# Document Clustering based on the Similarity of Data with Efficient Time Consumption

Saidesh Kumar Padmala
Research Scholar
University College of Engineering, Osmania University
Hyderabad, India

## ABSTRACT

Text mining has becoming an emerging research area now-a-days which helps in extracting the useful information from large amount of natural language text documents. The necessity of grouping the documents for different applications is gaining comprehensive review of the techniques used to improve the efficient time consumption, challenges, research issues are presented. The techniques presented in the review are k-means clustering, fuzzy c means clustering, support vector machine classifiers, naive Bayes classifier, Hidden Markov Model (HMM). Furthermore, discussion of the advantages and disadvantages of each technique is contributed to a better understanding and compared with the existing techniques based on the efficiency and computational time.

## General Terms

Document clustering, text mining algorithms

## Keywords

Clustering, text mining, k-means clustering

## 1. INTRODUCTION

In day to day lives, a web is considered as the primary source for the text, and the availability of the information is consistently increasing. The organization information which is approximately around 80% is stored in unstructured formats. So, this shows that about 90% of the information is stored in an unstructured form. The retrieval of the useful knowledge from the massive amount of the textual data which is used in the human analysis is apparent [1]. Manual analysis of the unstructured data is highly impossible, and thus the text mining techniques are being developed to mechanize the process of analyzing the information. Text mining is an emerging concept at the combination of several areas including data mining, natural language processing and information retrieval [2]. The text mining extracts the information from the data sources through the explorations and identification of unusual patterns. This is a multidisciplinary field which performs clustering, categorization, extraction of data and machine learning [3]. The conversion of unstructured data text into structured data items by employing a set of algorithms is the process of text mining. The unstructured nature of the documents makes the text mining tasks more difficult than their data mining counterparts. The steps of text mining are as follows, a) Extract information from unstructured data, b) Extracted information converted into structured data, c) Pattern identified from structured data, d) Analyze the pattern. e) Extract the valuable information, f) Store in the database. The reason for applying the data mining methods to text document collections is to structure them. The existing techniques for structuring collections are assigning the documents, classification or structuring the documents, clustering and thus information and arrange them into significant subgroups for analysis. The benefit of clustering is that documents can appear in multiple subtopics by making sure that the useful document is not omitted from the results [3]. By organizing the similar documents together, large collections of documents can be easily navigated browsed and organized. Document clustering has found applications in many fields such as knowledge discovery, business applications etc. [6]. The system can also summarize individual documents. In this case, it runs a summarization algorithm that extracts the most relevant sentences from a document. The relevance of each sentence is determined by computing the average relevance of all the words in the (preprocessed) sentence. Document clustering has applications in an automatic organization of documents, topic extraction, fast information retrieval or filtering [7]. There is not at all a single operation from the collection of documents to the clustering of document collection, and it includes the number of stages that consists of four main stages [8].

- Preprocessing: The preprocessing is required before the document representation and should stop removing words such as ‘a’, ‘any’, ‘the’ as they are frequently used. The words need to be stemmed [9].
- Feature Extraction: It employs to produce the set of features by parsing each document and helps to remove the noise and reduce the dimensionality of feature space. The most commonly used feature selection metrics are term frequency and inverse document frequency [8].
- Document Representation: Most of the clustering approaches use the vector space model for document representation. The m*n matrix is represented by the collection of n documents with the m unique words where each document is a vector of m dimension [5].
- Document Clustering: At this stage, the target documents are grouped into different clusters by selected features [8].

As document clustering is a well-known feature from many years but still it is far from a solved problem and the challenges faced are appropriate feature selection, cluster labeling, selection of proper similarity measure and knowledge about the input parameters. The overall aim of the text mining is to extract information from the vast dataset and provide its users with an understandable form. The stages and process of clustering is shown in the below figure.

