One Optimized Choosing Method of K-Means Document Clustering Center

Hongguang Suo, Kunming Nie, Xin Sun, and Yuwei Wang

School of Computer and Communication Engineering, China University of Petroleum, Dongying, China
{suohg,nkm1985,sunxin1000,wyw1101}@163.com

Abstract. A center choice method based on sub-graph division is presented. After constructing the similarity matrix, the disconnected graphs can be established taking the text node as the vertex of the graph and then it will be analyzed. The number of the clustering center and the clustering center can be confirmed automatically on the error allowable range by this method. The noise data can be eliminated effectively in the process of finding clustering center. The experiment results of the two documents show that this method is effective. Compared with the tradition methods, F-Measure is increased by 8%.

Keywords: Document Clustering, K-means, Initial Center, Sub-graph Division.

1 Introduction

K-means has better scalability and higher implementation efficiency to the application of document clustering, and it can achieve good results and is superior to hierarchical clustering[1], but it is difficult to determine the cluster center and it is sensitive to the isolated point in document set. Aiming at the characteristic that the cluster center of the K-means needs to be assigned, there have been some improved methods[2].

Literature [3] made use of genetic algorithm to optimize K which is the number of initial center; Literature [4] mentioned that a new program for choosing the initial clustering center has been raised by the global k-means method, which adds a dynamic clustering center through making use of global searching process. Liu Yuanchao made use of the improvement of the Maximin Principle to decide the clustering number and clustering center[5]. In literature [6] the clustering center was determined by improving the CBC committee algorithm, and the shortcoming of the method is that it has a high time complexity and it needs to use the hierarchy algorithm when selecting a committee, which depresses the accuracy to some extent. Additionally, the algorithm needs to assign the number of the cluster manually, which is always a difficult task.

This paper presents a new initial center choice method, which aims at the characteristic that the K-means algorithm needs to assign the initial clustering center, based on sub-graph division.

2 Initial Center Choosing Method Based On Sub-graph Division

After setting up the similarity matrix of the text, we select the entire document nodes as the vertexes of the graph. For the two documents whose similarity is bigger than the

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 490–495, 2008.

Fig. 1. Changing Process of Sub-graph (θ=0.25 → θ=0.20)

current similarity threshold, we connect a line between the two corresponding nodes in the graph. Thus a disconnected graph will be formed (if the threshold is too small, the graph may be a connected graph).

In the descending process of the similarity threshold, when the threshold value is proper, some sub-graphs exceeding a dedicated text number in the disconnected graph will come forth. And these sub-graphs perform as: the similarity between the documents of sub-graphs is high; meanwhile, the similarity between the documents included in different sub-graphs is smaller than the current similarity threshold. The purpose of the text clustering is that it makes the similarity of data point in the same cluster maximum and makes the similarity of data point in the different cluster minimum. It is feasible to take these sub-graphs as the candidate initial cluster center.

In the changing process of the threshold, if the mutual similarities of the document of one cluster are lower in the similarity matrix, when the threshold falls to a dedicated value, a sub-graph will be formed possibly between the documents of this cluster, and then we can consider the vector center of the documents in this sub-graph as the candidate clustering centers, which reduces the data noise.

2.1 Preprocess of Clustering

During the document clustering, preprocess work is very important. The usual processing steps include: segmentation, stop word removal, word frequency statistic, feature selection and building vector space model.

In the process of our experiment, we select different number of high frequency words to experiment, and find that the use of 50 high frequency words would have the best experiment result. So, we select the first 50 high frequency words of each clustering document as the feature words of the clustering. Collecting statistic of all the words of the total document and taking them as column, and build the vector space model. The computing of text similarity uses cosine formula.

$$sim\ (dt\ ,dm\ ) = \frac{dt' \times dm}{\| dt \| \times \| dm \|} \tag{1}$$

2.2 Algorithms

The first stage: Cluster center

The main process of the algorithm is, firstly, we find out all the sub-graphs of the text set and sub-graphs formed in the current threshold from these sub-graphs. We deal with them and then depress the threshold, loop until satisfying the end situation. The input of the algorithm is the storage structure of the adjacent table of N clustering documents and the threshold of similarity is θ, β.

