International Journal of Control and Automation
Vol. 13, No. 2s, (2020), pp. 151-164

# Weighted Frequent Pattern based Agglomerative Clustering for large unstructured text data

K.V.Kanimozhi[1], Dr.Rajakumarkrishnan[2], Dr.M.Venkatesan[3]

[12] *VIT University, Vellore, 632404, India*
[3] *NIT Surathkal, 575025, India.*

***Abstract***

*Processing large amount of text using traditional clustering methods are key challenges.Research communities have proposed the various clustering approaches for analyzing unstructured data. Frequent item based clustering method is one of the mostly used clustering for text analytic domain. An approach based on Frequent Weighted Utility Itemsets (FWUI) and then clustering using the MC (Maximum Capturing) algorithm is one of the most effective methods for text clustering. However, the Maximum Capturing clusteringAlgorithm based on the similarity matrix leads to a lot of irrelevant clusters that aren't desired. In this work, Weighted Frequent Pattern based Agglomerative Clustering(WFUP_AC)is proposed for clustering large text data.First, the Term Frequency (TF) is calculated for each term in the documents to create a weight matrix for all documents. The weights of terms in documents are based on the Inverse Document Frequency. The WFUP algorithm is applied for mining Weighted Frequent Utility Pattern (WFUP) from a number matrix and the weights of terms in documents. Then based on frequent utility itemsets, a similarity matrix is obtained for each document where each entry equals to common frequent itemset between two documents. Then distance matrix is calculated from the similarity matrix, finally Hierarchical Agglomerative Clustering method is applied on the Distance matrix using complete linkage and cut the dendrogram as per the need. Our proposed method has been evaluated on two text document data sets like newsgroup and Reuters data sets with different size consisting of 100,300,500 and 1000 documents. The experimental results show that our method, weighted frequent pattern based agglomerative clustering (WFUP_AC) improves the accuracy of the text clustering compared to MC clustering methods using FIs(Frequent Itemset) and FWUIs.*

*Keywords: Text Clustering, Frequent Pattern Mining, Minimum Support, Agglomerative Clustering, Unstructured data*

## 1. Introduction

As the number of Internet users surges every year, text information on the Internet has exploded and network is flooded with vast amounts of textual data.

Text clustering technology for massive web content has received widespread attention.

Text clustering is described as one of the most efficient techniques used in text mining domain, machine learning, and pattern recognition. With the rapid increase in the amount of electronic information on internet web pages and modern applications, text analysis requires complex techniques to deal with numerous text documents. Typically the Vector Space Model is used to represent text documents. Hence the documents are represented in a multi-dimensional space. The position value of each dimension corresponds to a weight value. Text clustering algorithms alone do not perform any feature selection. This leads to avast number of features being used for text document clustering, which also includes a lot of uninformative text features.

Since text documents contain a lot of uninformative, redundant, unevenly distributed and noisy features, a feature selection is required be for a clustering algorithm can be applied. A feature selection algorithm aims to determine the most informative features in text documents. This not only removes uninformative features but also reduces the complexity of the clustering algorithms, and yields better clusters of text documents.

ISSN: 2005-4297 IJCA

Text clustering can be formally defined as the application of cluster analysis on textual documents. It uses unsupervised learning and natural language processing (NLP) to understand and categorize unstructured, textual data. Automatic document organization, topic extraction, information retrieval, and filtering are some of the key applications oftext clustering. All these require text clustering (sometimes also known as document clustering) to be done quickly and accurately. Due to its important roles in many applications text clustering is widely studied in text mining. In text clustering, feature extraction is an essential process for creating an accurate clustering model. Vector space model (VSM), the term frequency-inverse document frequency (TF-IDF) measure, and latent semantic analysis (LSA) are widely used in traditional text clustering to represent documents. However, these methods are confronted with high data sparseness and complex semantics, as well as large noise interference. Some researchers extended semantics of words through external links or world knowledge bases, while these methods would suffer from inefficiency when processing large-scale corpora.

There have been various techniques used to do text clustering. Earlier, researchers used probabilistic models to do the task. Later, the focus shifted towards Frequent Term based clustering and lately they focused on clustering using frequent pattern mining and maximum capturing. In this article, we propose a novel approach to do the task using a combination of frequent pattern mining and hierarchical clustering. Experiments were performed on standard 20 newsgroup datasets with varying amounts of documents to accommodate scalability. Results show that the f-score obtained from this approach outperforms the result obtained using the maximum capturing algorithm.The rest of the paper is organized as follows. In Section 2, briefly discuss about the related work. Proposed Methodology is discussed in Section 3, which comprises of preprocessing step, the WFUP algorithm, and the Agglomerative Clustering Algorithm(WFUP_AC). In Section 4, Results and Analysis are discussed and concluded the paper in Section 5.