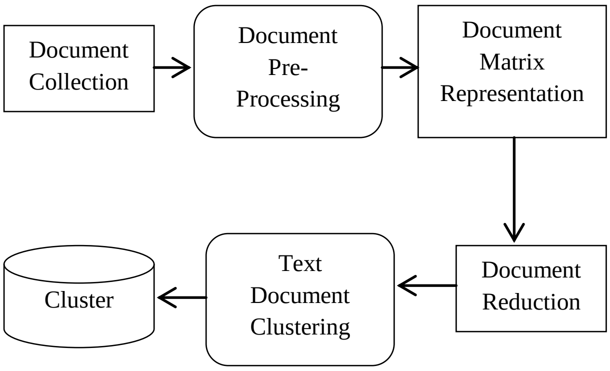

**Fig 1: The stages of cluster processing**

## 2. RELATED WORK

There are several techniques proposed by different authors regarding various text mining and document clustering techniques. These techniques are explored and discussed in the succeeding sections.

Chourasia et al. [10] proposed sequential minimal optimization algorithm, K nearest neighbor classifier and best first trees to predict the survivability for breast cancer patients. The Weka toolkit is used for experimenting with the three data mining algorithms mentioned above. The Weka tool is an ensemble of tools for data clustering, classification and visualization. The proposed method shows that the attributes considered for analysis are not the direct indicators of breast cancer in the patients. The sequential minimal optimization is more accurate classifier in comparison to BF tree.

Shen et al. [11] introduced an integrated system of text mining and case-based reasoning to retrieve the design for green buildings. By using CBR and text mining principles an integrated system is proposed to support the decision making in green building design process. As the green building has been increasing dramatically over years, the selection and application of the green building technologies under different situations usually puzzles designers. The effectiveness can and accuracy is increased by the proposed algorithms and help it to incorporate into green buildings.

Cutting et al. [12] presented a document browsing technique which employs document clustering as the primary operation. Fast clustering algorithms is also used which helps in interactive browsing paradigm. The document browsing method called scatter/gather for using document clustering and to implement scatter/gather fast document clustering is a necessity. Scatter/Gather is particularly helpful in situations in which it is difficult or undesirable to specify a query formally.

Turtle et al. [13] adapted an inference network model to support the use of multiple document representation schemes. The use of Bayesian inference networks for information retrieval represents an extension of probability-based retrieval research. The retrieval model presented provides a framework used to integrate several document representations and search strategies.

## 3. TEXT MINING AND DOCUMENT CLUSTERING

In this, a detailed review on the analysis of text mining process and techniques involved in document clustering is executed based on the similarity indices with effective time consumption and they are as follows,

For the purpose of clustering algorithms documents are represented using the vector space model. In this model, each document d is considered as a vector where each document is represented by the TF factor,

$$Dtf= (tf1, tf2…., tfn) \;……. (1)$$

Where, tfi is the frequency of the ith term of the document. The similarity between the documents must be measured in such a way that clustering algorithm to be used. The computation of similarity between the documents can be measured and defined as,

$$Cosine\ (d1, d2) = (d1, d2) /||d1||\ ||d2||………(2)$$

Where, indicates the vector product and ||d|| is the length of vector d.

For K means clustering the cosine measure is used to compute which document centroid is closest to a given document.

### 3.1 Genetic Algorithm for Text Mining

Applying genetic algorithms for text mining in searching better document descriptions is used from many years. The final goal is information retrieval researchers define genetic algorithm objective function based on the retrieval performance of past queries [14]. The work is to attempt and generalize the document descriptions beyond the specificity of one domain and thereby to accomplish the past queries cannot be used. So, a genetic algorithm is used for designing the co-occurrences of the documents. Genetic algorithms borrow their process from the Darwin natural process of survival. The adaptation of genetic theory to text categorization problem, the documents act as the chromosomes of the population. In text analysis, documents are represented by a set of index terms. Two most used processes in the genetic algorithm is the stemming and stop words. The stop word process helps in eliminating the significant words like ‘the’ and ‘a’. The stemming process helps in extracting the roots of the word to make a single term. For example, ski, skies and skiing can be considered as ski. The genetic algorithm generates new solutions by recombining the genes of the current best solutions. This is accomplished through the crossover and the mutation operators. The information retrieval is concerned with the classification processes and the selective recovery of information. For the text type of information, the situation is it consists of indexing a collection of documents with the keywords and then matching the index terms with the terms of a user query. When assessing the effectiveness of a retrieval process, the recall is measured by the number of relevant documents retrieved over the total number of relevant documents and the precision is measured by the number of relevant documents retrieved over the total number of documents retrieved. The effectiveness of the retrieval depends on the quality of the query and the quality of the index terms.

