American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

A Modified Hierarchical Agglomerative Approach for Efficient Document Clustering System

May Thu Lwin[a]*, Moe Moe Aye[b]

[a,b]*Department of Computer Engineering and Information Technology, MTU, Mandalay, Myanmar*
[a]*Email: maythu.chocolate@gmail.com*
[b]*Email: moeaye255@gmail.com*

Abstract

In today's world, the increasing volume of text documents has brought challenges for their effective and efficient organization. This has led to an enormous demand for efficient tools that turn data into valuable knowledge. One of the techniques that can play an important role towards the achievement of this objective is document clustering. The main function of document clustering is automatic grouping of documents so that the documents within a cluster are very similar, but dissimilar to the documents in other clusters. This research proposes a Modified Agglomerative Hierarchical Clustering (MAHC) algorithm based on hierarchical method. In many traditional systems, the number of term frequency is considered to create data representation matrix. However, a modified algorithm creates data representation matrix based only on occurrence of items, not on frequency of items. The proposed algorithm can increase the quality of clustering because it can merge the related or similar documents into the same cluster efficiently. Moreover, the proposed algorithm can reduce the processing time than the existing methods. In this paper, the performance of clustering between the proposed and original clustering algorithm was compared and evaluated by using F-measure.

Keywords: Document Clustering; Agglomerative Hierarchical Clustering (AHC) algorithm; Similarity Measures; F-measure; Optimized Bubble Sort Algorithm.

1. Introduction

Nowadays, the internet has become the largest data repository, facing the problem of information overload.

---

* Corresponding author.

The existence of an abundance of information makes a tedious process in searching information for the average user. This has led to an enormous demand for efficient tools that turn data into valuable knowledge. Researchers from numerous technological areas, namely, pattern recognition, machine learning, data mining etc. have been searching for eminent approaches to fulfill this requirement. As a result, document clustering system plays a vital role towards this achievement. Document clustering is an unsupervised approach of data mining. Document clustering groups similar documents to form a coherent cluster, while documents that are different have separated apart into different clusters. Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis [1]. Many algorithms are available in the literature for performing data clustering. Out of these, two major categories of algorithms are commonly used for document clustering are: "Partitioning" and "Hierarchical". Partitioning clustering algorithm divides the documents into fixed partitions, where each partition represents a cluster. The commonly used partitioned clustering technique is k-means algorithm, where k is the desired number of clusters. The disadvantage of this method is that the number of clusters is fixed and it is very difficult to select a valid k for an unknown data set. Hierarchical clustering produces a hierarchical tree of clusters called dendrogram. The hierarchical clustering techniques can be divided into two parts - agglomerative and divisive. In Agglomerative Hierarchical Clustering (AHC) method, starting with each data point as individual cluster, at each step, it merges the most similar clusters until a given termination condition is satisfied. In Divisive Hierarchical Clustering (DHC) method, starting with the whole set of data points as a single cluster, the method splits a cluster into smaller clusters at each step until a given termination condition is satisfied. The time complexity of most of the hierarchical clustering algorithms is quadratic and the algorithm can never undo what was done previously [2]. This paper presents a modified algorithm based on an agglomerative hierarch approach for document clustering system. This paper is organized as: Section II presents a brief survey of various techniques used for document clustering so far. Section III provides procedures and algorithms used in document clustering. Section IV explains about the process of the proposed work. Experimental results and discussions are shown in Section V. Finally, the conclusion of this paper and future work is in Section VI.

2. Literature Review

For a long time the concept of clustering has been around. It has more than a few applications, mainly in the situation of information retrieval and in organizing web possessions. The research in clustering ultimately goes ahead to automatic indexing to index as well as to recover the electronic proceedings. The crucial intend of clustering is to supply a grouping of similar records. Recently, T. Su and C. A. Murthy [3] proposed a new hierarchical approach, Clustering Using Extensive Similarity (CUES) measure, for document clustering. It introduced a new document similarity using extensive similarity measure. In this approach, two documents are considered to be similar if they share minimum number of common words and they have almost same distance with every other document in the corpus. There are three features of the proposed clustering method. First, the algorithm can identify two dissimilar clusters and will never merge them. Second, the algorithm can be stopped if the distance between two clusters becomes very high, since at each step CUES checks the cluster distance to merge two clusters. Third, there is no need to input the desired number of cluster prior to implement the algorithm. It is experimentally found on several test data sets that the proposed algorithm performs significantly better than the traditional document clustering techniques according to F-measure and normalized mutual

information (NMI). M. Paul and P. Thangam [4] proposed a modified hierarchical clustering algorithm by using multiviewpoint-based similarity measure (MVS). The MVS uses the different viewpoints unlike the traditional similarity measure that uses only a single viewpoint. The MVS increases the accuracy of clustering than the traditional similarity measures. They compared the proposed work with k-means clustering using MVS and found that the performance of hierarchical clustering using MVS is higher.