## 2. Related Work

Researchers have exploited various ways of text clustering, including the use of popular text-domain clustering algorithms, the use of the essence of word patterns/context, and probabilistic approaches[1].Frequent Term based clustering[3] is an approach to do text clustering that provides higher accuracy and faster processing compared to bisecting K-means[9]. This research inspired a new line of text clustering methods with some modifications, Total Sequence Clustering (CMS) and Frequent Word Sequence Clustering (CFWS), etc. However, there were drawbacks of using these approaches as FTC caused isolated documents, FIHC couldn’tresolve cluster conflicts, CMS depended on the efficacy of information representation and CFWS yielded insignificant clustering outcomes. To overcome these problems frequent item- sets (FIs) with the Maximum Capture (MC) approach were used. Later, the weights of terms were also incorporated to improve the performance of MC.Term Frequency-Inverse Document Frequency (TF-IDF)[7] as weights from a list of text documents to mine Frequent weighted utility itemsets (FWUIs)[10][5].Then the resulting FWUIs were used for text clustering with the MC approach. This approach enhanced the output of the MC.

A model for clustering geographic locations is proposed in [13] based on geo social network data with the inclusion of extension towards temporal information derived from checkins. Text document clustering is as discussed in [14] which includes three features using feature selection algorithms with feature weight and dynamic dimension reduction for the text clustering problem.The paper also uses Genetic algorithm, harmony search algorithm and particle swarm optimization algorithm where in the number of features are used to improve the performance using a dynamic reduction method.Text classification in natural language processing is proposed in [15] by building a graph convolutional network based on document word relations and word co-occurrence called as text graph convolutional network for the corpus using various classification techniques. A neural feedback clustering algorithm is demonstrated in [16] with the combination of bidirectional long short memory and convolutional neural network with k means clustering technique which incorporates feature extraction and clustering as a united process where

ISSN: 2005-4297 IJCA

clustering results are used as feedback information to dynamically perform optimization of the parameters of the networks.

Text document clustering based on frequent word meaning sequences as proposed in [17] with clustering based on frequent word sequences and clustering based on frequent word meaning sequences. A word (meaning) sequence is considered frequent if it occurs more than a certain percentage of the documents in the text database. The frequent word meaning sequences provides valuable and compact information about the text documents. A fast and effective cluster based information retrieval using frequent closed itemsets is proposed in [18] with an intelligent cluster based information retrieval which combines k means clustering with frequent closed itemset mining for the extraction of document clusters and also to find frequent terms in each cluster .The patterns which are discovered in each of the cluster are then used to select the most relevant document cluster to answer each of the query asked by the user. A fast density clustering algorithm for dynamic data streams as proposed in [19] uses ant colony stream clustering which demonstrates an online, bio inspired approach for clustering dynamic data streams. The proposed algorithm uses ant colony stream clustering algorithm which is a densitybased clustering algorithm where by the clusters are determined as high density areas of the feature space separated by low density areas. The tumbling window is used to accept a stream and rough clusters are incrementally formed during the single pass of the given window. This approach also identifies clusters as group of microclusters. A density based spatial clustering of applications with spatial textual information on social media is proposed in [20] which integrates the existing DBSCAN algorithm and the heterogeneous textual information to avoid noisy regions having numerous POI-irrelevant geo tags.
Geo social clustering of places from check in data is discussed in [21] to cluster places not only based on their locations but it also included semantics. The check ins are introduced which provides insights into the community structure of the people who are visiting the given place which is leveraged and also integrated into the proposed geo social clustering framework called a GeoScoop. The paper also used extensive versions of iterative procedure of expectation maximization and the DBSCAN algorithm.A grid based DBSCAN algorithm for clustering extended objects in radar data is proposed in [22] as the algorithm is modified to deal with non-equidistant sampling density and the clutter of the radar data.To be robust against the clutter the paper also uses varying sampling resolution to perform an optimized separation of objects at the same given time. Density based spatial clustering on twitter data is also discussed in [23] where POI (point of interest) tags are considered as POI relevant and POI irrelevant tweets,with the density based algorithm proposed shows much higher clustering quality than the DBSCAN in terms of F1 score and its variants. Density based clustering using Geo geo social network data is proposed in [24] which show how the density based clustering algorithm can be extended to be applied on locations which are being visited by the users of the geo social network. The paper uses spatio temporal information and considers social relationships between the users who visited the clustered places. A combine clustering and frequent itemsets mining to enhance biomedical text summarization is discussed in [25] which proposes a novel biomedical text summarization system that combines two data mining techniques i.e., clustering and frequent itemsets mining. The paper uses kmeans algorithm to cluster similar sentences and later the apriori algorithm is applied to determine the frequent itemsets among the clustered sentences. The various features from each cluster are selected to build the entire summary using the discovered frequent itemsets.

## 3. Weighted Frequent Pattern based Agglomerative Clustering

The unsupervised learning techniques such as frequent pattern mining and clustering are integrated together and developed novel frequent patternbased clustering to analyze large amount of unstructured text data. The concept of frequent pattern originates from association rule mining which uses frequent pattern to find association rules of items in large transactional databases. A frequent pattern is a set of frequent items, which co-occur in transactions more than a given threshold value called minimum support. Recent studies on frequent pattern in text mining fall into two categories. One is to use association rules to conduct text categorization and the other one is to use frequent pattern for text

ISSN: 2005-4297 IJCA

clustering. In this paper weighted frequent pattern based agglomerative text clustering is proposed process large amount of text data.

