Application of Formal Concept Analysis in Overlapping Document Clustering

Yong Liu[1,2*], Yu-Peng Hu[1], Jie-Cai Zheng[1,3], and Xue-Qing Li[1]

[1]School of Computer Science and Technology, Shandong University, Jinan, China
{liuyong, huyupeng, zhengjiecai, xqli}@sdu.edu.cn

[2]Department of Computer Engineering, Changji University, Changji, China

[3]School of Sport Communication and Information Technology, Shandong Sport University, Jinan, China

Abstract. Clustering is a technique for classifying objects into multiple clusters through the calculation of similarity. However, there are overlapping phenomena in clustering because the objects may belong to two or more categories in the real world. We propose the overlapping clustering method FCAC (Formal Concept Analysis overlapping Clustering) based on formal concept analysis. Using Formal Concept Analysis to find objects with the same features, replace the original text document with the concept vector of formal context output for clustering to achieve the effect of Overlapping Clustering. At the same time, because the number of concept vectors is much smaller than the number of feature words in the data set, the problem that the traditional clustering algorithm vectors grow too quickly has been improved. The FCAC method can be combined with various classical clustering algorithms such as K-means and HAC (Hierarchical Agglomerative Clustering) to fully apply existing and future research results. The experimental results show that this method can achieve the effect of overlapping clustering, and effectively reduce the computational space occupied by two-dimensional vectors, and improve the efficiency of clustering.

Keywords: Formal concept analysis, Document clustering, Text mining, Concept vector

1 Introduction

Clustering is an important data preprocessing method in the field of machine learning and data mining (Tseng,2005). Its purpose is to obtain the intrinsic distribution structure of valuable sets from the unlabeled dataset, thereby simplifying the description of them (Fan,2016). Clustering is an unsupervised learning method that attempts to find its distribution or pattern in untagged datasets (Lv, 2016). In general, we consider that data points in the same cluster have greater similarities than data points in different clusters. Clustering is to classify similarly characterized objects into the same cluster. In the research of traditional clustering, an object belongs to a cluster, which is called disjoint clustering (Dillon,2001). In the study of overlapping clustering, allowing one object to belong to multiple clusters. There are many applications in real life that are overlapping clusters (Omran, 2007., Mulder,2013., Yu,2016). For example, in biology, fraction of proteins may exist in the structure of multiple protein complexes. Text clustering is text mining. The main purpose is to discover important patterns in text documents and group them into clusters (Handfield,2013).

Text clustering is a kind of text data mining. Its main purpose is to discover important patterns in text documents and group them into clusters. In the reference (Pérez-Suárez,2018., Shah,2012., Popat,2014), the clustering algorithm is divided into more than ten categories such as Partitioning Method, Hierarchical Clustering Method and Model Clustering Method. Among them, k-means algorithm (Krishna, 1999., Capó,2017) and F-measures algorithm (D'Hondt,2010) are more famous. Other commonly used clustering algorithms are Hierarchical Agglomerative Clustering (HAC) (Castellanos,2017., Ackermann,2014), COBWEB, and so on. However, most clustering algorithms can only allocate documents to no more than one cluster. For a clustering algorithm that allows two or more clusters to share the same document, they are often called overlapping clusters (Mulder,2013).

Graph-based clustering algorithms are one of the common methods for building overlapping clusters. Such methods usually represent each data point as a node. The weights between the nodes are the similarities between the data, and similarity determines which nodes need to be merged or split. (Bonchi and Francesco et al.,2014) proposed the original data is first represented as data points, and the similarity of a pair of data points (u, v) can be expressed as $s(u, v)$, $l(u)$ and $l(v)$ respectively. The set of clusters where the data points u and v are labeled, and $H(l(u),l(v))$ represents the Jaccard similarity of $l(u)$ and $l(v)$. When clustering, $|H(l(u),l(v))-s(u,v)|$ is calculated for each pair of data points (u,v) and summarized, combined with local search to try to select the best clustering result. The study is based on a pair of data points as a reference unit, a single data point can be mapped to multiple labels (Labels) at the same time, so as to achieve the purpose of overlapping clustering. Wang Y et al. (2016) proposed the clustering method to first represent the data points and similarity as graphs, and the data points are connected by edges, which represent similarities between data points. When clustering using this method, the cluster is represented by a star-shaped structure, and the number of neighboring nodes is calculated for each node to judge the influence of this node on the entire graph, and the weight of the influence of the node is referenced. Nodes with lower weights can belong to multiple nodes with high weights at the same time, which is also an important factor for implementing overlapping clustering in this method. Zhang X et al.(2015) improved

