INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 3)
Available online at: www.ijariit.com

Review of different Text documents clustering techniques

Jyoti Verma
vermajyoti051@gmail.com
Deenbandhu Chhotu Ram University of Science and Technology, Sonipat, Haryana

Neetu Verma
verma21neetu@gmail.com
Deenbandhu Chhotu Ram University of Science and Technology, Sonipat, Haryana

ABSTRACT

Along with the rapid and fast development of the Internet, there is a prodigious increase in the use of data and information. The aggressive growth of data has led us to an information explosion era, where the data cannot be easily maintained. Also, there is an increase in the use of electronic data and the information is stored in electronic format in the form of text documents such as news articles, books, digital library and so on. Clustering of the text documents has become an important technology over the internet. Text Clustering is mainly described as a grouping of the similar documents a large collection of unstructured documents. Text document clustering is the most widely used method for generalizing a large amount of information. In this paper, we tried to compare some existing text document clustering techniques on the basis of few criteria like time, accuracy and performance.

Keywords: *Text mining, Clustering, Document clustering, K-means, Dimension reduction*

1. INTRODUCTION

There is a huge amount of data is increasing day by day which is difficult to handle. It may be in the structured form or unstructured form. So, for converting this data into a structured or meaningful form, there is need for data processing applications. Data can be either image, text, audio, video, graphics, spatial form etc. Among with all common forms of data, we are handling with the data which is in the form of text documents. All text form data are mainly news stories which we are reading, posting and messaging on social media. So, Text mining process has a great significance nowadays [1].

For handling a large volume of text data, a technology is being introduced which is called as text mining. Text categorization, text clustering, text classification etc. are the different functionalities of text mining. The term text mining is used to explain either a single process or a bunch of processes. In text mining, we can obtain previous undefined information by extracting this information automatically form the different-different digital data sources [2].

1.1Areas of Text Mining

In text mining, we mined the structured data from the unstructured and semi-structured data by applying certain patterns. In this activity text clustering, text classification, text summarization, information extraction, information retrieval, association, visualization methods are involved. By using theses, structured data achieved.

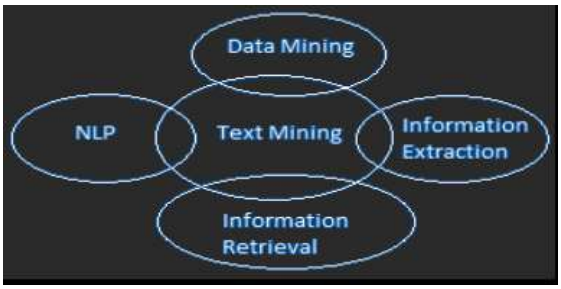


Fig.1: Text mining areas [3]

1.1.1Information Retrieval

Information retrieval is the activity of obtaining or retrieving useful information from the certain documents according to user's need. So, this is also designated full form of document retrieval. This activity is followed by text summarization stage. User posed a query which is focused on information retrieval. It is also called as information extraction stage. Information retrieval system helps in to check the set of text documents that are relevant to the particular query. It is very difficult to apply text mining algorithms on the large collection of text documents. Information retrieval can increase the speed of analysis significantly by decreasing the number of documents for analysis [3].

1.1.2Data Mining

Data mining is the process of extracting information from a large amount of dataset. Data may be in the structured form or unstructured form, then apply the certain pattern on this

data and convert this data in a meaningful form or in knowledge. The knowledge or information which is extracted is used in many applications like fraud detection, market analysis, credit card, astrology etc. Knowledge can be mined from data by using data mining techniques like clustering, classification, association, regression etc. [3].

1.1.3 Natural Language Processing

Natural language processing is a way of computers for analyzing, understanding and deriving the meaning of human language in a smart and useful way. With the help of natural language processing, developers can organize the knowledge and convert it into structured form to perform certain tasks like sentimental analysis, text summarization, speech recognition etc. It allows the machine to understand how human speaks. With the help of this accuracy of machine increases. It summarizes the blocks of text, creates chat box, automatically generates keyword text, identifies the type of entity which is extracted, performing tokenization etc. [3].

1.1.4 Information Extraction

Information extraction is the phenomenon of extracting structured information from unstructured and semi-structured documents automatically. These documents are in the machine-readable form. This task perform sometimes with the help of natural language processing. For example, when the email extracts only the data from the message which you add to your calendar. It will extract information from many sources medical records, social media, news group online, co-operate sectors, government sector etc. There are many techniques which are used for information extraction. Applying patterns on data of text documents and collect the knowledge or structured data [3].