3. Document clustering system

In document clustering system, many researchers apply various clustering algorithms to build clusters of similar documents on their purposes. To apply any clustering method on text documents, researchers need to follow several steps which is shown in figure 1.

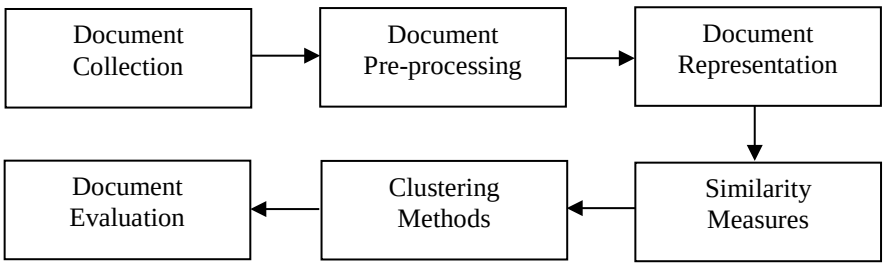

Figure 1: various steps used in document clustering

Document collection: To perform clustering process, the first input is the document folder or document datasets. Each document consists of several lines and each line further contains several words. Document pre-processing: The input to this step is a plain text document and the output is a set of pre-processed tokens. The concept of pre-processing is used to prune all the words and terms with poor information. In this process, the two popular methods namely, stop word removal and stemming algorithm, are used. Stop words are words that carry no information and meaningless. (i.e., Pronouns, Prepositions, Conjunctions,). Stop word removal removes stop linking words like “if”, “but”, “the”, “to”, “also”, “then” from the documents. Stop words may be eliminated using a list of stop words. The second process is stemming a word. Stemming is the process of reducing words to their stem or root form. Stemming also removed prefixes and suffixes of each word. Words are stemmed using Porter’s suffix-stripping algorithm. For example, *production, produce, produces and product* will be mapped to the stem *produc* [5]. Document representation: After pre-processing, documents need to be represented into some mathematical form. In this work, documents are converted into vector space model (VSM). Under VSM, n documents with m words are represented as n × m document by word matrix. The three different representations proposed and widely used in research area are word, term and N-gram representation [6]. Document similarity measures: The nature of similarity measures plays an important role in the failure or success of clustering methods. Since the performance of the clustering system relies on the choice of an appropriate measure, many researchers have taken elaborate efforts to find the most meaningful similarity measures over a hundred years. Common distance/similarity measures used in document clustering are Cosine similarity, Euclidean distance, Jaccard coefficient, Pearson correlation and so on [7].

Clustering methods: The goal of clustering is to reduce the amount of data by categorizing or grouping similar data items together. Clustering methods can be classified into six categories: partitioning method, hierarchical method, density-based method, grid-based method, model-based method and constraint based method [2]. Document evaluation: To measure the cluster quality and goodness, document evaluation can be divided into two approaches, internal quality measure and external quality measure. The function of internal quality measure is to attain high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar). The external quality measures commonly used in document clustering are purity, entropy, F-measure and so on [8].

4. Agglomerative hierarchical clustering techniques

Many clustering algorithms are available in literature, but hierarchical clustering methods are applied in this work. Hierarchical clustering method can be classified into two approaches, namely, agglomerative hierarchical clustering (AHC) and divisive hierarchical clustering (DHC). AHC, which has bottom-up approach, is commonly used in clustering problems than DHC which has top-down approach. In AHC method, each observation starts in its own cluster and pairs of the clusters are merged as one move up the hierarchy. This process is repeated until a minimum number of clusters have been reached, or, if a complete hierarchy is required then the process continues until only one cluster is left. The procedure of AHC algorithm [2] is as follows:

1. Compute the similarity between all pairs of clusters, i.e., calculate a similarity matrix whose ij^{th} entry gives the similarity between the i^{th} and j^{th} clusters.
2. Merge the most similar (closest) two clusters.
3. Update the similarity matrix to reflect the pairwise similarity between the new cluster and the original clusters
4. Repeat steps 2 and 3 until only a single cluster remains.