* Corresponding Author

the agglomerative hierarchical clustering method, proposed Fuzzy Agglomerative Hierarchical Clustering (FAHC), and used a fuzzy mechanism to select the merged initial cluster. It uses the similarity between documents to determine its ownership cluster. FAHC defines similarity thresholds and difference thresholds, and treats each document as a cluster with a membership degree. The cluster to which the document d_i belongs is A_i, and $M_{A_i}(d_i)$ indicates the degree to which the document d_i belongs to the cluster A_i. Initially, each document belongs to each cluster with a degree of 1. When d_i and d_j are smaller than the similarity threshold, the clusters need to be merged and the corresponding degree of membership is updated. The merged new group and the next neighbor cluster continue to calculate the membership. The difference between degree and similarity, if less than dissimilarity, merge the clusters and update the degree of membership. Repeating these steps until the degree of membership of multiple documents is less than the threshold indicates the completion of this clustering process. Yu et al. (2016) proposed a three-way decisions Overlapping Clustering Algorithm Based on the Relation Graph (TDC-RG).TDC-RG uses three decision theory as the basis of clustering theory. The difference compared to Binary-decision method is that if an object can only belong to or does not belong to a certain cluster, it may also belong to the third case may not belong. This method determines the degree of membership of the cluster to which the object belongs by setting the Upper Bound and Lower Bound of the cluster to determine the cluster to which the object belongs. Moreover, there are other studies that also involve overlapping clustering (Capó, 2017., D'Hondt,2010., Castellanos,2017., Moertini,2018), but these studies have limited performance evaluation for overlapping clusters.

In the research of clustering using text as data source, the Vector Space Model (VSM) is still the more commonly used clustering method. After the text is converted into a computable vector, clustering is performed using similarities such as K-means clustering algorithm or hierarchical clustering algorithm. Reviewing related research, we found that these clustering methods are still inadequate. First of all, for the cluster analysis of texts, a document is usually represented as a single vector. If no direct clustering is performed, the document cannot be classified into multiple clusters. In the study of Yu and Mulder D et al., (Yu,2016., Mulder,2013) it was proposed that the clustering results should have partial overlap, otherwise if they are only classified into a single cluster, then the other meanings of this document are ignored. Moreover, another issue of concern is the problem of high vector space dimensionality when translating documents into VSMs using the Term Frequency–Inverse Document Frequency (TF-IDF) method, mainly because There are more vocabularies for documents.

In this paper, we combine Formal Concept Analysis (FCA) (Wille,1982) to assist with clustering. We use the concept lattice of the FCA to create a hierarchical conceptual structure. Each node in the concept lattice is a formal concept that contains a set of documents and keywords. This approach not only represents the relationship between documents and keywords, but also finds documents and content with common keywords, and gathers together the same concepts. If we output the concept as a vector instead of the original document for clustering, then a document has the opportunity to be divided into more than two clusters. Moreover, the number of concept vectors is much smaller than the number of data set keywords, and the vector space for calculating similarity is greatly compressed. This article will select some text data sets in the UCI Machine Learning Repository as experimental data sets to test the clustering effects of different clustering methods.

The remainder of this paper is organized as follows. Section II presents our document clustering approach in detail. Section III provides experimental results and comparisons with traditional methods. Section IV states our conclusions and provides suggestions for future work.

2 Proposed Approach

Based on the study of traditional clustering methods, this paper proposes an FCAC (FCA Overlapping Clustering Methods) based on the help of FCA theory to assist text file clustering. We use the FCA to define a single text document as a formal concept, transform each formal concept into a concept vector, and use an algorithm such as HAC and K-means to cluster concept vectors to build overlapping clusters. The clustering method architecture diagram is illustrated in Figure 1.

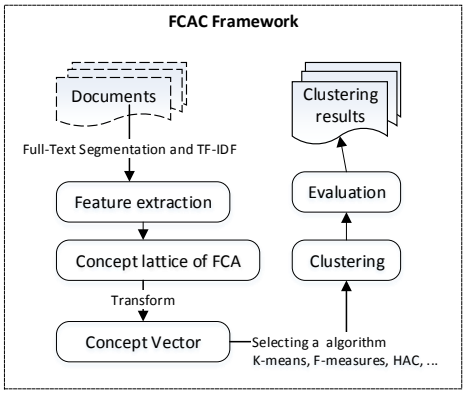

Fig. 1. Architecture of text clustering method based on FCA

We use the process shown in Figure 1 to achieve overlapping clustering. In this process, we first use the existing tools of Natural Language Processing (NLP) to perform full-text segmentation of original text documents and filter stop words. Then we use Term Frequency (TF) to calculate the frequency of each vocabulary, according to the frequency to select the higher frequency vocabulary, as a representative of the document keywords. Next, we use the FCA to convert the original data into concept lattices. Select the appropriate node layer from the concept lattice to transform it into a concept vector to replace the original document. Then we use the traditional algorithm to complete the clustering.