Step 1. Finding out all the sub-graphs of the text set by using of the depth graph traverse method, if the document numbers of all the sub-graphs are smaller than β, turn to **Step 5**;

Step 2. If N sub-graph *sub-graph*$\{W_1, W_2, W_3, \ldots, W_n\}$ are disposed for the first time, we take the N sub-graphs as the candidate initial cluster center of the text set, signature as *old*$\{M_1, M_2, M_3, \ldots, M_k\}$ and compute the vector center of each cluster center. Turn to **Step 5**, or else turn to **Step 3**;

Step 3. Collecting statistic of all the sub-graph *sub_graph*$\{W_1, W_2, W_3, \ldots, Wn\}$ produced by traverse, establishing the mapping between it and the old cluster center *old*$\{M_1, M_2, M_3, \ldots, M_{k1}\}$, and then dealing with these sub-graphs in *sub_graph*$\{W_1, W_2, W_3, \ldots, W_n\}$ respectively.

For the sub-graph that has mapping relationship $W_i \rightarrow \{M_i, M_j \ldots\}$ with the old cluster center, if the increased document number of sub-graph W_i, num(W_i)-num(M_i+M_j) is smaller than β, we consider that the cluster center of M_i, M_j doesn't change, and we diverse the newly increased document to the old cluster center having the higher similarity. If the increased document number is bigger than β, we need to judge whether the new element is a new cluster center or not.

For the sub-graph that doesn't have mapping relationship with the old cluster center, taking the sub-graph as the candidate cluster center and turn to **Step4**;

Step4. Firstly, we must judge whether the similarity between each candidate cluster center and other candidate cluster center that doesn't have mapping relationship with the old cluster center is bigger than θ or not, if the similarity is bigger than θ, the two candidate cluster centers can be emerged as one cluster center M_{k+1}, meanwhile, we must judge whether the M_{k+1} is a new cluster center entered to *old*$\{M_1, M_2, M_3, \ldots, M_{k1}\}$. If the maximum of the similarity between M_{k+1} and old cluster center is small than the present threshold of sub-graph division, M_{k+1} needs to be entered to *old* $\{M_1, M_2, M_3, \ldots, M_{k+1}\}$;

Step 5. Reducing the similarity threshold of the sub-graph division by 0.05;

Step 6. Repeating the process **Step 1 ~ Step 5** until the threshold reducing to the dedicated value η, the judged cluster center *old*$\{M_1, M_2, M_3, \ldots, M_k\}$ will output and the program is end.

The sub-graph in the algorithm refers to the graph that the number of the interrelated document is over β. In the forth step, if the candidate cluster center W_i and the old cluster center M_i are in the same cluster, after the vector center of W_i being emerged, the similarity between W_i and M_i will be bigger than before, and the similarity is bigger than the similarity between any of the element in W_i and the old cluster center M_i, so, we take the candidate cluster center, which has a lower similarity than the present sub-graph division threshold, as a new cluster center.

The second stage: Clustering

We conduct a k-means clustering by use of the clustering centers, which are found out in the first stage.

Step 1. Take the k cluster center producing in the first stage as the initial cluster center;

Step 2. Assign each document to the most similar cluster according to the average of object in the cluster;

Step 3. Update the average value of the cluster;
Step 4. Repeat **Step 1～Step 3** until the cluster division does not change again;

2.3 Time Complexity Analysis of Algorithms

Prior to the implementation of the algorithm, we need to build the similarity matrix of the document. The time spending is bigger, but most of the document clustering algorithms needs to build the similarity matrix in advance [6]. In the process of building similarity matrix, the storage structure of adjacent table can be built according to the dedicated threshold. The time complexity of the graph's traverse algorithm is $O(n+e)$ in the implementation process, taking d as the feature dimension of text set and taking k as the number of cluster center. The time complexity of the second step is nd and the time complexity of dealing with the newly created sub-graph is $k^2 d$, so the time complexity of the algorithm is $O(n+e+nd+k^2 d)$。 The time complexity of the algorithm can be increased with the increasing of the dimension numbers. Consequently, when we handle the large data set, the time complexity of the algorithm is high.

3 Experiments

We conduct a series of experiments. The first group selects the different articles that have already been categorized from www.sina.com.cn. These articles are classified artificially, so it is convenient to compare the test result. There are 7 classes of documents: studying abroad, real estate, music, automobile, military affairs, college entrance examination, sport. We take out 20 documents for experimentation from each category of these documents. The second group makes use of the classification corpus which is provided by Lee Lurong of FuDan University, selecting 6 classes from these corpus for the experiment: Energy、Electronics、Medical、Communication、Philosophy、Literature.

The evaluation criteria of the experiment result used most commonly F-measure values [7]. Let *P* be average precision. Let *R* be average recall. F-measure is defined to be *2RP/(R+P)*.