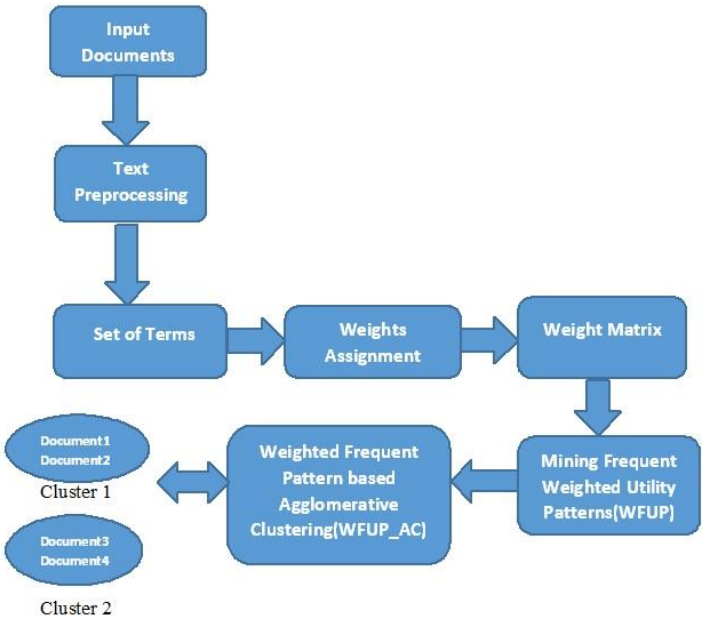


Figure.1 Flow of Weighted Frequent Patten based Agglomerative Clustering

Figure 1.shows the flow of weighted frequent item set based agglomerative clustering for clustering text documents. Here the input is text documents which are preprocessed using the text preprocessing methods and set of terms are generated. Based on the importance of term in the document, a weight is assigned to term. Then the weight matrixes of all the terms are computed. Third, frequent utility patterns are extracted from the weight matrix as the benefit for the contents of documents. Finally, agglomerative clustering is applied to cluster the documents based on the frequent patterns.

### 3.1 Preprocessing of Text

Preprocessing step is utilized to change each archive into a set of words. Right off the bat tokenization is performed which changes our documents into a set of tokens. Tokenization is the procedure of tokenizing or parting a string, content into a rundown of tokens. After Tokenization ,Lemmatization and Stemming were applied to our corpus of words. Lemmatization is the way toward changing over a word to its base form. Lemmatization considers the unique circumstance and changes over the word to its significant base structure which really exists in the language dictionary. Word net lemmatizer is used to lemmatize the words having a place with our corpus. With stemming, words are diminished to their promise stems. A word stem need not be a similar root as a lexicon-based morphological root, it simply is an equivalent to or littler type of the word. Porter stemmer is utilized for stemming. Stop words are expelled from our arrangement of Tokens the same number of those simply exist to improve the syntactic feeling of the sentence and aren't useful enough to give the specific circumstance. The terms present in the documents are converted in to vector using TF-IDF.

The TF-IDF [11] of a word is a score indicating the importance of that word/term in a document with regard to a collection of documents. This score is the product of Term Frequency (TF) and Inverse Document Frequency (IDF).Term Frequency (TF), annotated as tf(t, d), is the number of occurrences of a term t in a document d, as computed by the following formula:

$$tf(t,d) = \frac{n(t,d)}{n(d)} \qquad (1)$$

where n (t, d) is the occurrences of term t in document d and n (d) is the total number of occurrences of all terms in document d.

IDF, annotated as idf (t, D), measures the informativeness of the term t in a collection of corpus D. It is calculated as the logarithmically scaled inverse fraction of the number of documents in a corpus that contain the term t.

$$idf(t, D) = log \frac{|D|}{|\{d\ D|t\ d\}|}$$

where |D| is the number of documents in D and df (t, D) is the number of documents in D containing term t.

$$df(t, D) = |\{d\ D|t\ d\}|$$

IDF score of a term indicates the importance in the collection of documents D, i.e., rare terms have high scores and frequent terms have low scores.

### 3.2 MiningWeightedFrequentUtilityPatterns(WFUP)

Associate rule mining [2] was proposed by Agrawal and Srikanth to mine frequent pattern from the large transaction data. The association rules were generated from the frequent patterns using Apriori algorithm. They have proposed two user defined threshold like min-support for frequent item set and min-confidence for generating association rules. Frequent pattern based association rule mining was proposed by Han, Pei, and Yin to mine frequent patterns without generation of candidate itemset , thus saving memory and improves the processing time. Weighted Frequent UtilityPattern(WFUP) algorithm is applied for mining frequent utility itemsets in a weighted matrix. Two factors transaction weighted utility (twu) and weighted utility support are important for theWFUPalgorithm.

The Transaction Weighted Utility[12]twu of a transaction tk is calculated as follows

$$twu(t_k) = \frac{\sum_{i,j \in S(t_k)} w_j * x_{k_{ij}}}{|t_k|}$$

where $x_{k_{ij}}$ is the quantity of item ij,$w_j$is the weight of item ijin transaction tk and |tk| isthe total number of items in transaction tk.twu score of each documenttransactionis computed using the equation (3). Maximum vocabulary (distinct tokens) are limited to 10000 most frequent words to ignore words having very less frequency in the overall corpus of the document dataset. It also helps for efficient calculation of Frequent Itemset mining which would take very large time to compute because of the exponential nature of the algorithm.

The weighted utility support wus of an itemset X is calculated using equation (4)

$$wus(x) = \frac{\sum_{t_k \in t(x)} twu(t_k)}{\sum_{t_k \in t} twu(t_k)} \qquad (4)$$

TheWeighted Frequent Utility Pattern(WFUP) algorithm is described below.