In order to make the description of this research method more complete, we make the following conventions:

(1) D represents a set of all documents, that is $D = \{d_1, d_2, \ldots, d_n\}$;
(2) KW represents a set of all keywords, that is $KW = \{kw_1, kw, \ldots, kw_n\}$;
(3) RC represents the result of clustering, which represents the set of all clusters, that is $RC = \{C_1, C_2, \ldots, C_n\}$;
(4) FC represents the formal concept, which represents the set of all the concepts in the concept lattice, that is $FC = \{fc_1, fc_2, \ldots, fc_n\}$;

$X(d_i)$ indicates the category of the cluster to which the document d_i belongs.

In order to select a representative feature from the original document object, it is usually necessary to calculate the weight value of the feature in the object, and select a feature with a higher weight value in the object to represent the band object. (Alelyani,2013)In this paper, the experiment uses text documents as data sets. Through the steps of retrieving, restoring, and selecting candidate terms, feature vocabularies representing specific documents, ie keywords, are selected from the candidate terms.

The selection of keywords in a text document first requires full text segmentation. The number of vocabularies obtained by a document via a full-text segmentation may be as high as several hundred, or even more. However, the frequency of many words appearing in the articles is not high. Only the words with higher frequency are suitable for representing the content of the documents. The words with low frequency may become noise when the similarity is calculated. Therefore, the feature words must be strictly screened. In the study of reference (Kuo,2010), TF was used to rank the vocabulary. The higher the TF is, the higher the frequency of the vocabulary is, and it is more appropriate to represent the content of the document. After sorting according to the TF level, the study obtained the first 20 features. The vocabulary is more representative of the content of the vocabulary. We refer to the research conclusions of this document and use the same method to extract the feature vocabulary of each article no more than 20, as the keyword of the document where the feature vocabulary is located.

TF is a common feature selection method in text clustering research. We refer to the research results of (Kuo,2010., Bashir,2016), and use TF to calculate the frequency of vocabulary appearing in the affiliated document. The formula (1) defines the relevant calculation method.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_w n_{w,j}}, \qquad (0 < i < w) \tag{1}$$

In formula (1), the numerator is the number of occurrences of the vocabulary t_i in the document d_j, the denominator is the sum of the occurrences of all vocabularies in the document d_j, and w is the total number of all vocabularies in the document.

2.1 Formal Concept and Concept Lattice

FCA is a mathematical theory proposed by the German mathematician R.Wille (1982). It is a branch of applied mathematics. The theory provides an effective tool to support data analysis, and its core is concept lattice. Each node of the concept lattice is a formal concept. It consists of the extent of the concept and intent. The extent is all the objects covered by the concept, and the intent is the attribute description of the concept. The concept lattice vividly and concisely reflects the generalization and specialization relations between concepts through Hasse diagrams. We use the features of the FCA concept lattice structure to transform text documents into conceptual vectors, making a document not only represented by a single vector, but it can be classified into multiple clusters after passing through the clustering algorithm. In addition, because FCA can find common feature words between documents, the original huge number of keywords in vector space is replaced by the concept vector, which can effectively reduce the size of the vector space dimension.

In order to make the description of research methods more complete, the following important definitions in the FCA are listed below:

Definition 1 (Formal Context (Wille, 1982) A formal context is a triple $\mathcal{K} = (G, M, I)$ where G denotes a set of objects, M a set of attributes, and I a binary relation defined on $G \times M$. $I \subseteq G \times M$ is a binary table which assigns an attribute to an object. We write gIm or $(g, m) \in I$ to mean that object g has attribute m.

The visual presentation of the formal context needs to be implemented with the help of a data matrix. If $(g, m) \in I$, the intersection of the row $g \in G$ and the column $m \in M$ is represented by ×. Table 2 shows a typical formal context consisting of 8 objects and 19 attributes. The × in the table indicates the relationship between the corresponding objects and attributes.

Definition 2 (Formal concept (Wille, 1982)). A formal concept of a formal context $\mathcal{K} = (G, M, I)$ is a pair (A, B) where $A \subseteq G$ is called the extent, $B \subseteq M$ is called the intent of the concept, and A is the maximal set of

objects sharing the whole set of attributes in B (and vice versa). The set of all formal concepts of the context $\mathcal{K} = (G, M, I)$ is denoted by $\mathcal{C}(G, M, I)$. A formal concept is considered to be identified by its extent and its intent.