Table 1. The Experiment Result of the Algorithms

True k values	The Number of Initial Centers Generated on Different Group of Corpus					
	1	2	3	4	5	6
2	2	2	2	2	2	2
3	3	3	3	3	2	3
4	4	4	4	4	4	3
5	5	5	4	5	5	5
6	5	5	6	5	5	6

The threshold of the algorithm is selected according to the experiment experience. When θ is smaller, there will be many density areas and the mutual similarity of these areas will be lower. We conducted several experiments by taking definite standard text set as the training text set. After training, we let θ=0.25, β=N/5 in the process of the experiment (N is the document number of each cluster).

For the different values of k, we use different k cluster's combination to conduct experiment and each k is conducted 6 groups of different experiments. We can see that the experiment results are unanimously with the standard result from table 1. When analyzing the cluster that the division number is lack, it is easy to find that the document theme distributing of the cluster is too loose. When k=6, the military cluster is not found.

We select 2～6 clusters from the 7 clusters to conduct experiment, and three group of experiments are conducted to every k. Corresponding to each group of experiment for each K, the right column data is the cluster result that using the sub-graph division method.

Table 2. Comparision of Two Different Cluster Initial Center Choosing Method (Our Method B)

Experiment Times	Clusters of Experimental Text Set									
	2 Clusters		3 Clusters		4 Clusters		5 Clusters		6 Clusters	
	A	B	A	B	A	B	A	B	A	B
First Group	0.90	1.00	0.88	1.00	0.85	0.92	0.83	0.90	0.74	0.81
Second Group	0.95	1.00	0.83	0.95	0.80	0.86	0.72	0.90	0.75	0.89
Third Group	0.90	0.92	0.87	0.93	0.80	0.82	0.83	0.84	0.71	0.81

From table 2, we can see that the experiment result by the method of automatically determining the cluster center is superior to the method of artificially designating the cluster center. In the first stage of the cluster, the vector center of the cluster are the combination of the documents that are divided into the sub-graph, so each cluster in the test result has a high recall rate. If the number of selected cluster center is equal to the number of the original cluster of document set, the precision of each cluster will be promoted, so the accuracy of the cluster result is higher.

The second experiment (Table 3.) is a comparison between our method in this paper and the original cluster method C. In the original method, the cluster center is generated randomly, and we select the average of the three experiments as the result of this experiment.

Table 3. Comparison of Our Method With Original Clustering Method (Our Method B)

Experiment Times	Clusters of Experimental Text Set							
	2 Clusters		3 Clusters		4 Clusters		5 Clusters	
	B	C	B	C	B	C	B	C
First Group	0.78	0.76	0.64	0.66	0.69	0.54	0.4	0.49
Second Group	0.95	0.73	0.68	0.67	0.59	0.45	0.54	0.44
Third Group	0.69	0.52	0.81	0.59	0.57	0.62	0.59	0.55

From the experiment results of table 3, in most cases, this method can achieve better results. The F-measure value of the automatically determining the cluster center method is promoted by 8% than the original K-means cluster method. We analyze the reasons for several poorer test results and it is mainly because that the selected feature words of experiment are correlated with each other, such as electronics cluster and communication cluster. The other reason is that the length of individual document is too short. When selecting the key words of clustering, the effective words that can distinguish the document are less.

4 Discussion and Conclusions

A method of determining the cluster center is presented, by which the potential cluster center can be discovered using sub-graph division. In the process of searching the cluster center, it removes the noise data successfully. The method, which can improve the results of the cluster remarkably, is proved effective. When our method applying in the short content subjects, the effect of the cluster will decline because of the small amount of information contained in the text. In future, we will continue to research it deeply.

References

1. Zhao, Y., Karypis, G.: Criterion Functions for Document Clustering Experiments and Analysis. J. Department of Comp. Sci & Eng University of Minnesota, 01–40 (2001)
2. Khan, S.S., Ahmad, A.: Cluster center initialization algorithm for K-means clustering. J. Pattern Recognition Letters 25(11), 1293–1302 (2004)
3. Casillas, A., González de Lena, M.T., Martínez, R.: Document clustering into an unknown number of clusters using a Genetic Algorithm. In: A. International Conference on Text Speech and Dialogue TSD, 43–49 (2003)
4. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means algorithm. Pattern Recognition. J 23, 451–461 (2003)
5. Liu, Y., Liu, X., Liu, B.: An adapted algorithm of choosing initial values for k-means document clustering. J. High Technology Letters 16(1), 11–15 (2006)
6. Zhao, W., Wang, Y., Zhang, X., Li, J.: Variant of K-means algorithm for document clustering: optimization initial centers. J. Computer Applications 25(9), 2037–2040 (2005)
7. Liu, Y., Wang, X., Xu, Z., Guan, Y.: A Survey of Document Clustering. J. Journal of Chinese Information Processing 20(3), 55–62 (2005)