---

ISSN: 2005-4297 IJCA

Algorithm: Weighted Frequent Utility Pattern(WFUP)

**Input**: Documents to be clustered and a minimum threshold wus value of $min_{wus}$.
**Output**: A set of frequent weighted utility itemsets whose weighted utility support (wus) is greater than threshold $min_{wus}$.

---

**Method**:
**begin**
**for**each document t ∈ T, each term i ∈ I do
    Compute TF – IDF (I, t, T)
**end for**
**for**each document t ∈ T do
    Compute twu(t)
**end for**
**for** term i ∈ I do
    Compute wus (i)
    P←{ i∈ I and wus (i) ≥ $min_{wus}$}
    U← MWIT – WFUP (P, $min_{wus}$)
**end for**
**end**
**MWIT – WFUP** (Itemsets P, $min_{wus}$)
**begin**
**for**term $w_i$∈ P do
    W ←W∪$w_i$
    P←Ø
    **for**($w_j$∈ P and j > i) do
        X = $w_i$∪$w_j$
        Compute wus (X)
        **if**wus (X) ≥ $min_{wus}$then
            $P_i$ ←$P_i$∪ X
            U←MWIT – WFUP ($P_i$, $min_{wus}$)
        **end if**
    **end for**
**end for**
**end**

---

After running this algorithm, a set of weighted frequent utility patterns occurring in our entire corpus of documents will be generated.

**Step:1** Let us consider set of documents belonging to two topics.

Document set T is {d0, d1, . . ., d7} and itemsets I is {word1, word2, word3, word4, word5}. In this, database d0= {1,3,5,0,2} means document d0 contains one word1, three word2, five word3, two word5 and none of word4.

**Step:2** Calculate tf of each word w.r.t to a document.
For example tf(word1) in documents d0,d3 is given as :
tf(word1,d0) = 1 / (1+3+5+0+2) = 0.09090
tf(word1,d3) = 5 / (5+3+1+2+1) = 0.4167
Similarly we can compute tf of each word wrt to every document and can obtain the following matrix:
**Step:3** Inverse Document Frequency (IDF) is a unique score that indicates the importance of a word in a database,

ISSN: 2005-4297 IJCA

| | word1 | word2 | word3 | word4 | word5 |
|---|---|---|---|---|---|
| **0** | 1 | 3 | 5 | 0 | 2 |
| **1** | 4 | 5 | 3 | 0 | 0 |
| **2** | 0 | 0 | 2 | 1 | 3 |
| **3** | 5 | 3 | 1 | 2 | 1 |
| **4** | 3 | 2 | 1 | 0 | 0 |
| **5** | 0 | 0 | 0 | 4 | 0 |
| **6** | 0 | 3 | 0 | 0 | 0 |
| **7** | 1 | 3 | 4 | 0 | 0 |

Table 1. Input document set

Hence it indicates the weight of the word as shown in table.1. For example ,
idf(word1,D) = log(8/5) =0.204120 idf(word2,D) = log(8/6) = 0.124939

**Step4:**Using the tfs and idfs,transaction weighted utility (twu) of each document can be calculated as shown in table 2 and table 3.For example,
twu(d0)=(0.0909*0.204120 + 0.2727*0.124939 + 0.4545*0.124939+ 0.1818*0.425969) / 4
= 0.04617.

| | word1 | word2 | word3 | word4 | word5 |
|---|---|---|---|---|---|
| 0 | 0.090909 | 0.272727 | 0.454545 | 0.000000 | 0.181818 |
| 1 | 0.333333 | 0.416667 | 0.250000 | 0.000000 | 0.000000 |
| 2 | 0.000000 | 0.000000 | 0.333333 | 0.166667 | 0.500000 |
| 3 | 0.416667 | 0.250000 | 0.083333 | 0.166667 | 0.083333 |
| 4 | 0.500000 | 0.333333 | 0.166667 | 0.000000 | 0.000000 |
| 5 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| 6 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 7 | 0.125000 | 0.375000 | 0.500000 | 0.000000 | 0.000000 |

Table 2. TF of each word wrt a document

| | |
|---|---|
| 0 | 0.046717 |
| 1 | 0.050444 |
| 2 | 0.108542 |
| 3 | 0.046638 |
| 4 | 0.054843 |
| 5 | 0.425969 |
| 6 | 0.124939 |
| 7 | 0.044945 |
| 8 | 0.903037 |

Table 3.twu of each document

**Step5:**
Using the knowledge of tfs and twu we can calculate the weighted utility support (wus) for each word in the corpus using the following formula:

wus(word1) = (0.046717 + 0.050444 + 0.046638 + 0.054843+ 0.044945) / 0.903037 = 1.2

Thus, the frequent patterns obtained for the above example are: {word1}, {word4}, {word5}, {word3}, {word2},{word1, word3}, {word1, word2}, {word3, word5}, {word2, word3}, {word1, word2, word3}. Similarly, for each document we can compute the twu.