Concepts are computed on the base of a Galois connection defined by two derivation operators denoted by 0:

$$A' \coloneqq \{m \in M | \forall g \in A, gIm\}$$
$$B' \coloneqq \{g \in G | \forall m \in B, gIm\}$$

A concept (A, B) verifies a closure constraint so that $A' = B and B' = A$.

Definition 3 (Concept lattice (Wille, 1982)). The ordered set $\mathcal{C}(G, M, I, \leqslant)$ of all formal concepts of (G, M, I) is a complete lattice called the concept lattice of (G, M, I).

Concept lattices can be visualized as line diagrams. In a line diagram, each node represents a formal concept and an edge represents a subsumption relation between two concepts. Figure 2 illustrates a line diagram of a concept lattice related to the formal context of Table 2.

According to the formal context, formal concepts, and definition of concept lattice, we find that the larger the super concept, the more objects with common features and the fewer common features. When the objects have the same features, it means that the objects may describe similar topics. Therefore, the largest parent concept in the concept lattice will be used in the experiment in this paper to form concept set C, and the concept will be converted to Concept Vetcor via the Vector Space Model (VSM).

Table 1. An example of articles and feature words

Article	Font size and style
d1	China, Import commodities, Tariff, Crude Oil, Steel
d2	Zambia, Import commodities, Tariff, Chemicals, Foodstuffs
d3	Import commodities, Foodstuffs, Food security, Health
d4	Foodstuffs, WHO, Health
d5	Nigeria, WHO, Immunization, Polio
d6	Liberia, Racial conflict, WHO, Humanism
d7	Tariff, Tobacco, Auto
d8	Health, Tobacco

In this article, we use the article as the object, the document's feature words as the object's attribute, and Table 2 shows an example of a list of articles and feature words (the information source is from UCI's reuters-21578).After the FCA calculates the common feature words, the concept lattice Hasse diagram shown in Figure 2 is obtained, each node in the diagram is a formal concept. Because the concept lattice is a hierarchical tree structure, the closer the nodes to the top of the concept lattice, the more objects with common attributes and the fewer common attributes, and the concepts have super-concept and sub-concept relationships. For example, fc_8 and fc_9, objects and attributes are {Foodstuffs, Health} and {Health, Tobacco}, with super-concept fc_3={Health}, where {d_3, d_4, d_8} is the union of two concept objects, {Health} is the intersection of two conceptual attributes. In Figure 2, we use CL_i to represent a set of concepts in the concept lattice hierarchy level i.

Table 2. The binary context of Table 1

	d1	d2	d3	d4	d5	d6	d7	d8
China	×							
Zambia		×						
Nigeria					×			
Liberia						×		
Import commodities	×	×	×					
Tariff	×	×					×	
Foodstuffs		×	×	×				
WHO				×	×	×		
Health			×	×				×
Crude Oil	×							
Steel	×							
Chemicals		×						
Food security			×					
Immunization					×			
Polio					×			
Racial conflict						×		
Humanism						×		
Tobacco								×
Auto							×	