### 3.2 Bee Colony Optimization Algorithm

The Bee colony optimization algorithm (BCO) is considered as one of the fast, robust and efficient global search techniques in tackling practical problems in data clustering which is raised in many of the applications. The partition clustering method is used for large data sets. The attempt is to divide the data set into a set of disjoint clusters without the hierarchical structure. In order to overcome the problem in the traditional approaches of partition clustering the new techniques have been proposed and that bone method is the optimization methods that try to optimize a predefined function that is very useful in data clustering. Optimization techniques define a global function to capture the quality of best partitioning and will try to optimize a predefined function. The one optimization technique which is followed is bee colony optimization which is a nature inspired metaheuristic models [16]. The performance of the BCO algorithm has been compared with those of other well-known heuristic algorithms such as genetic algorithm, differential evolutionary algorithm, and particle swarm optimization algorithm for unconstrained optimization problems. The BCO algorithm belongs to the class of population-based techniques which is considered to be applied to find solutions for difficult combinatorial optimization problems. The major idea behind the BCO is to create the multi agent system which solves hard combinatorial optimization problems. The contribution made by the BCO algorithm is as follows.

- A basic bee colony-based clustering algorithm solves the clustering problem with the ancient BCO method. The basic algorithm has some problems regarding basic behaviors of the BCO algorithm that causes bee to follow one solution.
- An improved BCO algorithm by introducing cloning and fairness concepts into the BCO algorithm. The second proposed algorithm is based on the improved BCO method and referred to as IBCOCLUST which proposes a better modeling for the specific application of clustering.
- Hybrid clustering algorithms uses k-means and the IBCOCLUST algorithms. The hybrid methods improve the k-means algorithm by making it less dependent on the initial parameters.

## 3.3 MapReduce Based Fuzzy E-Mean Clustering Algorithm

The present clustering algorithms cannot handle big data and thus the scalable options are necessary. So, the fuzzy clustering algorithms outperform the hard-clustering approaches in terms of accuracy. The algorithm is applied in many areas such as digital image processing, image segmentation etc. Most analytical fuzzy clustering algorithms are based on the optimization of the basic c-means objective function, or some modification of the objective function [17]. The optimization of the c-means functional represents a nonlinear minimization problem, which can be solved by using a variety of methods including iterative minimization. The most popular method is to use the simple Picard iteration through the first-order conditions for stationary points, known as the Fuzzy C-Means (FCM) algorithm. An optimal c partition is produced iteratively by minimizing the weighted within group sum of squared error objective function. Parallelization of algorithms is needed in order to enable big data processing. In order to make the development of parallel applications easier, Google introduced a programming paradigm called MapReduce that uses the Map and Reduce primitives that are present in functional programming languages. In order to consider the mapping of the FCM algorithm to the map and reduce primitives, it is necessary for FCM to be partitioned into two jobs. The first MapReduce job calculates the centroid matrix by iterating over the data records.

