Optimization of Initial Centroids for K-means Algorithm Based on Small World Network

Shimo Shen, Zuqiang Meng

College of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
shenshimo@126.com, mengzuqiang@163.com

Abstract. K-means algorithm is a relatively simple and fast gather clustering algorithm. However, the initial clustering center of the traditional k-means algorithm was generated randomly from the dataset, and the clustering result was unstable. In this paper, we propose a novel method to optimize the selection of initial centroids for k-means algorithm based on the small world network. This paper firstly models a text document set as a network which has small world phenomenon and then use small-world's characteristics to form k initial centroids. Experimental evaluation on documents croups show clustering results (total cohesion, purity, recall) obtained by proposed method comparable with traditional k-means algorithm. The experiments show that results are obtained by the proposed algorithm can be relatively stability and efficiency. Therefore, this method can be considered as an effective application in the domain of text documents, especially in using text clustering for topic detection.

Keywords: k-means; text clustering; small world network; SNN

1 Introduction

Clustering is useful in a wide range of data analysis fields, including data mining, document retrieval, image segmentation, and pattern classification. The goal of clustering is to group data into clusters so that the similarities among data members within the same cluster are maximal while similarities among data members from different clusters are minimal [1].

Among the different classes of clustering algorithms, the distance-based methods are the most popular methods in a wide variety of applications, while two best-known methods for distance-based clustering are the partition clustering algorithm and hierarchical clustering algorithm. K-means algorithm is one of the most widely used distance-based partitioning algorithms, and that separates data into k mutually excessive groups [2].

K-means algorithm is very popular because of its ability to cluster huge data set and its simplicity. However, K-means algorithm is quite sensitive to the initial cluster centers picked during the clustering, and does not guarantee unique clustering because different results obtained with randomly chosen initial centroids. The final cluster centroids may not be the optimal ones because of the k-means algorithm converges

into local optimal solutions. The initial centroids affect the quality of k-means algorithm, especially in documents clustering. Therefore, it is quite important for k-means algorithm to have good initial cluster centroids.

There are several methods to reduce the sensitivity of initial centroids picked during clustering proceeding.

Cutting,etl[3] use group average agglomerative clustering algorithm to select initial centroids.

Likas[4] proposed the global k-means algorithm which is an incremental approach to clustering which dynamically adds one cluster center at a time through a deterministic global search procedure consisting of N (with N being the size of the dataset)executions of the k-means algorithm from suitable initial positions.

Arthur and Vassilvitskii[5] proposed k-means++ algorithm, a specific way of choosing centers for the k-means algorithm, which choose the centers by weighs the data points according to their squared distance from the closest center already chosen, and improves both the speed and the accuracy of k-means. However, the k-means++ clustering method sometimes generates bad clusters because it depends on the selection of the first initial center. The first initial center is chosen uniformly at random from data points set.

Onoda.etl [6] proposed a seeding method based on the independent component analysis for the k-means clustering method. This method can be summarized as two steps. First, k independent components IC_i ($i=1,\ldots,k$) were obtained from given data x. Second, the initial centers were selected according by k independent components. This method is useful for Web corpus.

In the paper, we propose a novel method to optimize the selection of initial centroids for k-means algorithm based on small world network. A novel network which has the small world phenomenon is built by connecting similar documents, and then we use small-world's characteristics to form k initial centroids for k-means algorithm.

The rest of the paper is organized as follows. Section 2 introduces Vector space model and small world network. Selecting initial centroids for k-means algorithm based on small world network is proposed in Section 3. Section 4 is the algorithm of improved k-means clustering algorithm. Section 5 presents extensive experimental results. Conclusion follows in Section 6.

2 Preliminaries

2.1 Text Document Representation

Vector Space Model. The vector space model is used to represent the document as vector with a list of words and a list of weights of each word occurs. The vector of document is defined as:

$$d_i = \{w_{ij}\},\ j = 1.2.....m, i = 1.2.....n \tag{1}$$

Where

w_{ij} is the weight of jth term in ith document
m is the total number of unique terms appearing in the document
n is the number of documents in the document collection.