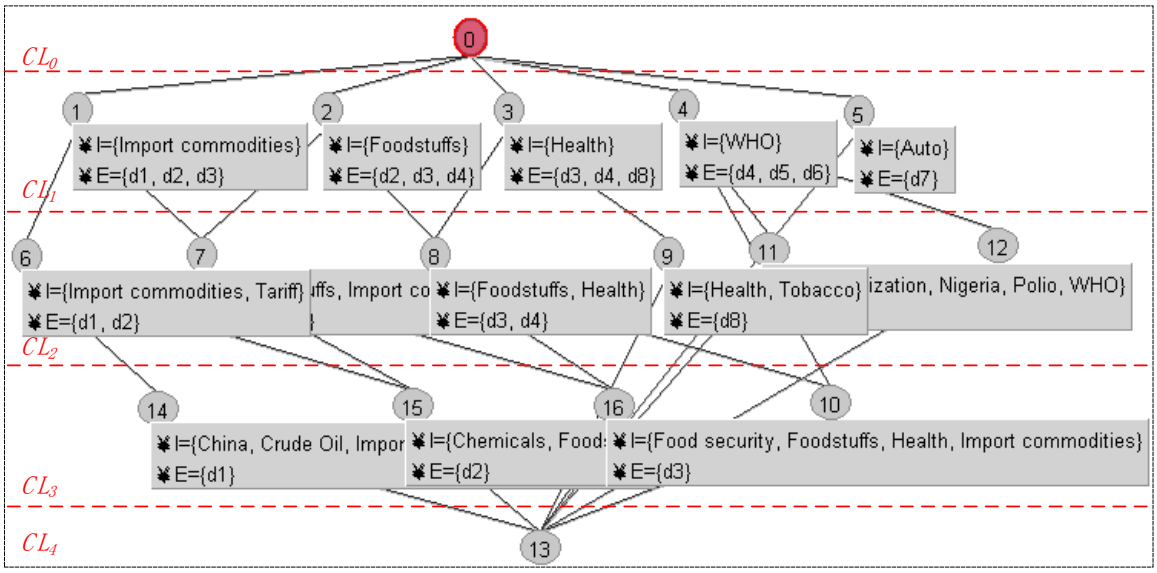

Fig. 2. The concept lattice Hasse diagram of Table 2

2.3Concept Vectors

Use According to the formal context, formal concepts, and definition of concept lattice (Wille, 1982), we find that the larger the parent concept, the more objects with common features and the fewer common features. When the objects have the same features, it means that the objects may describe similar topics (Pei,2016). Therefore, the largest parent concept in the concept lattice will be used in this paper to form the set of concepts. Since the super-concept contains the common characteristics (attributes) of all sub-concepts, we choose all the concepts starting from CL_1 and the matrix M is used to convert the concept into concept vectors. The matrix M is defined as follows.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_w n_{w,j}}, \qquad (0 < i < w) \tag{2}$$

If the number of clusters k is expected to be greater than the number of concepts in CL_1, CL_2 is selected downward and the process will continue until CL_i is greater than k. The matrix M obtained from Figure 2 is shown in Table 3.

Table 3. Matrix of Concept Vectors

Concepts	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
fc1	1	1	1	0	0	0	0	0
fc2	0	1	1	1	0	0	0	0
fc3	0	0	1	1	0	0	0	1
fc4	0	0	0	1	1	1	0	0
fc5	0	0	0	0	0	0	1	0

2.4 Conceptual Vector Clustering Method

At the beginning of the clustering process, all the concepts of CL_1 are first converted into M-matrices, and then each concept vector is designated as a cluster, and the initial clustering result RC can be obtained. If $|RC|$ is less than the specified number of clusters k, then re-select all the concepts of the next layer. If $|RC|>k$ is established, the degree of similarity between clusters is calculated, and the initial cluster is merged or distributed according to the algorithm until clustering is completed. The detailed process is shown in Algorithm 1.

Algorithm 1 Concept Vector Clustering Algorithm

Input: Concept Lattice; k; i=1;
$RC \leftarrow CL_i$;
IF$|RC| < k$
FOReach$CL_i \in Concept\ Lattice$
i++ ;
$RC \leftarrow CL_i$;
IF$|RC| \geq k$
break ;
END FOR
END IF
IF$|RC| > k$
// Clustering by using algorithms sush as K-means, HAC, etc.
END IF
Output: RC

In Algorithm 1, First initialize the concept lattice, specifies the number of clusters *k*, and sets the initial level *i=1*.Line 1 first specifies all CL1 concepts as initial clustering results. The third line determines whether clustering can be started. If the current cluster number is smaller than the specified cluster number *k*, the concept of the next layer is re-selected, and the loop body is entered. The determination of the loop body is determined by the judgment of the seventh row. Line 10 of the code determines whether the current number of clusters is greater than the specified number *k*. If it is greater than the specified number of clusters *k*, use an algorithm such as HAC or *K*-means to merge or assign clusters until |*RC*| meets the specified number of clusters *k*. Line 13 outputs the final clustering result.

Each concept in CL_i is initially considered a cluster and is represented as a concept vector, as shown in Table 2, six concept vectors are obtained. For example, the concept vector of fc_1 is [1,1,1,0,0,0,0,0].When the concept vector is ready, any clustering algorithm can be used for clustering. Because documents can appear in multiple concept vectors, overlapping clusters can be obtained after clustering. In this article, we also modified the HAC algorithm proposed by reference [16].Instead of using full links, average links, or single links, we use united operators to combine clusters, which we call United Hierarchical Clustering (UHAC).For example, two similar formal concepts fc_1 and fc_2, after UHAC agglomeration, C_{new} result base is the union of C_1 and C_2 [1,1,1,1,0,0,0,0], replacing the original C_1 and C_2.

The clustering method proposed in this paper uses the concept vector instead of the original document to calculate the similarity, which can reduce the computational space occupied by the two-dimensional vector. Suppose the number of original files is |D|, the total number of keywords is |KW|, if you do not use the concept vector, the size of the two-dimensional vector matrix is |D|×|KW|, and the size of the matrix *M* used in this paper is |D|×|CL_i|. Because the formal concept is based on a common feature word between documents, |CL_i| is less than |KW|, especially in large document corpora, the size of the vector space is significantly reduced. This means that you can save memory and increase your computing power from the size of your space.