1.1.5 Text summarization

Text which is produced from more than one text documents which convey important information in the original text document is called as a summary. It is in a shorter form than the original text document. The purpose of text summarization is converting the original text into the shorter form with semantics. The benefit of using text summarization is that due to this reading time will reduce. This method is categorized in two forms abstractive and extractive. Abstractive summarization is to express the main concept of a document in the form of clear natural language by understanding document's main concept. Extractive summarization is selecting the main sentences, paragraphs, quotes etc. from the original text document and concatenating all these and make the document's short form [4].

1.1.6 Unsupervised learning methods

Unsupervised learning methods are techniques looking for hidden shape out of unlabeled information. They do not want any training segment, consequently may be applied to any text documents without manual effort. Clustering and topic modeling are the two normally used unsupervised methods algorithms used in the context of text documents. Clustering is the task of segmenting a collection of files into partitions where files inside the equal organization (cluster) are more just like every other than the ones in different clusters. In topic modeling a probabilistic model is used to determine a tender clustering, wherein every document has a chance distribution over all of the clusters instead of hard clustering of documents. In topic models each subject matter can be represented as chance distributions over phrases and every file is expressed as opportunity distribution over topics [2].

1.1.7 Supervised Learning Methods

Supervised learning methods are the techniques which are not looking for hidden shapes out of unlabeled information. Supervised learning methods requires training segments. There may be a large variety of supervised methods along with nearest neighbor classifiers, decision trees, rule-based totally classifiers and probabilistic classifiers [5].

1.1.8Text streams and social media mining

These days' social media become an important source of knowledge. Through social media thousands of people share their views and ideas. Huge amount of posts generated daily from these open broadcasting platforms. The information of these posts are stored in various formats of text documents. Social networks, particularly Facebook and Twitter create massive volume of text records constantly. They offer a platform that lets in customers to freely specific themselves in a huge range of subjects. The dynamic nature of social networks makes the process of textual content mining difficult which needs unique capability to handle poor and non-preferred language [6].

1.1.9 Opinion mining and Sentimental Analysis

With the trends of e-commerce and online shopping, a huge amount of text is created and continuous to grow about different-different opinions of the user about the products. A huge amount of information is stored in processing and adjusting the price of products and services. By mining, such data find out important data and opinion about a topic which is significantly fundamental in advertising and giving online opinion [7].

1.1.10Biomedical Text Mining

The biological research field is growing rapidly. It makes difficult to access and needed data. For searching a reliable information in a particular domain is difficult and there is no sustainable method to retrieve and analyze the very large amount of datasets in an effective manner. The goal of biomedical text mining is to allow researchers to identify needed information more efficiently from sources such as Medline abstracts, abbreviations from scientific texts, etc. They use various manipulating methods and techniques to achieve the required results. This paper reviews various techniques proposed by various researchers such as new techniques, algorithms, tools, and methodologies like MyMed, DiscoTEX, dictionaries, etc. These techniques are assessed on the extracted information and knowledge in an efficient manner to find out the relationships among the information. These techniques are employed to lessen the burden of information overload by applying it to the vast data source) [8].

1.2 Text Clustering

Clustering is a separation of data into groups of similar objects. Each group of similar objects called as a cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate

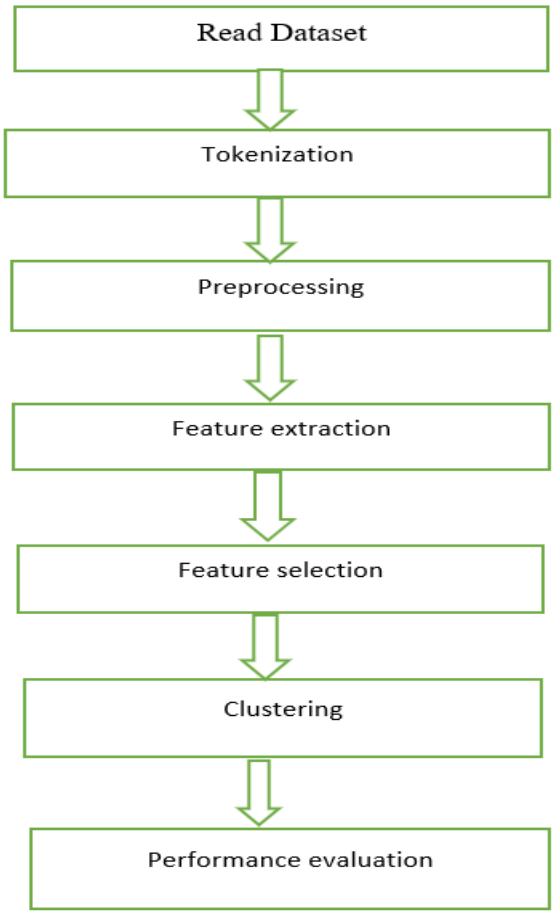
distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering.

Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership.

Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so-called training set, i.e. a set of data correctly labeled by hand, and then replicates the learned behavior on unlabeled data. The goal of a document clustering scheme is to minimize intra-cluster distances between documents while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering. The large variety of documents makes it almost impossible to create a general algorithm which can work best in case of all kinds of datasets **[2]**

1.2.1Text Clustering Process

The stages of text clustering are discussing as following points:



a. Read Dataset

The first step of Text clustering is read dataset. In this point of clustering, read the dataset from the collection of documents. These documents are of different types or formats like .html, .mat, .pdf, .docx, web contents etc. [9].

b. Tokenization

After text documents collection, next step is tokenization. In tokenization, converting the text document's string data in the form of tokens [10].

c. Pre-processing

In next step, Text preprocessing is to be done. In text preprocessing, four steps are involved.

- Lemmization- In this step, get the base form of each word.
- Stop word removal- In next step remove the stop words or can say that less meaningful words like example, etc.
- Special symbol removal – In this step remove the special symbols ',/;'[]<>?:"{}_*&^%'.
- Case conversion – In this step, convert all the words into lower case.

After performing all above steps, the output is taken for next step of feature extraction [11].

d. Feature Extraction

For describing the large dataset, feature extraction reduces the number of resources. When performing analysis on a large number of resources large memory and power required. Due to this consumption, performance reduces. For overcoming this problem, feature extraction is used which improves the accuracy [12].

e. Feature selection

Feature selection also known as variable selection, is the process of selecting a subset of important features for use in model creation. The main assumption when using a feature selection technique is that the data contain many redundant or irrelevant features. Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context. Feature selection technique is a subset of the more general field of feature extraction [3].In this weighting schemes are used which are described below:

e.1. Unsupervised term weighting schemes

Most of the unsupervised term weighting schemes is from the information retrieval field. These methods are very useful when the training documents are not labeled by their class labels. The traditional term weighting methods borrowed from IR, such as binary, term frequency (TF), TF-IDF, and its various variants are unsupervised schemes. The TF-IDF proposed by Jones and its variants are the most widely used term weighting schemes for text classification. Some of the variants of TF are Raw term frequency, log (TF), log (TF+1), or log (TF)+1. If ni is the number of documents containing the term and N is the number of documents in the collection then, the variants of IDF are 1/ni, log(1/ni), log(N/ni), log(N/ni)+1 and log(N/ni-1)[1]. In a novel inverse corpus frequency (ICF) based technique is proposed which computes the document representation in linear time [13].

e.2. Supervised term weighting schemes

Supervised term weighting schemes have been developed specially for text categorization due to the reality that a supervised expertise on the magnificence labels of the training samples is furnished. All the supervised term weighting schemes make use of this class information in unique methods. Supervised term weighting schemes are further classified into subcategories, based totally on whether or not the load estimates relevancy of a time period in preserving document content material or the relevancy of a term in setting a document as a member of a category. So, it will be greater powerful to name the weighting schemes which can be used to measure the relevance of a time period in maintaining the file content material as time period-document relevance measures and people which may be used to degree the time period relevance in categorizing a record as term elegance relevance measures[13].

f. Clustering

Clustering is an unsupervised process to classify the text documents in groups by using different clustering algorithms. In a cluster, same descriptions or designs are grouped got from different documents. Clustering is carried out in top-down and bottom-up behavior. The similarity degree of the content of the documents in the same cluster should be to the most, while in different clusters to the least. Various methods of clustering are distribution, density, centroid, hierarchical and k-means [3].

f.1.1. Clustering methods

Here, clustering methods are performed on text documents. For performing text document clustering following methods are used. On the basis on which clusters made. These methods are explained below:

f.1.1.1 K-means

K- means is one of the simplest unsupervised learning algorithms that solve the well -known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more[14]. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in the i^{th} cluster.

