamental supposition is that these terms, still if they had at all discerning power, would appearance moreover little cluster not to be helpful. A predefined value is normally used, e.g. a little portion of the number of terms during the corpus. Occasionally terms if take place as well regularly (e.g. in 40% or further of the files) are further deleted.

2. PROBLEM FORMULATION

Given a document collection D={d1,d2,d3,…,dN} which contains N documents, there is need to sub-group [2] the documents based on the semantic of the text contents present in a Document , assuming we require K such sub-groups, the clustering process generates C={c1,c2,…,ck} clusters, with each ci being non empty.

Document clustering is still a developing field which is undergoing evolution. It started off on the popular vector based approach where documents were treated as a bag of words and clustering criteria was the presence of common words in the documents. Several modifications were applied on this method to improve this method as the result set would only provide us information on what words were present in a group of documents, not the actual content or context of the documents. There was a need of more intuitive ways of clustering that would provide us sound knowledge of the content present inside the documents.

3. RELATED WORK

Clustering is mainly effected by three factors during generation of good quality of cluster. They are data representation, similarity function and clustering mechanism which is applied [3, 4, 5, 6]. Vector space model is well known method for data representation used in document clustering. In this method all documents are collected in form collection of vectors where each vector is a document data. Every word has TF and IDF values. Similarity between any
document pair is calculated by jaccard function [7, 8] similarity measure. In this method only word represents whole document hence not finds many valuable information from word proximity [4]. Other methods based on Vector space model also not consider incremental processing [9]. Survey of web clustering engines that incremental processing increases the effectiveness while useful in the clustering schemes [10]. The most related work that takes into account the information about proximity of words and phrase based analysis in an incremental way is Suffix Tree clustering (STC) [11].

Frequent item sets finding problem is explain in detailed paper [12]. Normally frequent items are derived by association rules mining. But as technology growing this method also capable to calculate frequent item form documents and also for documents clustering and other mining works. Paper [13] mention method for clustering documents by calculating neighbor function value from given document set. Another method for given in paper [14] for maximum frequent item set length finding. Document classification is also done by Gaussian membership value between documents. This value is helpful for clustering.

In [15], the categorization of documents is complete by allowing for Gaussian relationship and making use of it to achieve clusters by discovery term. Every cluster is defined by its term behavior find by Fuzzy Gaussian relationship if cluster created. A novel technique known as Maximum Capturing is projected for text file clustering in [16]. Maximum Capturing involved two actions as decision document clusters and giving cluster members. In [17], algorithm to search for a pattern in a text is proposed which can be used to search for component of interest in the component repository.

Hierarchical Clustering is prefered over non-hierarchical Clustering because in non-hierarchical clustering a central point, also called centroid, needed to be choosen randomly and the distance from that point is calculated to group documents (with less distance) in one cluster[18]. Finding this central point poses an enormous challenge. that's why non class-conscious strategies aren't extremely popular. A comparative work on both the methods is done by [19]. Florian Beil et. al. introduced two clustering algorithms FTC(non-hierarchical) and HFTC(hierarchical) in [19], based on the concept of frequent Term Set and analysed their behaviour.

They used the association rule mining to identify the frequent terms in documents to group them into clusters. Agglomerative Hierarchical Clustering (AHC) is a widely used bottom up clustering algorithm. Many researchers have efficiently used this method in a information retrieval and data and Web Clustering [20]. There is a group of methods that is used for Agglomerative hierarchical Clustering. Most common of these methods are Single Linkage, Complete Linkage, cluster Average Linkage and Wards technique. All of these methods differ in the approach of similarity calculation that guides the selection of the most similar pair of clusters.

Recent research has developed new methods for estimating the semantic distance between both terms and words. Rudi L. Cilibrasi et. al. introduced a replacement similarity, called Normalized Google Distance(NGD) [20], to effectively capture the semantic similarity between words and phrases based on information distance and Kolmogorove Complexity. Later on, Alberto J Evangelista et. al. reviewed the work of Rudi L. Cilibrasi to boost their distance operate through elimination of random data[21]. We tend to adopt this technique to estimate the similarity between among terms clusters rather than just two words.

4. DOCUMENT CLUSTERING ALGORITHMS

Clustering is thus preformed after the documents matching the query are identified. Consequently, the set of thematic categories is not fixed they are created dynamically depending on the actual documents found in the results. Secondly, as the clustering interface is part of a search engine, the assignment of documents to groups must be done efficiently and online. For this reason it is unacceptable to download the full text of each document from the Web – clustering ought to be performed based solely on the snippets returned by the search service.