2.5 Analysis and Evaluation Method

The traditional clustering assessment method F-measures (D'Hondt,2010., Kou,2014) is not suitable for the verification calculation of overlapping clusters, because the same document object may be repeatedly calculated by multiple clusters and categories, and this way will make Precision and Recall two The index is too high to accurately reflect the effect of overlapping clusters. In this paper, we refer to B Cubed (Amigo,2009., Sironi,2016) proposed by Amigo Eand Sironi A et al. as an overlapping cluster evaluation method. This method calculates Multiplicity Precision and Recall to evaluate the results of overlapping clusters by comparing the ratio of the intersections of clusters and categories between objects. Among them, Multiplicity Precision (MP) evaluates at least one identical cluster between two objects and has the same clustering ratio; Multiplicity Recall (MR) evaluates at least one same correct clustering between two objects and has the same cluster proportion. The specific definition is shown in formula (3)(4).

$$MP(o,o') = \frac{Min(|C(o) \cap C(o')|, |L(o) \cap L(o')|)}{|C(o) \cap C(o')|} \quad (3)$$

$$MR(o,o') = \frac{Min(|C(o) \cap C(o')|, |L(o) \cap L(o')|)}{|L(o) \cap L(o')|} \quad (4)$$

In the above formula, (o, o') is any two objects, $C(o)$ is the cluster set of objects o, and $L(o)$ is the correct cluster label.After calculating Multiplicity Precision and Multiplicity Recall for each document object, the average of Multiplicity Precision and Multiplicity Recall is called Precision BCubed (P_BC) and Recall BCubed (R_BC) as two evaluation indexes to measure the performance of overlapping clustering algorithm, as shown in formula (5)(6).

$$P\_BC = Avg_o[Avg_{o',C(o)\cap C(o')\neq\emptyset}(MP(o,o'))] \quad (5)$$

$$R\_BC = Avg_o[Avg_{o',L(o)\cap L(o')\neq\emptyset}(MR(o,o'))] \quad (6)$$

3 Experiments and Results

We use the Reuters-21578 data set provided by UCI as an experimental data set. In the overlapped clustering study using this dataset as experimental data, the OClustR method proposed by Pei X and Li W et al.(Pei,2016., Li,2017) exhibits a better clustering effect. Therefore, we use the data set used by OClustR to take two sets of documents for experimentation based on the LEWISSPLIT attribute and TOPICS category of each file in Reuters-21578.One group of documents Reu-Te is based on the LEWISSPLIT attribute of TEST and at least one TOPIC, and the other group of Reu-Tr is based on the LEWISSPLIT attribute of TRAIN and possesses at least one TOPIC.

In the document processing project, JATE (Java Automatic Term Extraction) is used as a POS tool to remove stop words and calculate the weight of TF-IDF. The same feature interception as OClustR is used, and term is used as a feature word. Since the number of feature words and the feature value ω may differ depending on different data sets or different numbers of feature words, in the experiment, we use the Grid Search (2004) calculation method to test the parameter combination of the number *N* of feature words and the feature value ω, and try to use the FCAC method (the specific clustering step uses UHAC algorithm), HAC and K-means algorithm to cluster the concept vector. The HAC here is divided into three HAC algorithms, Single-linkage, Average-linkage

and Complete-linkage, due to the different clustering distance calculation standards. We denote by HAC_S, HAC_A, and HAC_C respectively.

When the parameters of the number of feature words N and the weight value ω are uniformly set to 5 and 1, respectively, the experimental results for the datasets Reu-te and Teu-Tr are shown in Table 4, Figure 3, and Figure 4.We focus on the Precision and Recall indicators. The accuracy of the FCAC method is significantly better than that of other algorithms. The Recall of the FCAC method is slightly higher than other algorithms.