ISSN: 2005-4297 IJCA

### 3.3 Frequent pattern based Agglomerative Text Clustering

Hierarchical clustering generates a nested partitions of clustering structure.Hierarchical clustering algorithms are either top-down or bottom-up. The clustering starts with singleton sets of each point in an agglomerative clustering algorithm. That is, each data point is its own cluster. At each time step, the most similar cluster pairs are joined according to the selected similarity measure, and this step is repeated either until all data points are included in a single cluster or until some predetermined criteria are met. Maximum Capturing(MC) algorithm is one of the most effective methods for text clustering. However, the Maximum Capturing clustering Algorithm based on the similarity matrix leads to a lot of irrelevant clusters that aren't desired. To overcome this problem,weighted frequent pattern based agglomerative hierarchical clustering is proposed to find the desired clusters using frequent terms from the documents.

---

Algorithm:Weighted Frequent Pattern based Agglomerative Clustering(WFUP_AC)

**Input**: Set of Documents D, Frequent Patterns F.

**Output**: Clustered Documents

---

**Method**

**begin**

- Compute a similarity matrix A, such that $A_{ij}$denotes the common frequent itemsets of documents $d_i$, $d_j$.
- Find the max of similarity matrix, max = max(A)
- Compute distance matrix using similarity matrix

  **for** s ∈ A do:

  D←max – s

  **end for**

  **for** n down to 2 do

  - Search matrix D for a closest pair (I,j) of clusters
  - Replace clusters I and j by an agglomerative cluster h
  - Update to reflect deletion of i and j and to exhibit revised dissimilarities between h and all remaining clusters.

**end for**

- Decide where to cut in the hierarchy of agglomerated clusters.
- According to the cut clusters are formed

**end**

---

The weighted Frequent Pattern based Agglomerative Clustering algorithm(WFUP_AC)is applied on frequent patterns and clustered are produced for the documents based on the terms present in the documents. The step by step process is given below:

•

**Step1:**Construct the similarity matrix A, where $A_{ij}$ithe common itemsets between documents $d_i$and $d_j$.The similarity matrix for the above example is shown in table 4.

ISSN: 2005-4297 IJCA

[[0,7,3,9,7,0,1,7],
[7,0,1,7,7,0,1,7],
[3,1,0,4,1,1,0,1],
[9,7,4,0,7,1,1,7],
[7,7,1,7,0,0,1,7],
[0,0,1,1,0,0,0,0],
[1,1,0,1,1,0,0,1],
[7,7,1,7,7,0,1,0]]

Table 4. Similarity Matrix

**Step2:**Compute the distance matrix by taking max of similarity matrix and subtracting it from every element of the similarity matrix. Note, here we made diagonal of both similarity and distance matrix as 0s, since the distance between 2 same documents should be 0.The distance matrix for the above example will be :

[[0,2,6,0,7,2,9,8,2],
[2,0,8,2,2,9,8,2],
[6,8,0,5,8,8,9,8],
[0,2,5,0,2,8,8,2],
[2,2,8,2,0,9,8,2],
[9,9,8,8,9,0,9,9],
[8,8,9,8,8,9,0,8],
[2,2,8,2,2,9,8,0]]

Table 5. Distance Matrix

**Step3:** Agglomerative clustering is applied on the above distance matrix using complete linkage. Linkage function is used to group objects into hierarchical cluster tree, based on the distance information generated in the previous step. Objects/clusters that are in close proximity are linked together using the linkage function.

**Step 4:**The formed dendrogram for the data is shown in figure 2:

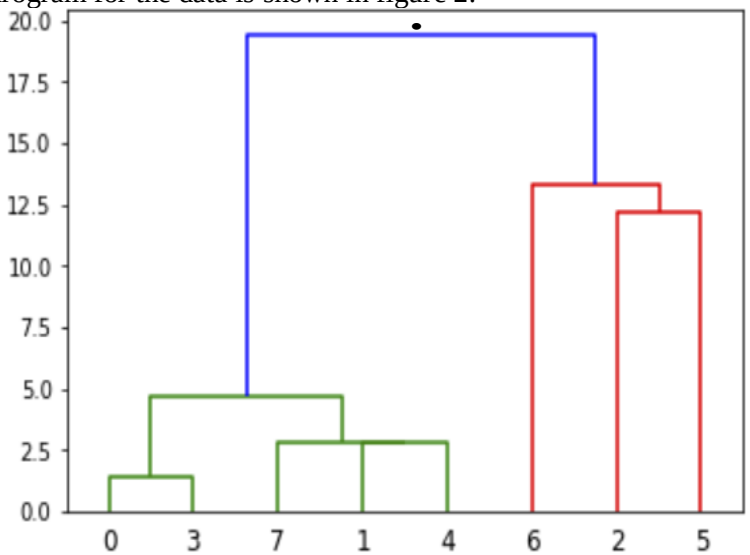


Figure 2. Dendrogram

ISSN: 2005-4297 IJCA

Next, need to determine where to cut the hierarchical tree into clusters. This creates a partition of the data. Since the documents were from 2 topics so we cut at level 4 to form 2 clusters. Thus, documents 2,5,6 form 1 cluster and 0,1,3,4,7 form the 2nd cluster.