Hierarchical clustering initializes a cluster system as a set of singleton clusters (agglomerative case) or a single cluster of all points (divisive case) and proceeds iteratively with merging or splitting of the most appropriate clusters until the stopping criterion is achieved. The appropriateness of clusters for merging/splitting depends on the similarity/dissimilarity of cluster elements. To merge or split subsets of points rather than individual points, the distance between individual points, the distance between individual points has to be generalized to the distance between subsets. Such derived proximity measure is called a linkage metrics [9]. This method is very simple but needs to specify how to compute the distance between two clusters. Three commonly used methods for computing this distance are listed below: Single linkage method: The similarity between a pair of clusters is the maximum of the similarities between all pairs of documents such that one document is in one cluster and the other document is in other cluster. This method is also called “nearest neighbor” clustering method as shown in equation 1.

$$d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \left| P - P' \right| \qquad (1)$$

Complete linkage method: The similarity between a pair of clusters is calculated as the minimum of the similarities between all pairs of documents. This method is also called “furthest neighbor” clustering method and it is defined in equation 2.

$$d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \left| P - P' \right| \qquad (2)$$

Average linkage method: This method process clusters such that each document in a cluster has greater average similarity with the other documents in the cluster than with the documents in any other cluster. This method takes into account all possible pairs of distances between the objects in the clusters, and is considered more reliable and robust to outliers. This method is also known as Unweighted Pair Group Method using Arithmetic means (UPGMA) and it is defined in equation 3.

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} \left| p - p' \right| \quad (3)$$

where $\left| p - p' \right|$ represents the distance between two objects, p and p' and n_i and n_j represents the number of objects in clusters, C_i and C_j.

5. Optimized bubble sort algorithm

The bubble sort (BS), sometimes referred to as sinking sort, is a simple and straightforward sorting algorithm. It starts with compare with the first two elements and if the first element is greater than the second then swaps it. It continues for each pair of elements to the end of data set. It again starts with the first two elements and repeating until no swap has occurred in the last pass. The algorithm gets its name from the way smaller elements “bubble” to the top of the list. The positions of the elements in bubble sort play an important role in determining its performance. Large elements at the beginning of the list are quickly swapped, while small elements at the beginning move to the top extremely slowly. This has led to these types of elements being named rabbits and turtles, respectively. In optimized bubble sort (OBS), after every pass, all the element after the last swap are sorted, and do not need to checked. The difference between BS and OBS may not be clear with a small size of input array, but with a large size it is very clear that OBS is better than BS.

The procedure of this algorithm is as follows:

1. Procedure bubble sort (A: list of sortable items) [10]
2. n = length (A)
3. repeat
4. swapped = false;

5. for i=1 to n-1 inclusive do
6. if A[i-1] > A[i] then
7. swap (A[i-1], A[i])
8. swapped = true;
9. end if
10. end for
11. n = n-1
12. until not swapped
13. end Procedure

6. The proposed method

In many traditional systems, data representation matrix is constructed based on the number of word frequency. After that, document similarity between pairs of document is calculated and cluster merging process is performed by using linkage methods. However, the main idea of the proposed clustering algorithm is to construct the data representation matrix based on the items in documents without considering the count of words/ terms. In the proposed algorithm, the linkage methods used for cluster merging process in the traditional AHC algorithm are not needed to perform. Instead of this, Jaccard coefficient is used for calculating the similarity between pairs of documents. Based on Jaccard's similarity scores, the closest documents are merged into the same cluster. The procedure of the proposed MAHC algorithm is as follows:

1. Accept k-input number.
2. Sort all words in each document according to lexical order.
3. Sort all documents in ascending order by document-length.
4. If $(X \cap Y) = X$ or $(X \cap Y) = Y$ then merge these closest clusters.

   Else (i) Calculate similarity measure.

   (ii) Merge the most similar (closest) two clusters.
5. Repeat step 3 and 4 until k-clusters.

At first, the number of cluster, k, is specified as input parameter. In the next step, overlapping words in each

document are removed because the number of occurrence of word is neglected. After removing the overlapping words, all words in each document are sorted into the lexical order using Optimized Bubble Sort algorithm. In step 3, all documents are sorted by document-length in ascending order by using Optimized Bubble Sort algorithm again. Based on the ascending order, the uppermost document is the shortest length document and the lowermost is the longest one. In step 4, if the shortest length document is subset of one of all documents, this two subset documents are merged into the same cluster. If not, the similarity between the shortest document and all other documents is calculated by using Jaccard coefficient and then two most similar documents are merged into the same cluster according to Jaccard similarity scores. Finally, step 3 and 4 are repeated until k-cluster is reached.