Table 4. Performance of different methods on the two datasets

Method	Precision (Reu-Te)	Recall (Reu-Te)	Precision (Reu-Tr)	Recall (Reu-Tr)
FCAC	0.5404	0.8838	0.5275	0.7872
HAC_S	0.4937	0.8541	0.4797	0.7053
HAC_A	0.4937	0.8733	0.4797	0.7729
HAC_C	0.4931	0.8817	0.4788	0.7587
K-means	0.4684	0.8247	0.4600	0.7138

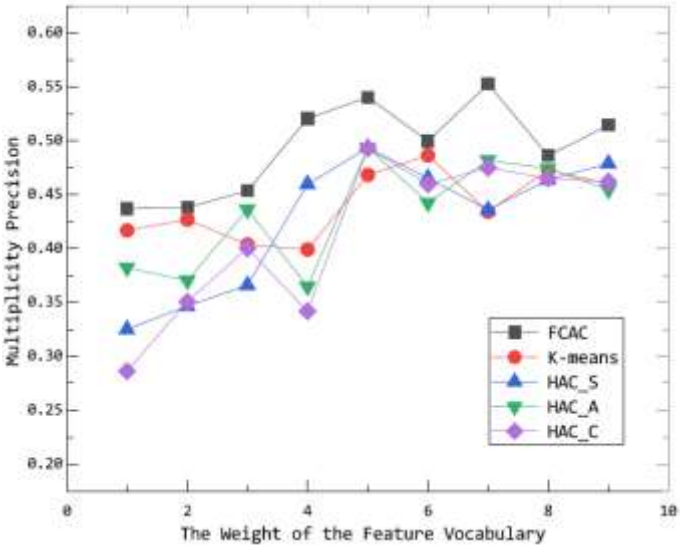

Fig. 3. Precision curves of different methods on the Reu-Te dataset.

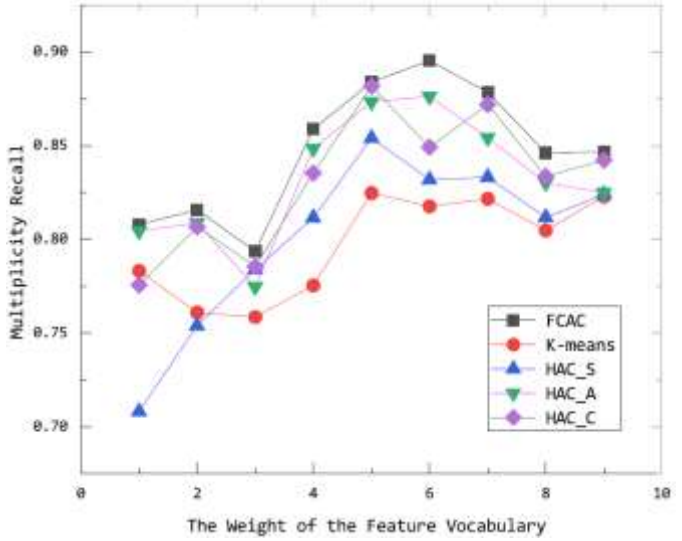

Fig. 4. Recall curves of different methods on the Reu-Te dataset.

4 Conclusions and Future Works

In this paper, we propose an overlapping clustering algorithm that uses FCA to convert documents into concept vectors. Using the proposed method, the size of the vector space can be compressed effectively, and the memory and computational capabilities of the algorithm are reduced. The experimental results show that the effect of overlapping groups can be achieved, and the two indicators from Precision and Recall are better than the commonly used clustering algorithms. The method can be combined with multiple clustering algorithms, and the basic clustering algorithm can be adjusted for different data sets and different feature words parameters, thereby making the algorithm more flexible.

The method proposed in this paper has not considered the concept of Fuzzy Clustering. The current method only focuses on the classification of objects into multiple clusters, and does not know the "degree" of the articles

belonging to the cluster, but the "degree" of understanding the objects related to the cluster is very important. For example, an article describes football and basketball. Assume that after overlapping clustering, both the “basketball” and “football” clusters have this article object. However, we are unable to observe from the clustering results this article describes the degree of football. That is, if an object is more related to a feature word, it needs to be described by "degree of membership."Therefore, the degree of relevance between objects and clusters deserves further study.

In addition, this study has not yet considered the part-of-speech and semantics of the document feature words, combined WiKipedia or WordNet to establish the semantics for the feature words, and made the clustering results closer to reality. For example, if a feature word such as "dollar" is selected, but "dollar" is not suitable for naming the cluster, you can consider using WordNet to find the cluster name whose superlative "commerce" gives more meaning to the documents. At the same time, we can also try to find the semantic differences between similar feature words through WiKipedia to increase the accuracy of clustering results. This will also be one of the main contents of future attention and further research.

5 Acknowledgement

I should like to express my deepest gratitude to all those whose kindness and advice have made this work possible. This work is supported by the Key Research and Development Program of Shandong Province (2015GGX101009).

References

Ackermann M R, Blömer J, Kuntze D, et al. Analysis of Agglomerative Clustering[J]. Algorithmica, 2014, 69(1):184-215.

Alelyani S, Tang J, Liu H. Feature Selection for Clustering : A Review[J]. Encyclopedia of Database Systems, 2013, 21(3):110-121.