**4. Results&Analysis**

Two types of data sets were considered for the experiments. One is the newsgroup documents and second one is Reuters document data sets.To evaluate our proposed approach of text clustering using WFUP and agglomerative clustering, the F-measure(equation.5) is adopted, which is the harmonic function of Precision and Recall. Precision and Recall are computed as follows:

$$P(i,j) = \frac{n_{ij}}{n_j}$$

$$R(i,j) = \frac{n_{ij}}{n_i}$$

$$F(i,j) = \frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)} \quad (5)$$

Where $P_{ij}$ is the precision of clusterj in class i,$R_{ij}$ is the recall of class i in cluster j, n ij is the number of documents of class i in cluster j, n j is the number of documents of cluster j, and n i is the number of documents of class i. The F-measure of cluster j in class i, $F_{ij}$ . Generally, the higher the F-measure, the better the clustering performance is for the data set. The proposed method is implemented in a system using python's framework and Ubuntu 18.04—64 bit, with Intel Core i7 and 16GB RAM. The proposed Weighted Frequent Pattern based Agglomerative Clustering(WFUP_AC)algorithm performance is compared with the Maximum Capturing(MC) algorithm for the two data sets. Figure 2, shows the performance of WFUP_AC with MCon newsgroup document data sets. In the newsgroup documents, 20 newsgroups data are considered. The dataset is a collection of newsgroup documents. Data contains a set of messages that were posted on the particular newsgroup. There is a list of 20 different newsgroups, broadly of 5 different topics computer, sports, science, politics, and religion.The experiments were performed by selecting 2 topics from newsgroups namely sports and computer for better analysis and exploration. To ensure scalability we have gradually increased the number of documents as an input for clustering from 100 documents to 1000 documents. The min support also varied at the different document level for analyzing the best cluster point.

ISSN: 2005-4297 IJCA

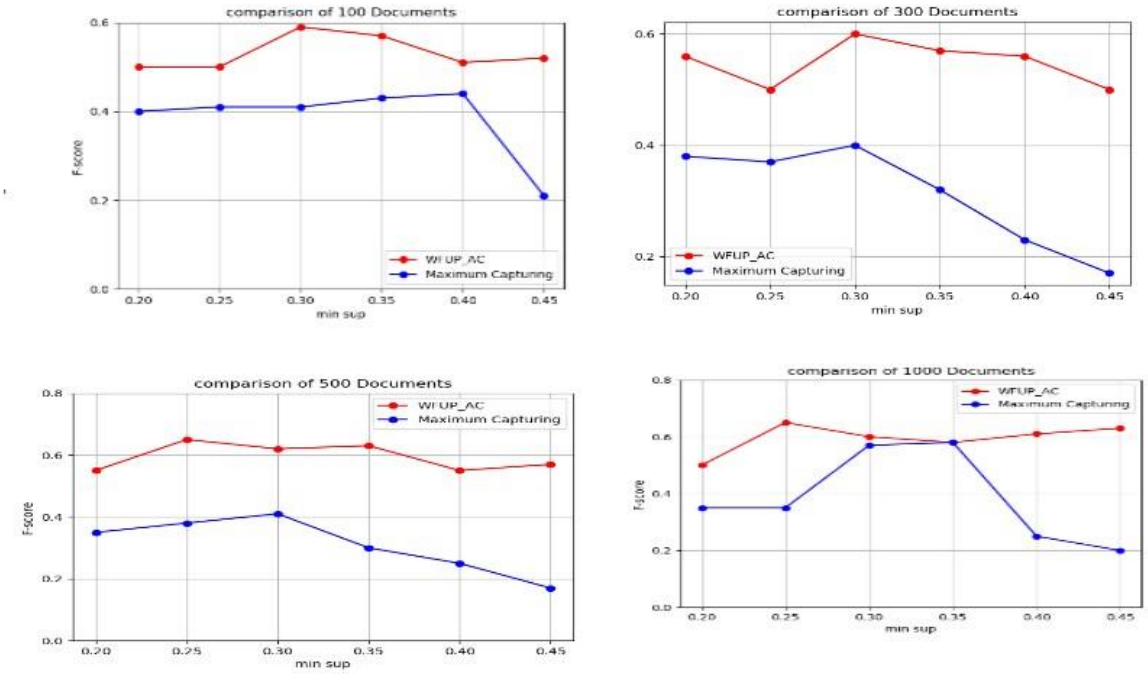


Figure 3. The performance of WFUP_AC(Agglomerative) Algorithm with Maximum Capturing Clustering Algorithm for News group data sets