4.1 Agglomerative Hierarchical Clustering (AHC) Algorithm

The basic process of hierarchical clustering:

1. If there are n items them make n clusters and allocate every item toward a cluster. Every cluster be supposed to contain only single item.
2. Discover the most same match up of clusters and combine them into a new one cluster, thus now there is one cluster less in result set.
3. Compute similarities between the new clusters and each of the old clusters.
4. Go over steps 2 and 3 until the entire documents are clustered within desire number of cluster with size N

The main problem with AHC [22], that, they are very slow with large amount of data provided and also very sensitive with halting criterion that is, by mistake it can merge valuable clusters into one cluster. Also they do not scale well. They can never undo what was previously done. With outliers it performs poorly.
4.2 K-Means Algorithm
This algorithm is based on the center locations [23]. It first finds out the k cluster center location. Then each data point finds out which center is closest to it. Each center finds the centroid of the points and jumps to there. The main benefit of K-means algorithm is that, it is capable to produce overlapping clusters. Its main disadvantage is that it is most effective when the desired clusters are approximately spherical with respect to the similarity measure used. There is no reason to believe that documents should fall into approximately spherical clusters.

4.3 Suffix Tree Clustering (STC)
STC includes two main steps [24]. First it searches for all sets of documents that share a common phrase. They are found by suffix tree data structure. In second step we merge these phrases into cluster. The merge process is dependent on the percentage of the documents that contain both phrases. It also allows overlapping clusters. STC uses simple cluster definition. Also, STC is a fast incremental linear time algorithm which makes it suitable for web search clustering. It is faster than K Means. The main benefit of Suffix Tree Clustering is that it uses phrases to provide concise and meaningful descriptions of groups. But needs some thresholds for cluster formation and they turn out particularly difficult to tune. Its main disadvantage is it removes longer high quality phrases and use only shorter phrases. Finally, if a document does not include any of the extracted phrases or just some parts of them, it will not be included in the results although it may still be relevant.

4.4 Semantic Hierarchical Online clustering (SHOC)
The Semantic Online Hierarchical Clustering [25] is a web search results clustering algorithm that uses variation of the Vector Space Model called Latent Semantic Indexing (LSI) and uses phrases in the process of clustering. Unlike STC, SHOC improves the quality of label. STC gives incomplete labels while SHOC gives complete phrases. With SHOC documents can belong to several clusters. SHOC includes two key concepts: Complete phrases and definition of continuous clusters. It should meet the three requirements: Semantic, Hierarchical, and Online. It has three steps:
1. Data collection and cleaning
2. Feature extraction and
3. Identifying and organizing clusters.
Problems with SHOC
The problem with SHOC [25] is that it provides only vague comments on the values of thresholds of their algorithm and the method which is used to label the resulting clusters. It uses the singular value decomposition. So it may create unintuitive, random continuous clusters. It might be because of the input snippets used.

4.5 Lingo Algorithm
The Lingo algorithm is applied by the web searcher and this method is mainly uses whole phrases and LSI [26]. Lingo is an improvement of SHOC and suffix tree and different most of other methods, it initially find out descriptive labels for the current clusters and after that allocates the text files into suitable cluster. single drawback with this method is that the topic partition part typically requires arithmetical conversions that require a most calculating time, using Singular Value Decomposition.

5. ADVANCED CLUSTERING TECHNIQUES
5.1 Personalized Web Search Engine Using Suffix Tree Clustering
For make uses of clustering methods for maintained of web document efficiently as required effective users’ search paper [27] gives schemes. Searching software make use of links, images and associated information from global internet. Text files already accessed again used for STC systems by this searching capability of search engines are increased. All result of search engines are maintained by page rank methods it uses hyperlinks and information structure to store web pages. Thus software make a pattern of users’ document accessing on internet and this is monitored and checked according users previous searches. The web documents are fetched by using crawlers [1] and are further organized in to clusters which aid the user in finding the required details faster. Extended suffix tree clustering algorithm is proposed which analyzes the retrieved web documents semantically descriptive meaningful phrases as cluster labels.