Amigo E, Artiles J, Gonzalo J. Combining Evaluation Metrics with a Unanimous Improvement Ratio and its Application to the Web People Search Clustering Task[J]. 2009.

Bashir S, Afzal W, Baig A R. Opinion-Based Entity Ranking using learning to rank[J]. Applied Soft Computing, 2016, 38:151-163.

Bonchi, Francesco, Gullo, et al. Core decomposition of uncertain graphs[M]. 2014.

Capó M, Pérez A, Lozano J A. An efficient approximation to the K -means clustering for massive data[J]. Knowledge-Based Systems, 2017, 117:56-69.

Castellanos A, Cigarrn J, Garca-Serrano A. Formal concept analysis for topic detection[J]. Information Systems, 2017, 66:24-42.

D’Hondt J, Vertommen J, Verhaegen P A, et al. Pairwise-adaptive dissimilarity measure for document clustering[J]. Information Sciences, 2010, 180(12):2341-2358.

Dillon, Inderjit S, Modha, et al. Concept decompositions for large sparse text data using clustering[J]. Machine Learning, 2001, 42(1-2):143-175.

Fan Z, Chen S, Zha L, et al. A Text Clustering Approach of Chinese News Based on Neural Network Language Model[J]. International Journal of Parallel Programming, 2016, 44(1):198-206.

Handfield L F, Chong Y T, Simmons J, et al. Unsupervised Clustering of Subcellular Protein Expression Patterns in High-Throughput Microscopy Images Reveals Protein Complexes and Functional Relationships between Proteins[J]. PLoS Computational Biology,9,6(2013-6-13), 2013, 9(6):e1003085.

Kou G, Peng Y, Wang G. Evaluation of clustering algorithms for financial risk analysis using MCDM methods[J]. Information Sciences, 2014, 275(11):1-12.

Krishna K, Murty M N. Genetic K-means algorithm[M]. IEEE Press, 1999.

Kuo T F, Yajima Y. Ranking and selecting terms for text categorization via SVM discriminate boundary[C]// IEEE International Conference on Granular Computing. IEEE, 2010:496-501 Vol. 2.

Lavalle S M, Branicky M S. On the Relationship between Classical Grid Search and Probabilistic Roadmaps[J]. International Journal of Robotics Research, 2004, 23(23):673-692.

Li W, Joo J, Qi H, et al. Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph[J]. IEEE Transactions on Multimedia, 2017, 19(2):367-381.

Lv Y, Ma T, Tang M, et al. An efficient and scalable density-based clustering algorithm for datasets with complex structures[J]. Neurocomputing, 2016, 171(C):9-22.

Moertini V S, Suarjana G W, Venica L, et al. Big Data Reduction Technique using Parallel Hierarchical Agglomerative Clustering[J]. Iaeng International Journal of Computer Science, 2018, 45(1):188-205.

Mulder D, Wim. Optimal clustering in the context of overlapping cluster analysis[J]. Information Sciences, 2013, 223(223):56-74.

Omran M G H, Engelbrecht A P, Salman A. An overview of clustering methods[J]. Intelligent Data Analysis, 2007, 11(6):583-605.

Pei X, Chen C, Gong W. Concept Factorization With Adaptive Neighbors for Document Clustering[J]. IEEE Trans Neural Netw Learn Syst, 2016, PP(99):1-10.

Pérez-Suárez A, Martínez-Trinidad J F, Carrasco-Ochoa J A. A review of conceptual clustering algorithms[J]. Artificial Intelligence Review, 2018(6):1-30.

Popat S K, Emmanuel M. Review and Comparative Study of Clustering Techniques[J]. International Journal of Computer Science & Information Technolo, 2014.

Shah N, Mahajan S. Document Clustering: A Detailed Review[J]. International Journal of Applied Information Systems, 2012.

Sironi A, Turetken E, Lepetit V, et al. Multiscale Centerline Detection[J]. IEEE Trans Pattern Anal Mach Intell, 2016, 38(7):1327-1341.

Tseng V S, Kao C P. Efficiently mining gene expression data via a novel parameterless clustering method[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2005, 2(4):355-365.

Wang Y, Wang Y X, Singh A. Graph Connectivity in Noisy Sparse Subspace Clustering[J]. Computer Science, 2016, 17(3):689–708.

Wille R. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts[J]. Orderd Sets D Reidel, 1982, 83:314-339.

Yu H, Zhang C, Wang G. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-Based Systems, 2016, 91(C):189-203.

Zhang X, Xu Z. Hesitant fuzzy agglomerative hierarchical clustering algorithms[J]. International Journal of Systems Science, 2015, 46(3):562-576.