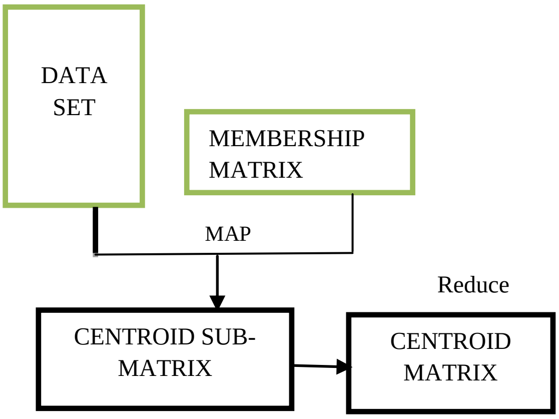

**Fig 2: First MapReduce job [17]**

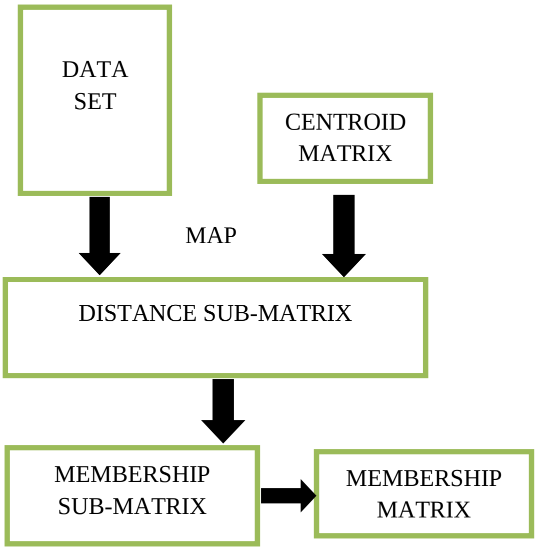

**Fig 3: Second MapReduce job [17]**

The second MapReduce job compared to the first involves more computations to be executed and is shown in the above figures. The accuracy of the MR-FCM algorithm was measured in terms of purity and compared to different clustering algorithms. Furthermore, the MR-FCM algorithm scales well with increasing data set sizes as shown by the scalability analysis conducted.

## 3.4 Text Similarity Algorithm

The text search similarity algorithms are subjected to two evaluation techniques. As the alignment algorithms are slow for searching a document collection, it works in two phases. At the first stage, a vector based TSS algorithm is used. In the preliminary processing step, word-count vectors are created from library documents. Non-essential ‘stop words’ are removed from the vector, as they are not useful in the IR process. At run-time, a query consisting of a paragraph of natural language text is submitted to the system. The query is immediately converted to a vector representation expanded by stemming and then compared with the vector representation of each library document by one of the several similarity functions. IN this comparison of several TSS algorithms with one another utilizing both user testing and industry-standard evaluation tools is compared

## 4. CLUSTERING AND MINING ALGORITHMS

Various clustering algorithms used for classification are proposed by the authors and is explained in the following section. Partitioning algorithms are widely used in the database literature in order to efficiently create clusters of objects. Hierarchical clustering is usually portrayed as the better clustering approach but is now reduced because of the its quadratic time complexity. There are two types of k clustering method and the new type of clustering such as bisecting k means is considered as the best performance for the measure of cluster quality. The bisecting K-means algorithm starts with a single cluster of all documents and has the following process,

- Picks up a cluster to split
- Finds out 2 sub clusters using the basic K -means algorithm
- Continue step 2 the bisecting step for few times and split that produces the clustering with high similarity.
- Repeat steps 1, 2, and 3 till the target cluster is reached.

The three different agglomerative hierarchical techniques for the clustering documents are intra-cluster similarity technique, centroid similarity technique and UPGMA. In Intra cluster similarity technique the similarity of all the documents in a cluster to the cluster centroid is observed. In centroid similarity it defines the similarity of two clusters to be the cosine similarity between the centroids of the two clusters Apart from these the naive Bayes classifier and support vector machine is used for the text classification and is used in clustering and these two classifier techniques are the best for text classification.

In k-medoid algorithm a set of points from the original data around the clusters are built. The main goal of the algorithm is to determine an optimal set of representative documents around which the clusters are built. Each document is assigned to the closest representative from the collection. The algorithm works with an iterative approach in which the set of k representatives are successively improved. One general disadvantage of k-medoids clustering algorithms is that they require a large number of iterations in order to achieve convergence and are therefore quite slow. The second key disadvantage is that k-medoid algorithms do not work very well for sparse data such as text [18]. Another clustering algorithm is k-means clustering algorithm which uses a set of k representatives around which the clusters are built. One of the advantages of the k-means method over the k-medoids method is that it requires an extremely small number of iterations in order to converge. The main disadvantage of the k-means method is that it is still quite sensitive to the initial set of seeds picked during the clustering [18].