With the vector space model, the text data is converted into structured data that computer can handle, and the similarity between the two documents is converted into the similarities between the two vectors.

Cosine Similarity computation between documents. The cosine similarity is often used to measure the similarity between two documents in text mining. For text matching, the document vectors d_i and d_k are the TF-IDF vectors of the documents. Cosine similarity between d_i and d_k is defined as:

$$SC(d_i, d_k) = \frac{\sum_{j=1}^{t} d_{kj} d_{ij}}{\sqrt{\sum_{j=1}^{t} (d_{ij})^2} \sqrt{\sum_{j=1}^{t} (d_{kj})^2}} \tag{2}$$

Where
d_{ij}(or d_{kj}) is the jth term in d_i (or d_k) document.

The cosine measure gives values between 0 and 1, and the more common words of documents have, the bigger of cosine similarity is.

Definition 1. **Cosine Similarity Threshold λ.** We suppose that cosine similarity threshold λ to measure the similarity between two documents. While cosine similarity of documents is greater than λ, these documents are considered as the same class, and vice versa. This method is good at handling noise, outliers and reducing the dimension of the network.

Shared Nearest Neighbor Similarity. SNN (shared nearest neighbor similarity) [7], an approach to similarity two documents, with the value is the number of the same points that similar to the two documents.

For two points, x and y, the shared nearest neighbor definition of similarity between in the manner indicated by algorithm as follows.

1) Find the k-nearest neighbors of x and y;
2) If x and y are not among the k-nearest neighbors of each other then
3) Similarity $(x, y) \leftarrow 0$;
4) Else
5) Similarity $(x, y) \leftarrow$ number of shared neighbors;
6) End if.

SNN can be any positive integer, the more SNN is, the more relevant to each other the two texts are.

Definition 2. SNN Threshold ν. We suppose that SNN threshold ν to measure the similarity between two documents.

2.2 Small World Network

The small world phenomenon came from the research of sociologist Milgram carried out in 1967 to trace the shortest path in the U.S. social network.

Watts spent a deep studying on the small world phenomenon in 1998, and proposed that a small world network has characteristics of high concentration and short path [8].

Cluster Coefficient. The cluster coefficient [9] C is defined as follows. Suppose that a vertex i has K_i neighbours; then at most $K_i (K_i - 1)/2$ edges can exist between them (this occurs when every neighbour of i is connected to every other neighbor of i). φ_i denote the actual number of edges that exist between K_i and neighbor nodes. The cluster coefficient of vertex i is C_i:

$$C_i = \frac{\varphi_i}{K_i (K_i - 1)/2}. \tag{3}$$

The cluster coefficient of the whole network is C:

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{4}$$

Network with small-world nature have high cluster coefficient feature.

Power law distribution of node degrees features. Ferrer's[8] study also showed that: the lexical co-occurrence network also has a scale-free features ,the degree of network node distribution is close to the power law. The probability $P(k)$ of having a node with degree k scales as $P(k) \approx k^{-r}$, where r is a constant. It reflects that the status of the connection between each node in the network is severe heterogeneity. In this network only a very small number of nodes have many connections with other nodes and become hub nodes, and most of nodes have few connections. Hub nodes play a leading role in the operation of scale-free network. So we use this feature form the k seeds.

3 Selecting Initial Centroids for K-means Algorithm

The three major steps of the proposed selection of the initial centroids approach are described as follows.

Step 1. Model a document set as document network based on vector space model. Document is represented as the document space vector, and the similarity of documents was calculated by vector way. Two documents which similarity is greater than threshold λ were connected. Form document network and express as $G= \{V, E\}$, where V is a set of vertex and E is a set of edges or connection between documents. In

the document network, a vertex corresponds to a document that connected with one document at least. An edge $e = (i, j) \in E$ stand for a connection between vertices i and j.