5.2 Hesitant Distance Similarity Measures
Paper [28] presents new approach, Hesitant Distance Similarity Measures for Document Clustering. The proposed Hesitant Distance Similarity Measures
### Table 1. Comparison of Web Document Clustering Algorithm

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Cluster Diversity</th>
<th>Cluster labels</th>
<th>Scalability</th>
<th>Time Complexity</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agglomerative Hierarchical Clustering (AHC) [22]</td>
<td>not very robust towards outliers</td>
<td>Most frequent terms</td>
<td>Low</td>
<td>Single link and group average: O(n^2) Complete link: O(n^3)</td>
<td>Simple</td>
<td>- Slow when applied to large document collections. - Sensitive to halting criterion. - Poor performance in domains with many outliers.</td>
</tr>
<tr>
<td>K-menas [23]</td>
<td>Low, small (outlier) clusters rarely highlighted</td>
<td>One-word only, may not always describe all documents in the cluster</td>
<td>Low</td>
<td>O(nk^2) or O(nk^3) iterations</td>
<td>Efficient and simple. - Suitable for large datasets.</td>
<td>- Very sensitive to input parameters.</td>
</tr>
<tr>
<td>Suffix Tree Clustering (STC) [24]</td>
<td>Low, small (outlier) clusters rarely highlighted</td>
<td>One-word only, may not always describe all documents in the cluster</td>
<td>High</td>
<td>O(n)</td>
<td>- Incremental - Uses phrases to provide concise and meaningful description of groups.</td>
<td>- Snippets usually introduce noise. - Snippets may not be a good description of a web page.</td>
</tr>
<tr>
<td>Semantic Online Hierarchical Clustering (SHOC) [25]</td>
<td>Low</td>
<td>Label that describe the cluster</td>
<td>High</td>
<td>O(n)</td>
<td>- Uses Latent Semantic Indexing (LSI) and phrases in the process of clustering. - Uses suffix array to identify complete phrases. - Allows overlapping clusters. - Provides a method of ordering documents</td>
<td>- Provides only vague Comments on the values of thresholds of the algorithm and the method which is used to label the resulting clusters.</td>
</tr>
<tr>
<td>Lingo [26]</td>
<td>High, many (outlier) clusters highlighted</td>
<td>Longer, often more descriptive</td>
<td>Low, For More than about 1000 documents, Lingo clustering will take a long time and large memory</td>
<td>O(n)</td>
<td>- Readable cluster Labels. - Overlapping clusters. - Cluster accuracy</td>
<td>- Unable to generate a Hierarchical structure of clusters. - The implementation of lingo is fairly computationally expensive.</td>
</tr>
</tbody>
</table>

#### 5.3 Clustering with New Ranking and Similarity Measures

Another author of paper [29] gives a novel algorithm for document grouping this method is apply ranking feature for finding cluster of documents. STHAC method is combination of hierarchical agglomerative schemes using STC to enhanced cluster combine step. Simulation result also supports given algorithm effectivenes. This method improves F-measure from STHAC is comparatively batter then previous methods because of better filtering minimum value cluster, hence combination of clusters is more effective within this method. STHAC, Suffix Tree based Hierarchical and Agglomerative Clustering that has made four key contributions to the conventional. Suffix Tree Clustering to improve overall cluster quality and hence clustering performance.

Comparison between couples of files can be definite whichever explicitly or implicitly. Paper [30] gives a Multiviewpoint-based Similarity activity methodology, named MVS. Theoretical analysis and empirical examples show that MVS is doubtless additional appropriate for text documents than the popular circular function similarity. Supported MVS, 2 criterion functions, IR and IV, and their several agglomeration algorithms, MVSC-IR and MVSC-IV, are introduced. Compared with different progressive agglomeration strategies that use differing types of similarity live, on an outsized range of document knowledge sets and beneath completely different analysis metrics, the planned algorithms show that they might offer considerably improved agglomeration performance. The key contribution of this paper is that the elementary construct of similarity live from multiple viewpoints. Future strategies may create use of a similar principle, however outline different forms for the relative similarity in (10), or doesn’t use average however produce other strategies to mix the relative similarities in keeping with the various viewpoints. Besides, this paper focuses on divided agglomeration of documents. Within the future, it’d even be doable to use the planned criterion functions for stratified agglomeration algorithms.