The text classification in text mining is performed by naive Bayes classifier. The simplest model naives Bayes classifier is constructed by using the training data to estimate the probability of particular category given with the feature values of a new instance. The naive Bayes classifier is effective despite of the fact that assumption of conditional independence is generally not true for word appearance in documents [19]. The one more approach is support vector machines which have been recently gaining popularity in regards of learning community. An SVM is a hyperplane that separates the set of positive examples from a set of negative examples with the maximum margin and the below figure 4 shows the model.

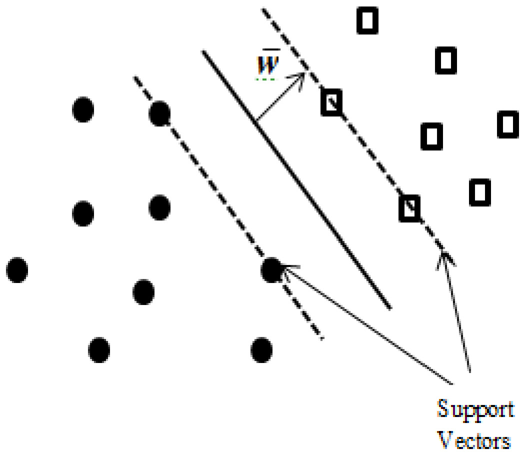

**Fig 4: Linear support vector machine [19]**

Training an SVM requires the solution of quadratic programming. The accuracy of the linear SVM is the best and the model is very simple. SVM'S work well as they create a classifier which maximizes the margin between the positive and negative examples [19]. SVM's are the most accurate classifier and linear SVM's are the best among all. Another approach of text clustering is by Hidden Markov Model. HMM is a simple case of dynamic Bayesian network, where the hidden states are forming a chain and only some possible value for each state can be observed. One goal of HMM is to infer the hidden states according to the observed values and their dependency relationships. A very important application of HMM is part-of-speech tagging in NLP. HMM is a simple case of dynamic Bayesian network, where the hidden states are forming a chain and only some possible value for each state can be observed. One goal of HMM is to infer the hidden states according to the observed values and their dependency relationships. A very important application of HMM is part-of-speech tagging in NLP [20]. Generally, an HMM can be considered as generalization of a mixture model where the hidden variables are related through a Markov process rather than independent of each other.

## 5. COMPARISON AND RESULTS

**Table 1. Table captions should be placed above the table**

| Sl No | Title | Techniqu e | Outcomes | Research Gap |
|---|---|---|---|---|
| 1 | Probabilistic models for text mining [20] | Hidden Markov Model | Helps in sequence analysis and in speech recognition | Computational time is very large |
| 2 | Inductive learning algorithms and representations for text categorization [19] | Support vector machine classifier | The training data is redundant and helps in supporting the boundary and is more efficient. | Choosing the appropriate kernel function is quite tricky |

| | | | | |
|---|---|---|---|---|
| 3 | A survey of text clustering algorithms [18] | K-means clustering | With a large number of variables. k-means may be computationally faster than hierarchical clustering | Difficult to predict the number of clusters (k-value) |
| 4 | MapReduce-based fuzzy c-means clustering algorithm [17] | Fuzzy c-means clustering algorithm | Helps in reducing the time taken for data and information retrieval from large dataset | Accuracy is not acceptable and needs to be improved. |