When drawing a graph of document network, the Fruchterman-Reingold Algorithm [10] was used to. Fig.1shows document network graph obtained after processing about 842 documents.

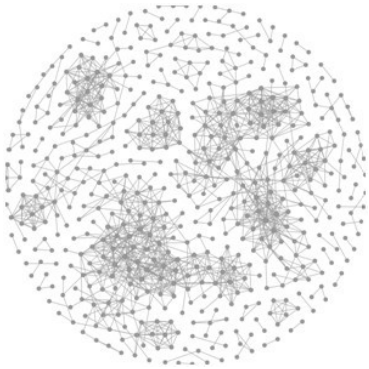

Fig. 1. Document network

In Fig.1, black nodes are documents, and two documents are linked while the similarity of them is greater than threshold λ. The cluster coefficient of the document network is C=0.382, while random network is C_{random} =1.55*10^(-4). It can be seen that $C >> C_{random}$, and the document network has high cluster coefficient. The distribution of degrees of document network is shown in Table 1.

Table 1. Distribution of nodes' degree

degree	2	3	4	5	6	7	8	9	10	11	12	13	14
number	112	53	37	25	15	26	21	11	5	7	4	1	1
P(k)	0.13	0.06	0.04	0.03	0.017	0.03	0.025	0.013	0.005	0.008	0.004	0.0011	0.001
$k^{-2.3}$	0.2	0.08	0.04	0.03	0.016	0.011	0.008	0.006	0.005	0.004	0.003	0.003	0.002

From Table 1, it shows that the probability P(k) of having a node with degree k scales as $P(k) \approx k^{-2.3}$, and the degree of network node distribution is close to the power law.

Fig.1 and Table 1 illustrate that the document network has a high cluster coefficient characteristic and the node degree distribution is close to the power-law distribution features which is of small world network. According to the distribution of power-law distribution, we can see only a small number of nodes with many connections to other nodes. These nodes can be thought as common documents in the network, and these documents play a leader role in the network. The content of these common documents are the common content that most probably appear in the dataset. So we can use the center of these nodes to represent the initial centroids of k-means algorithm.

Step 2. Split network based on its features with SNN. In order to get these common documents, we can segment network based on its features with the ν, and form network as show in Fig.2.

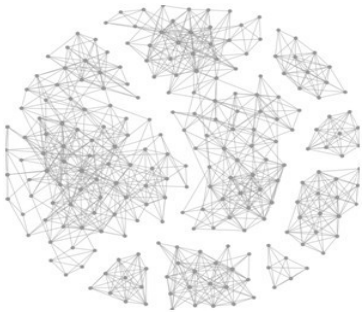

Fig. 2. Common document network

Step 3. Form initial centroids. Compute the centroid of each subgraph and make these centroids as initial centroids of clusters.

4 Proposed Algorithm

We use the method which proposed in section 3 to improve the k-means algorithm, and the algorithm is described as follows.

1. Set the cosine similarity threshold λ. Make a Link between two documents which similar is greater than λ, and form a network.
2. Initialize the classification sample collection, $M_1 = M_2 = \ldots = M_k = \{\}$.
3. Set the SNN threshold ν.
4. Find the maximum degree of nodes from sample collection and join in set M_1.
5. Identify the nodes with the nearest neighbor similarity is not less than ν in the sample set, and added to the collection M_1.
6. Repeat step 4 on the collection of all data points in the M_1 is no longer changes until the collection of M_1.
7. Delete the nodes that in set M_1 from sample collection.
8. Repeat steps 4-6, until M_k is generated.
9. Calculate the centers of $M_1, M_2, \ldots, M_k$ and make these centers as initial centroids of clusters.
10. Form k clusters by assigning each point to its closet centroid. Then recomputed the centroid of each cluster.
11. Repeat step10 until centroids do not change.