7. Evaluation metrics

F-measure [8] is harmonic mean of precision and recall. F-measure is commonly used in evaluating the effectiveness of clustering and classification algorithms. Let C = $\{C_1, C_2,\ldots, C_k\}$ be a clustering of document set D, $C^*=\{C_1^*, C_2^*,\ldots, C_l^*\}$ designate the "correct" class set of D.

The recall of cluster j with respect to class i, recall (i, j) is defined as

$$recall(i, j) = \frac{\left|C_j \cap C_i^*\right|}{\left|C_i^*\right|} \tag{4}$$

Then the precision of cluster *j* with respect to class *i*, precision *(i, j)* is also defined as

$$precision(i, j) = \frac{\left|C_j \cap C_i^*\right|}{\left|C_i\right|} \tag{5}$$

The F-measure combines both values according to the following formula,

$$F(i, j) = \frac{2 * precision(i, j) * recall(i, j)}{precision(i, j) + recall(i, j)} \tag{6}$$

Based on this formula, the F-measure for overall quality of cluster set C is defined by the following formula,

$$F = \sum_{i=1}^{l} \frac{\left|C_i^*\right|}{D} \max_{j=1,2,\ldots,k} \{F(i, j)\} \tag{7}$$

A perfect clustering solution will be the one in which every class has a corresponding cluster containing the exactly same documents in the resulting hierarchical tree, in which case the F-measure will be one. In general, the higher the overall F-measure values, the better the clustering solution is.

8. The experimental results and discussion

All the experiments are carried out by using mini-newsgroups dataset. Mini-newsgroups datasets is a subset of popular dataset 20-newsgroups dataset [11], which is a collection of news articles collected from 20 different sources. Three categories of this dataset: alt.atheism, comp.graphics, comps.os.ms-windows.misc, are considered in the experiments. As experiment 1, the different numbers of documents (50, 100 and 150) are considered to evaluate the clustering performance between the traditional AHC and MAHC. The number of documents is randomly selected from each category. The graphical representation of comparison between the two clustering algorithms is shown in terms of F-measure in figure 3 and in terms of time in figure 4.

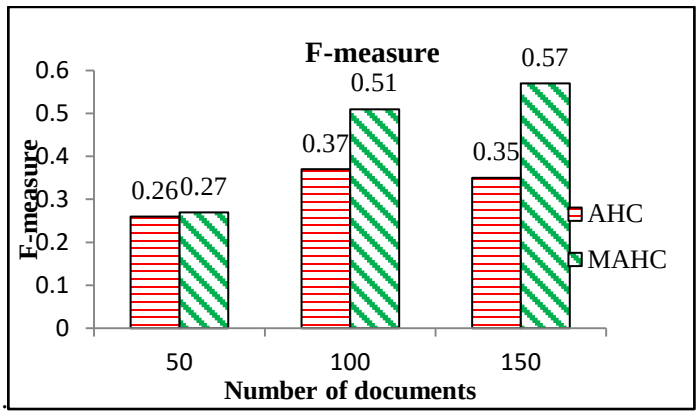

Figure 2: F-measure scores for two clustering methods

According to figure 2, for 50 documents, the F-measure scores between two clustering algorithms are similar. When the total number of documents is 100, the F-measure score of MAHC is significantly increased. In this case, it can be clear that the proposed MAHC algorithm can merge the similar documents into the same cluster efficiently than the traditional AHC. When the total number of documents is 150, the F-measure score of MAHC is still climbed up, but the F-measure score of AHC is slightly decreased. Therefore, it can be concluded MAHC can provide the better performance results than AHC even when the document sets become gradually higher.

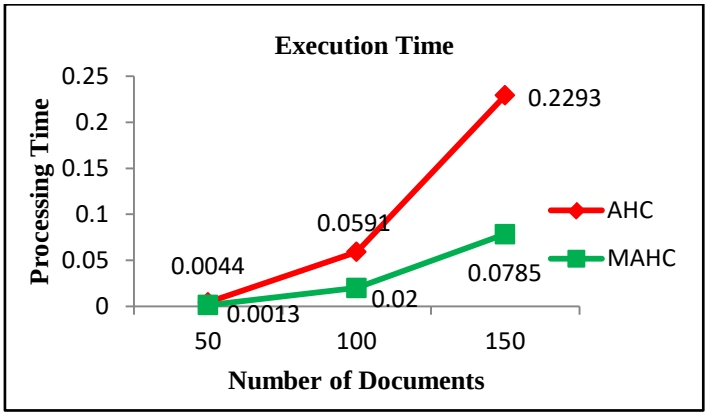

Figure 3: processing time for two clustering methods

In figure 3, the proposed MAHC algorithm reduces the computational time into the nearly one-third of the traditional AHC algorithm on the different document sets (50, 100 and 150). As experiment 2, only 300 documents are considered. 100 documents are randomly selected from each category. The graphical representation of comparison between the two clustering algorithms is shown in terms of F-measure in figure 4 and in terms of time in figure 5.