## 6. CONCLUSION

This paper presents a review of the document clustering and text mining algorithms with effective time consumptions. Also, this paper helps in determining the efficient algorithm for text classification and clustering. The techniques discussed in the review are support vector machines, k-means clustering, hidden markov model, fuzzy c means clustering and text similarity algorithm. A good clustering result requires appropriate similarity measure and proper selection of algorithm. Existing studies shows that the support vector machine classifier mining approach performs well and achieves accuracy more than the other techniques discussed. The support vector machine approach helps in deciding the probability of the text and outperforms the other classifiers. Therefore, an appropriate emphasis should be based on the accuracy rate and the computational time in all text mining and document clustering approaches

## 7. REFERENCES

[1] Gupta, Vishal, and Gurpreet S. Lehal. "A survey of text mining techniques and applications." Journal of emerging technologies in web intelligence 1, no. 1 (2009): 60-76

[2] Feldman, Ronen, and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)." In KDD, vol. 95, pp. 112-117. 1995.

[3] Sundari, D. Jasmine Guna, and D. Sundar. "A Study of Various Text Mining Techniques."

[4] Ghosh, Sayantani, Sudipta Roy, and Samir K. Bandyopadhyay. "A tutorial review on Text Mining Algorithms." International Journal of Advanced Research in Computer and Communication Engineering 1, no. 4 (2012): 7.

[5] Bisht, Sunita, and Amit Paul. "Document clustering: a review." International Journal of Computer Applications 73, no. 11 (2013).

[6] Lian, Wang, Nikos Mamoulis, and Siu-Ming Yiu. "An efficient and scalable algorithm for clustering XML documents by structure." IEEE transactions on Knowledge and Data Engineering 16, no. 1 (2004): 82-96.

[7] https://en.wikipedia.org/wiki/Document_Clustering\

[8] Shah, Neepa, and Sunita Mahajan. "Document clustering: a detailed review." International Journal of Applied Information Systems 4, no. 5 (2012): 30-38.

[9] Chaurasia, Vikas, and Saurabh Pal. "A novel approach for breast cancer detection using data mining techniques." (2017).

[10] Shen, Liyin, Hang Yan, Hongqin Fan, Ya Wu, and Yu Zhang. "An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design." Building and Environment 124 (2017): 388-401.Cutting, D. R., Karger, D. R., Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (2017, August). Scatter/gather: A cluster-based approach to browsing large document collections. In ACM SIGIR Forum (Vol. 51, No. 2, pp. 148-159). ACM.

[11] Turtle, Howard, and W. Bruce Croft. "Inference networks for document retrieval." In ACM SIGIR Forum, vol. 51, no. 2, pp. 124-147. ACM, 2017.

[12] Desjardins, Guy, and Robert Godin. "Combining relevance feedback and genetic algorithms in an internet information filtering engine." In Content-Based Multimedia Information Access-Volume 2, pp. 1676-1685. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

[13] Forsati, Rana, Andisheh Keikha, and Mehrnoush Shamsfard. "An improved bee colony optimization algorithm with an application to document clustering." Neurocomputing 159 (2015): 9-26.

[14] Lučić, Panta, and Dušan Teodorović. "Computing with bees: attacking complex transportation engineering problems." International Journal on Artificial Intelligence Tools 12, no. 03 (2003): 375-394.

[15] Ludwig, Simone A. "MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability." International journal of machine learning and cybernetics 6, no. 6 (2015): 923-934.

[16] Dumais, Susan, John Platt, David Heckerman, and Mehran Sahami. "Inductive learning algorithms and representations for text categorization." In Proceedings of the seventh international conference on Information and knowledge management, pp. 148-155. ACM, 1998.

[17] Aggarwal, C. C., and C. Zhai. "Probabilistic Models for Text Mining: In Mining Text Data." (2012): 257-294.

[18] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." In KDD workshop on text mining, vol. 400, no. 1, pp. 525-526. 2000.

[19] Yu, Hwanjo, and Sungchul Kim. "SVM tutorial—classification, regression and ranking." In Handbook of Natural computing, pp. 479-506. Springer, Berlin, Heidelberg, 2012.

[20] Teh, Yee W., Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Sharing clusters among related groups: Hierarchical Dirichlet processes." In Advances in neural information processing systems, pp. 1385-1392. 2005.