5 Experimental Analysis

We use documents datasets that are downloaded from `http://www.163.com` (November 27, 2011) for evaluation in our experiments. This dataset has 5 categories and conclude 842 documents. The categories come from the `http://www.163.com`, and contain the following types: News, Reading, Education, Science and Technology.

Experiment compared the traditional k-means and the proposed k-means on various evaluation measures. Details are described as follows.

5.1 Evaluation Measures

To evaluate the proposed approach, we employed three measures of total cohesion, accuracy and recall to evaluate the quality of clusters generated by different methods. The definitions are as follows.

Total Cohesion. It is an objective function, which measures the quality of a clustering, in this experiment our objective is to maximize the similarity of the documents in a cluster to the cluster centroid, this quantity is known as the cohesion of the cluster. The total cohesion is defined as follows:

$$\text{Total Cohesion} = \sum_{i=1}^{k} \sum_{x \in C_i} cosine(x, c_i) \tag{5}$$

where
k is the number of cluster
C_i is the ith cluster
$\mathbf{c}_i$ is the centroid of the ith cluster and $\mathbf{x}$ is the node that belong to C_i.

Purity. Purity measure the degree to which each cluster consists of objects of a single. For each cluster, the class distribution of the data is calculated first, i.e., for cluster C_i and class G_j of text .The corresponding purity of C_i and G_j are defined as:

$$precision(C_i, G_j) = n_{ij} / n_i \tag{6}$$

where
n_i is the number of objects in cluster C_i
n_{ij} is the number of objects of class G_j in cluster C_i . The purity of cluster C_i is as:

$$precision(C_i) = \max_j precision(C_i, G_j) \tag{7}$$

The overall purity of a clustering is as:

$$precision = \sum_{i=1}^{k} \frac{n_i}{N} precision\ (C_i) \tag{8}$$

where
N is the total number of text, k is the number of cluster.

Recall. Recall measure the extent to which a cluster contains all objects of a specified class. Similarly as purity, the recall rate of this method can be defined as:

$$recall = \sum_{i=1}^{k} \frac{n_i}{N} recall(C_i) \tag{9}$$

where

$recall(C_i) = \max_j recall(C_i, G_j)$ $recall(C_i, G_j) = n_{ij} / n_i$.

5.2 Setting of Threshold μ and ν

In order to find out more specific value for parameter μ and ν, we perform the proposed algorithm, with different values of μ and ν and the results are shown in Fig.3 and Fig.4.

Fig.3 shows the result of different values of λ when ν=2.

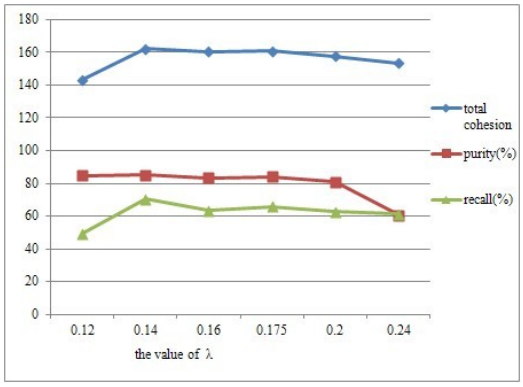

Fig. 3. The result of different values of λ

Fig.3 shows that better results are gotten when the value of λ is between 0.14 and 0.2. Specially, the three evaluation values have the highest value when λ=0.14. May be 0.14 is the suitable value to measure the similarity between two documents.

Fig.4 shows that the result of different values of ν when λ=0.14. The best results are gotten when the value of ν is 2.