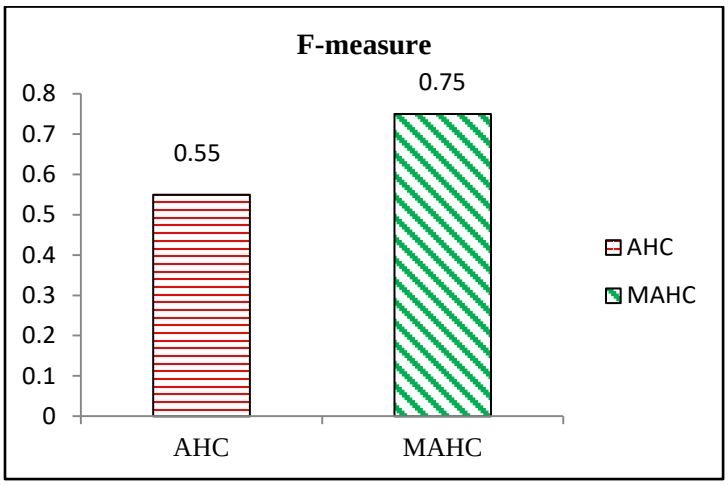

Figure 4: F-measure scores for two clustering methods

Figure 5: processing time for two clustering methods

According to figure 4 and 5, the proposed MAHC algorithm performs significantly better than the traditional AHC algorithm in terms of F-measure and the execution time. The proposed MAHC reduces the computational time into double than the traditional AHC algorithm and provides better cluster quality on the larger document set.

9. Conclusion and further extension

In this paper, a new modified hierarchical clustering algorithm for document clustering has been presented.

Existing systems consider the frequency of words/terms in documents to construct the data representation matrix. However, the focal point of this algorithm is to construct data representation matrix mainly based on the occurrence of word, not on the word frequency. Optimized Bubble Sort is used for sorting both of lexical order and document-length ascending order. By sorting this, the proposed algorithm does not need to perform linkage methods used in existing systems for cluster merging process. The sorted documents are used to calculate the document similarity by using Jaccard coefficient. Jaccard coefficient is one of the binary similarity measures and is suitable sufficiently to be employed in the word similarity measurement. To measure the efficiency and effectiveness of the proposed algorithm, the proposed algorithm is compared with the traditional AHC algorithm. The experimental results show that the proposed algorithm outperforms the clustering results than the traditional AHC algorithm in terms of execution time and F-measures. In future, the experiment can be measured by finding semantic relatedness between documents instead of document similarity and can also be conducted for large volume of datasets.

Acknowledgements

The author would like to especially express her deep appreciation to her supervisor, Dr. Moe Moe Aye, Associate Professor, Department of Computer Engineering and Information Technology, Mandalay Technological University, for her close supervision, helpful advice, encouragement and invaluable guidance. The author would also thank to her parents, all her friends and all the teachers who taught her throughout the whole life.

References

[1] J. Han and M. Kamber, "Data Mining concepts and techniques", Morgan Kaufmann Publishers, Second edition, 2009.

[2] R. Xu, “Survey of clustering algorithms”, IEEE Transactions on Neural Network, Vol.16, No.3, May, 2005.

[3] T. Su and C. A. Murthy, “A new hierarchical approach for document clustering”, Journal of Pattern Recognition Research, 66-84, August, 2013.

[4] M. Paul and P. Thangam, “A modified hierarchical clustering algorithm for document clustering algorithms”, International Journal of Advanced Research in Computer Engineering and Technology, IJARCET, Vol.2, Issue-6, June, 2013.

[5] Pushplasta and R. Chatterjee, “Analytical assessment on document clustering”, International Journal of Computer Network and Information Security, IJCNIS, Vol.5, 63-71, June, 2012.

[6] M. Shafiei and S. Wang and et.al, “Document representation and dimension reduction for text clustering”, IEEE 23rd International Conference on Data Engineering Workshop, April, 2017.

[7] A. Haung, "Similarity measures for text document clustering", Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC), April, 2008

[8] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets", CIKM, Virginia, USA, November 4-9, 2002.

[9] P. Berkhin, "Survey of clustering data mining techniques"

[10] J. Alnihoud and R. Mansi, "An enhancement of major sorting algorithm", The International Arab Journal of Information Technology", Vol.7, No-1, January, 2010.

[11] website: https://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html., last date accessed 2\2\2017.