#### 5.4 Multiviewpoint-Based Similarity Measure
5.5 Structural and Textual Feature Extraction for Semi-structured Document

Author of paper [31] give classification method for XML documents using two different methods. One is structural and second is content based feature selection method. This method is making best use of structured present in document for similarity measure between two structured documents as text information present in XML file. Author proposed a new framework for document classification, in particular XML document classification, based on extracting features from different aspects of the document. Extracted a feature vector that represents the document in a compact format; it captures valuable information that can be used later on to build an accurate classifier using any of the well known classification techniques. It also proposed a fast, robust, accurate, and novel approach for content-based document classification. Our proposed model is easy to implement for real world applications.

5.6 Correlation Similarity Measure

Paper [32] offers a novel document clustering technique derived from correlation indexing maintain. It at the same time take full advantage of the correlation among the files within the limited patches and reduce the correlation among the files outside these patches. Accordingly, a low down dimensional semantic subspace is resultant wherever the files equivalent to the matching semantics are near to both further. Extensive experiments on NG20, Reuters, and OHSUMED corpora show that the proposed CPI method outperforms other classical clustering methods. Furthermore, the CPI method has good generalization capability and thus it can effectively deal with data with very large size.

5.7 Hybrid XNOR similarity function

The paper [33] defines a new similarity function to compute similarity between any two software components or text files. An algorithm to cluster a set of given documents or text files or software components is designed which uses the proposed similarity function called hybrid XNOR to find the degree of similarity among any two entities. The input to algorithm is a similarity matrix and the output is the set of clusters.

6. SUPPORT VECTOR MACHINE

Support Vector Machines are linear classifiers based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. Additionally, SVMs are “a classification method that determines the maximum-margin hyperplane” [17]. This terminology is defined shortly. It should be noted that Support Vector Machines compete with Neural Networks as tools for solving pattern recognition problems.

Support Vector Machines learn their classification through a training data set of the form

\[ \{x_i, y_i\}_{i=1}^{l} \]

where \( x_i \) is the \( n \)-dimensional vector that describes the features of that instance and a label \( y_i \) that classifies the instance as belonging to one of two categories, 1 or -1 (‘positive’ or ‘negative’) respectively. Given sufficient training examples the Support Vector Machine is then able to classify previously unseen examples (instances of data), those with no predefined label, into one of the two categories [18].

In the case of basic linear classification a Support Vector Machine creates a maximum margin hyperplane that lies in a potentially transformed input space. Given binary choice training examples (labeled either ‘positive’ or ‘negative’), a maximum-margin hyperplane divides the ‘positive’ and ‘negative’ examples, such that the distance from between the respective class (in this case ‘positive’ or ‘negative’), to the hyperplane is maximized. This is termed maximising the margin.

This can also be viewed from a geometric standpoint. The Support Vector Machine attempts to construct a decision surface that bisects \( \mathbb{R}^n \) such that all instances belonging to the positive class appear on one side of the surface with all instances belonging to the negative class appearing on the other. See Figure 2. While this approach is not new to the classification field, where SVMs set themselves apart is in their implementation. In order to obtain a maximum margin between a class and the decision surface we must define a convex hull for that class and maximise the margin in respect to that hull. This is because the closest approach of a particular class to the decision surface may not be at a specific point, but a linear combination of points.

7. CONCLUSION

In this survey we had projected various clustering approaches and algorithms used for document clustering. The areas of document clustering have lots of issues, which need to handle. This paper gives interested readers a broad inspection of previous methods. In future new
method can be derived for improved results of existing method by applying various techniques or combination of better methods to improve searching result within best user preference match. Such new method must fulfill limitations present in current work by clustering or co-clustering methods. In future some approaches can be used for clustering result improvement such as classifier can be extends using efficient classification algorithms also fuzzy logics can be applied. Using support vector with this classifier may good method for classification whenever cluster formed. SVM network can be trained by input document feature and cluster label once then for new document matrix can be result by using trained network for new document. Hence time of feature matching will minimize.

8. References
List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

[8] A.B. Smith, C.D. Jones, and E.F. Roberts, Book Title, Publisher, Location, Date.
[10] A.B. Smith, C.D. Jones, and E.F. Roberts, Book Title, Publisher, Location, Date.
[16] A.B. Smith, C.D. Jones, and E.F. Roberts, Book Title, Publisher, Location, Date.
[18] A.B. Smith, C.D. Jones, and E.F. Roberts, Book Title, Publisher, Location, Date.
[22] A.B. Smith, C.D. Jones, and E.F. Roberts, Book Title, Publisher, Location, Date.
Knowledge Discovery (FSKD), Volume: 2, Pages 1187–1190, IEEE, 2011.