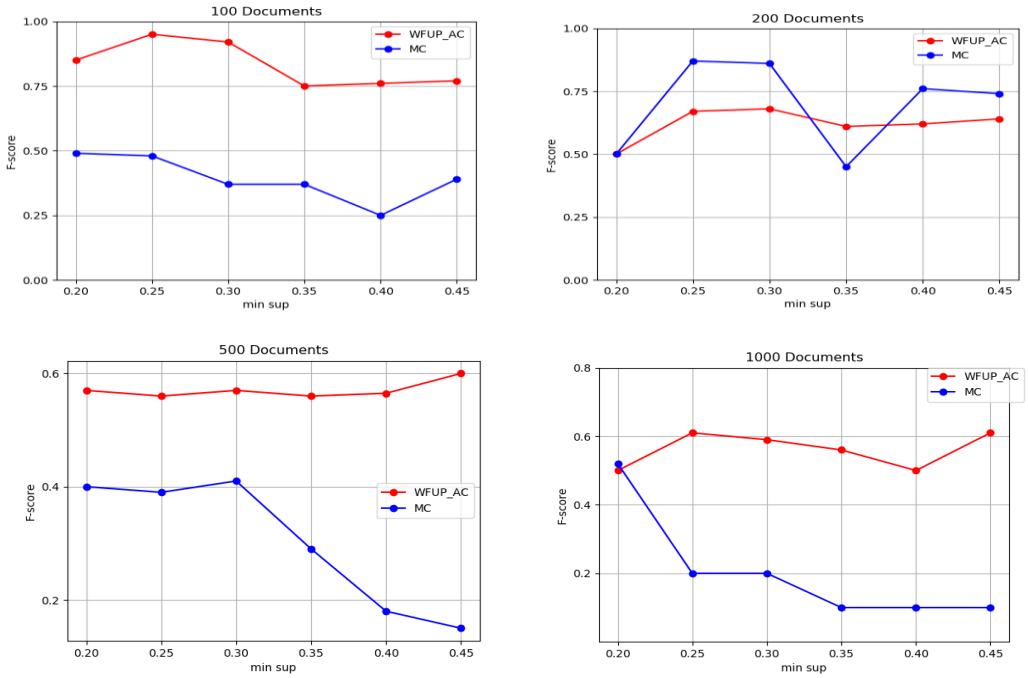


Figure 4. The performance of WFUP_AC (Agglomerative) Algorithm with Maximum Capturing Clustering Algorithm for Reuters Text data sets

ISSN: 2005-4297 IJCA

As evident from the graphs shown in figure 3, we can conclude that the algorithm is scalable and works well for any number of documents. Moreover, agglomerative clustering outperforms the maximum capturing algorithm in all the possible different scenarios. The F-score[8] of agglomerative is always greater than that of maximal capturing. We need to choose the value of min support optimally too. If we choose it too high then fewer frequent of patterns will be mined thus, the similarity matrix won't be able to capture the details properly. And if it's set too low then many frequent patterns will be formed resulting in an explosive similarity matrix that is not good. Through our experiments, we can safely conclude min support of 0.3 works well for these sets ofexperiments.

The proposed weighted frequent pattern based agglomerative clustering performance is tested with Reuters-21578 text documents data set. The Reuters-21578 collection is distributed in 22 files. Each ofthe first 21 files contain 1000 documents, while the last contains 578 documents. The experiment shows that F score of WFUPAC is higher than Maximal Capturing algorithm as shown in figure.3. The other observation is, the WFUP_AC is also working better than MC with large number of documents.Theproposed methodology produces desired cluster and avoids generating of more micro clusters.

## 5.Conclusion

In this article, frequent pattern mining is integrated with hierarchical clustering approach and a novel weighted frequent utility pattern based agglomerative clustering is proposed to process large amount of text document data sets. The text documents are preprocessed using tokenization,lemmatization and stemming and set of terms are generated as tokens. Weighted frequent utility pattern algorithm is applied on the terms and computed frequent patterns.Using these frequent patterns similarity matrix is constructed. The distance matrix is computed from the similarity matrix and hierarchical agglomerative clustering complete linkage is applied and desired clusters are obtained in the form of dendrogram. Based on the requirement dendrogram is cut at the desired level to obtain final clusters. The proposed WFUP_AC approach shows better results(better F-score)than using FWUI with Maximum Capturing. Also, our approach works well for large amounts of documents. Primarily this research was done on English Text. As future work, one can focus on using these techniques on other languages such as Vietnamese, French, Hindi, etc.

## References

[1] Vo, B., B. Le, and J. J. Jung. 2012. A tree-based approach for mining frequent weighted utility itemsets. In Proceedings of the 4th International Conference on Computational Collective Intelligence: Technologies and Applications—Volume Part I, ICCCI'12:114–23

[2] Wu, Dingming et al. "Clustering in Geo-Social Networks." IEEE Data Eng. Bull. 38(2015): 47-57.

[3] Laith Mohammad Abualigah, AhamadTajudinKhader, Mohammed Azmi Al-Betar,Osama Ahmad Alomari,"Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering", Expert Systems with Applications, Volume 84, 2017, Pages 24-36.

[4] Liang Yao et al "Graph Convolutional Networks for Text Classification" The Thirty-Third AAAI Conference on Artificial Intelligence.Hilton Hawaiian Village, Honolulu, Hawaii, USA (AAAI-19).

[5] Y. Fan, L. Gongshen, M. Kui and S. Zhaoying, "Neural Feedback Text Clustering With BiLSTM-CNN-Kmeans," in IEEE Access, vol. 6, pp. 57460-57469, 2018.

[6] Yanjun Li, Soon M. Chung, John D. Holt, "Text document clustering based on frequent word meaning sequences", Data & Knowledge Engineering, Volume 64, Issue 1, 2008, Pages 381-404.

[7] Youcef Djenouri, AsmaBelhadi, Philippe Fournier-Viger, Jerry Chun-Wei Lin, "Fast and effective cluster-based information retrieval using frequent closed itemsets", Information Sciences, Volume 453, 2018, Pages 154-167.

[8] C. Fahy, S. Yang and M. Gongora, "Ant Colony Stream Clustering: A Fast Density Clustering Algorithm for Dynamic Data Streams," in IEEE Transactions on Cybernetics, vol. 49, no. 6, pp. 2215-2228, June 2019.