'c' is the number of cluster centers.

Advantages:

1) Fast, robust and easier to understand.

2) Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.

3) Gives the best result when data set is distinct or well separated from each other.

Disadvantages

1) The learning algorithm requires apriori specification of the number of cluster centers.

S. No.	Attributes	K-means	Hierarchical clustering
1	Time complexity	O(n)	$O(n^2)$
2	Shapes of clusters	Hyper-spherical	Non-spherical
3	Repeatability	Repeatable results	Non-repeatable results
4	Support big data	Support	Not support
5	Prior knowledge	Require	Not require
6	Support Nominal\binary data	Not possible	Possible
7	Noise	Sensitive	No effect
8	Outlier points	Sensitive	No effect

2) The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.

3) The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).

4) Euclidean distance measures can unequally weight underlying factors.

5) The learning algorithm provides the local optima of the squared error function.

6) Randomly choosing of the cluster center cannot lead us to the fruitful result. Pl. refer Fig.

7) Applicable only when mean is defined i.e. fails for categorical data.

8) Unable to handle noisy data and outliers.

9) Algorithm fails for non-linear data set.

f.1.1.2 Hierarchical Algorithm

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. Hierarchical methods are usually classified into Agglomerative and Divisive methods depending on how the hierarchy is constructed.

Agglomerative methods start with an initial clustering of the term space, where all documents are considered representing a separate cluster. The closest clusters using a given inter-cluster similarity measure are then merged continuously until only 1 cluster or a predefined number of clusters remain.

Divisive clustering algorithms start with a single cluster containing all documents. It then continuously divides clusters until all documents are contained in their own cluster or a predefined number of clusters are found.
Agglomerative algorithms are usually classified according to the inter-cluster similarity measure they use. The most popular of these are single-link, complete-link and group average. In the *single link* method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters (one from each), in the *complete link* it is the maximum distance and in the *average link* it is correspondingly an average
Distance **[15].**

f.1.2. Differences between k-means and hierarchical clustering

g. Performance evaluation

After performing all above steps of document clustering, performance evaluated by measuring four factors time, precision, recall, and f-measure.

Precision – It finds out that how many selected items are relevant?

Recall – It finds out that how many items are selected those are relevant?

F-measure – It is a harmonic function of precision and recall. It is used for measuring accuracy.

2. APPLICATION

Text Mining can be applied in many areas. Some of the most common used areas are:

2.1 Web Mining

These days web contains a lot of information about subjects such as persons, companies, products, etc. [3] that may be of huge interest. Web Mining is an important application of data mining techniques to discover hidden and unknown patterns from the Web. Web mining is an important activity of recognizing term implied in large document collection say C, which can be denoted by mapping i.e. C →p. The first take toward any Web-based text mining effort would be to gather a substantial number of web pages having an observation of a subject. Then, the question becomes not only to find all the subject developments but also to separate out those that have the wanted meaning.

2. 2 Clustering

Clustering is an unsupervised process to classify the text documents in groups by using different clustering algorithms. In a cluster, same descriptions or designs are grouped got from different documents. Clustering is carried out in top-down and bottom-up behavior. In NLP, many types of mining tools and techniques are used for the determination of unstructured text. Various methods of clustering are distribution, density, centroid, hierarchical and k-mean [3].

2.3 Social Media

Text mining software packages are available for analyzing social media applications to monitor and analyze the online plain text from internet news, blogs, email etc. Text mining tools help to identify the number of posts, likes, and followers on the social media. This kind of findings shows the people reaction to different posts, news and how it gets spread around. It shows the behavior of people belonging to specific age group or communities having similarity and differences in views about the specific post.

2.4 Resume Filtering

Big companies and headhunters get thousands and lakh of resumes from job applicants every day. Obtaining information from resumes with high precision and revising is not an easy task [3]. Instead of constituting a restricted domain, resumes can be written in a multi tudinal formats (e.g. structured tables or plain texts), in different languages (e.g. Japanese and English) and in different file types (e.g. Plain Text, PDF, Word etc.). Moreover, writing styles can also be much varied. In the first manual scan of the resume, a recruiter looks for mistakes, educational qualifications, employment history, job titles, the frequency of job changes, and other personal information. Exactly getting this information will be the first step in ignoring resumes. Hence, the process of selection of resume is an important task in recruitment.