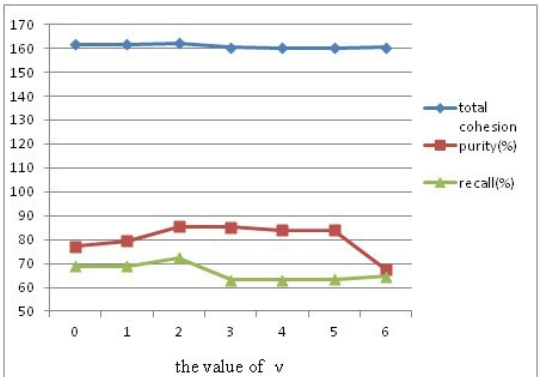

Fig. 4. The result of different values of v

5.3 Evaluation of Clustering Efficiency

To further evaluate the proposed approach, the Comparison of traditional k-means algorithm [2] and proposed algorithm with the best parameters on total cohesion, purity and recall have done. The traditional k-means algorithm is run 50 times, and the three better results are selected from all of the run results and are listed in Table 2. The proposed algorithm is run only one time because it can get a unique initial seeding. It is observed that proposed algorithm has high values in every evaluation.

Table 2. Cluster quality of different cluster algorithms.

The name of algorithm	Totalcohesion	purity(%)	Recall(%)
First of Basic k-means algorithm	142.54	47.41	50.37
Second of Basic K-means algorithm	146.7	63.3	53.7
Third of Basic K-means algorithm	145.55	53.5	48.4
Proposed algorithm	162.56	85.14	70.5

There are two aspects which can prove the stable results can be obtained by the proposed algorithm. On the one hand, theoretically, for the same text corpus, with any sequence of corpus and at any time to run, the document network which the corpus are model as to is constant, and then the common document networks are stable. Thus the center of these common document network (the initial centroids for k-means clustering algorithm) are stable， Therefore, the results are obtained by this method is stable. On the other hand, the experiment shows the result of proposed method that runs several times on the same corpus is the same.

6 Conclusion

In this paper, we improve the k-means algorithm in selection of initial centroids based on small world network. The proposed method solved the traditional k-means algorithm is sensitive to the initial centroids. The same clusters were obtained by the proposed method, using the same data, with any sequence of documents. And also the purity of clusters has been greatly improved with the proposed method. Although additional work for selection of initial centroids increase time complexity to O ($n^{\wedge 2}$), n is the size of document collection, the effective and steady clusters with O ($n^{\wedge 2}$) remain can be applied to practice. Therefore, this method can be considered as an effective application in the domain of text documents, especially in using text clustering for topic detection.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61063032) and the Science Foundation of Guangxi Education Department (No. 201012MS010).

Reference

1. Feldman, R. and Sanger, J. *The text mining handbook.* Beijing [M]: posts & telecom press, 2009:82-92.
2. Aggarwal, C. and Zhai, C. A survey of text clustering algorithms [M]: Springer, 2012:77-128.
3. Cutting, D., Karger, D. and Pedersen, J. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections[C]. ACM SIGIR Conference, 1992.
4. Likas, A., Vlassis,N. and Jakob,J.V. The global k-means algorithm algorithm [J]. Pattern Recognition, 2003, 36(2): 451–461.
5. Arthur,D. and Vassilvitskii,S. K-means++: the advantages of careful seeding[C]. ACM-SIAM Symposium, 2007.
6. Onoda, T., Sakai, M. and Yamada, S. Independent Component Analysis based Seeding method for k-means Clustering[C]. IEEE/WIC/ACM Conference, 2011
7. Tan, P., Steinbach, M. and Kumar, V. *Introduction to Data Mining* [M]: posts & telecom press.2011:385-387.
8. Cancho, R.F. and Sole, R.V. The small world of human language. The Royal Society of London，Biological Sciences(Series B), 2001, 268(1482) : 2261—2265.
9. WaRs, D.J. and Strogatz. S.H. Collective dynamics of small-world networks [J]. Nature，1998，393(6684): 440-442.
10. Thomas, M.J.F and Edward, M.R. Graph Drawing by Force-directed Placement [J]. Software: Practice and Experience. 1991, 21(11):1129–1164.