[9] M. D. Nguyen and W. Shin, "An Improved Density-Based Approach to Spatio-Textual Clustering on Social Media," in IEEE Access, vol. 7, pp. 27217-27230, 2019.
[10] S. Srivastava, S. Pande and S. Ranu, "Geo-Social Clustering of Places from Check-in Data," 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, 2015, pp. 985-990.
[11] D. Kellner, J. Klappstein and K. Dietmayer, "Grid-based DBSCAN for clustering extended objects in radar data," 2012 IEEE Intelligent Vehicles Symposium, Alcala de Henares, 2012, pp. 365-370.
[12] M. D. Nguyen and W. Shin, "An Improved Density-Based Approach to Spatio-Textual Clustering on Social Media," in IEEE Access, vol. 7, pp. 27217-27230, 2019.
[13] D. Wu, J. Shi and N. Mamoulis, " Density-Based Place Clustering Using Geo-Social Network Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 5, pp. 838-851, 1 May 2018." IEEE Transactions on Knowledge and Data Engineering, Vol. 30, No. 5, May 2018.
[14] OussamaRouane, HaceneBelhadef, Mustapha Bouakkaz, "Combine clustering and frequent itemsets mining to enhance biomedical text summarization", Expert Systems with Applications, Volume 135, 2019, Pages 362-373.
[15] Laxmi Lydia, E., Sharmili, N., Nguyen, P. T., Hashim, W., & Maseleno, A. (2019). Automatic document clustering and indexing of multiple documents using KNMF for feature extraction through hadoop and lucene on big data. Test Engineering and Management, 81(11-12), 1107-1130. Retrieved from www.scopus.com
[16] Reddy, S. S. R., Malathi, P., & Mahalakshmi, D. (2019). Efficient datacenter clustering in map reduce framework using cache index algorithm. Test Engineering and Management, 81(11-12), 5418-5422. Retrieved from www.scopus.com
[17] Vanitha, R. (2019). BOVW classification method with particle swarm optimization in big data. Test Engineering and Management, 81(11-12), 4529-4535. Retrieved from www.scopus.com
[18] Kumar, R., & Bhardwaj, D. (2020). An improved moth-flame optimization algorithm based clustering algorithm for VANETs. Test Engineering and Management, 82(1-2), 27-35. Retrieved from www.scopus.com
[19] Kumudha, & Shanmuga Prabha, P. (2020). Effective combination of biclustering mining and adaboost learning for breast tumor analyzation. Test Engineering and Management, 82, 2060-2063. Retrieved from www.scopus.com
[20] Lee, J. -. (2020). Distance-based 2-hop clustering algorithm in WSNs (wireless sensor networks). Test Engineering and Management, 83, 4299-4306. Retrieved from www.scopus.com
[21] Madhumitha, S., & Ashwini, S. (2020). Foresight-based and locality-aware task setup for complementing video transcoding over deep learning clustering. Test Engineering and Management, 82, 2015-2019. Retrieved from www.scopus.com
[22] Amru, M., & Bhavani Sankar, A. (2019). Analytical study on design and development of herichal clustering using adhs algorithm. International Journal of Advanced Science and Technology, 28(20), 1208-1213. Retrieved from www.scopus.com
[23] Bakeyalakshmi, P., & Mahendran, S. K. (2019). Integrated clustering based intrusion detection model (IC-IDM) in mobile ad-hoc networks. International Journal of Advanced Science and Technology, 28(12), 306-318. Retrieved from www.scopus.com
[24] Dinesh, G., Rajinikanth, C., Maheswara Venkatesh, P., & Vanitha, T. V. (2019). An optimized Dijkstra's SPF algorithm used multihop clustering for scalable IoT systems. International Journal of Advanced Science and Technology, 28(14), 298-306. Retrieved from www.scopus.com
[25] Joseph, J., & Kesavaraj, G. (2019). Evaluation of clustering algorithms for credit card data set using WEKA. International Journal of Advanced Science and Technology, 28(17), 392-400. Retrieved from www.scopus.com
[26] Joshi, R. R., Mulay, P., Lohia, A., Singh, A., & Nagar, A. (2019). Mapreduce4CFBA: Distributed incremental closeness factor based clustering algorithm (DICFBA) for analysis of chronic diseases on hadoop mapreduce. International Journal of Advanced Science and Technology, 28(1), 241-253. Retrieved from www.scopus.com
[27] Kiruthika, M., & Sukumaran, S. (2019). Multi-modal approach with deep embedded clustering for

social image retrieval. International Journal of Advanced Science and Technology, 28(17), 946-995. Retrieved from www.scopus.com

[28] Madhu, T., Sarma, S. S. V. N., & Murthy, J. V. R. (2019). Fault containment based self stabilized fuzzy relevance clustering algorithm. International Journal of Advanced Science and Technology, 28(14), 502-512. Retrieved from www.scopus.com

[29] Goar, V. K., Kuri, M., & Tanwar, P. K. (2019). Energy efficient base clustering algorithm to achieve optimum value of energy consumption. International Journal of Control and Automation, 12(5), 581-587. Retrieved from www.scopus.com

[30] Kaur, N., Verma, S., Kavita, & Yadav, A. L. (2019). Survey on classification and clustering schemes in big data using image processing. International Journal of Control and Automation, 12(4), 170-188. Retrieved from www.scopus.com

ISSN: 2005-4297 IJCA