2.5. Medical and life science

Users frequently exchange information with others about areas of interest or send requests to web-based forums, or ask the expert services [3]. Every people wants to understand particular diseases (what they have), to be told about new therapies, questioned for a second opinion before treatment. Additionally, these forums also indicate seismographs for medical and/or psychological requirements, which are correctly not met by present health care based systems. Medias like E-mails, e-consultations, and requests for medical advice through the network have been manually weighed using quantitative or qualitative methods. In order to help the medical experts and to make use of this seismograph function of expert forums, it would be helpful to distinguish visitors' requests instantly. So, particular requests could be directed to the expert or even answered semi-automatically, providing complete

monitoring. By creating ―frequently asked questions (FAQs) ‖ same alike patient requests [3] and their c] answers could be congregated, even before the particular expert responses. Machine-based conclusions could help the public to handle the mass of information and medical experts to give expert their feedback. An instant classification of amateur requests to medical expert network forums is a heavy task because these requests can be long and unstructured as an end of mixing, for example, personal experiences with laboratory data. It is a big challenge to find out a correct and important text to take a right decision from a huge biological repository. The records of medicals contain content varying in nature, complex, lengthy and technical vocabulary are used that make the knowledge discovery process difficult .The text mining tools in biomedical field gives an opportunity to obtain valuable information, their association and their relationship among various diseases. Text mining used in biomarker discovery pharmaceutical companies, clinical trade analysis, pre-clinical safe toxicity report studies, patent competitive intelligence and landscaping, mapping of genetically diseases and exploring the specified identifications by using different techniques.

3. CONCLUSION

For categorizing news articles, different-different text clustering algorithms can be used. Data set of text documents is obtained from the web and pre-processing techniques are applied to the data set in order to remove special symbols, stop-words, case- conversion, stemming and to do tokenization. Due to these, unnecessary words removed from the data set of text documents and provides a pre-processed dataset of these documents. Then, document clustering algorithm is applied for making clusters of news articles. By using a clustering algorithm cluster are formed. Similar news documents put in a cluster and dissimilar puts in another cluster. Results of clustering algorithm are improved by using k-means which increases accuracy and consumes less time. As Future work, such systems can be used to work for big data.

4. REFERENCES

[1] Thomas, A. M., & Resmipriya, M. G. (2016). An efficient text classification scheme using clustering. *Procedia Technology*, *24*, 1220-1225.

[2] Svadas, T., & Jha, J. (2015). Document Cluster Mining on Text Documents.

[3]Mrs.B.Meena Preethi, Dr.P.Radha, "A Survey Paper on Text Mining - Techniques, Applications And Issues" IOSR Journal of Computer Engineering (IOSR-JCE)

[4] N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, Nagercoil, 2016, pp. 1-7.

[5] Ge, L., & Moh, T. S. (2017, December). Improving text classification with word embedding. In *Big Data (Big Data), 2017 IEEE International Conference on* (pp. 1796-1805). IEEE.

[6] Kang, J., & Lee, H. (2017). Modeling user interest in social media using news media and wikipedia. *Information Systems*, *65*, 52-64.

[7] Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems*, *90*, 65-74.

[8] Savitha, R., & Porkodi, R. (2015). An Overview of Text Mining Techniques and Methodologies Used in Bioinformatics Domain.

[9] Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85.

[10] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, *4*(8), 966-974.

[11] Wongso, R., Luwinda, F. A., Trisnajaya, B. C., & Rusli, O. (2017). News Article Text Classification in Indonesian Language. *Procedia Computer Science*, *116*, 137-143.

[12] Patel, M. R., & Sharma, M. G. (2014). A survey on text mining techniques. *International Journal Of Engineering And Computer Science ISSN*, *2319*(7242), 5621-5625.

[13] Guru, D. S., & Suhil, M. (2015). A novel term_class relevance measure for text categorization. *Procedia Computer Science*, *45*, 13-22.

[14] Na, S., Xumin, L., & Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on* (pp. 63-67). IEEE.

[15] Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In *KDD workshop on text mining* (Vol. 400, No. 1, pp. 525-526).

