Text Clustering Algorithm Based on Random Cluster Core

Long-Jun HUANG[1,a], Meng-Zhen CHENG[1] and Yao XIAO[2]

[1]*School of Software, Jiangxi Normal University, Jiangxi, 330022*
[2]*Baidu Time Network Technology (Beijing) Co., Ltd, Beijing, 100081*
[a] *huanglong98614@163.com*

Abstract. Nowadays clustering has become a popular text mining algorithm, but the huge data can put forward higher requirements for the accuracy and performance of text mining. In view of the performance bottleneck of traditional text clustering algorithm, this paper proposes a text clustering algorithm with random features. This is a kind of clustering algorithm based on text density, at the same time using the neighboring heuristic rules, the concept of random cluster is introduced, which effectively reduces the complexity of the distance calculation.

1 Introduction

Cluster analysis as an important means of text information mining, is widely used in log analysis, statistics of public opinion and other areas .In recent years, the means of text and word frequency word bag clustering emerge in endlessly, these algorithms extracted a lot of valuable information available from the vast amounts of data, creating a huge amount of wealth.However, with the rapid development of computer and Internet industry, many companies store data reaches TB or even PB level, such background puts forward higher requirements for data mining algorithms. Some traditional text clustering methods cannot adapt to the massive and high dimension data mining tasks, because of the bottleneck of performance, so it is necessary to design some efficient and agile clustering methods[1].

Experts at home and abroad, according to the characteristic of text mining, many classical clustering scheme applied to the text field, among them, the k-means algorithm proposed by Queen Mac provides an efficient random clustering, but the method of obtaining the center of mass in iteration makes the clustering result very vulnerable to outliers. And The DBSCAN algorithm proposed by Arlia Massimo and Domenica accurately describes the density of clustering objects, and excludes the interference of noise points, but its O() time complexity often makes it helpless in the face of massive data clustering task[2]. In this paper, we combine the advantages of these two clustering schemes, and use the concept of relative distance to effectively reduce the time complexity of the distance calculation[3]. At the same time, this paper introduces a heuristic rule, which greatly reduces the computation of the distance of the whole object. This algorithm can not only efficiently get the clustering result in a short time, but also can ensure the correctness of the conclusion. Under the conditions of meeting the user's requirements, minimizing the cost of distance computation and the effect of noise points[4].

2 Related Notion

2.1 Text distance calculation based on DWC structure

Structure for storing a document's word bag information. After completing the full text scanning, remove the duplicated words and get a “dictionary”, which records the various words and numbers[5]. Referring to the dictionary, available for each document DWC structure [Doc:Word:Count] of the storage unit, the doc said document number, the word represents the number of a word in the dictionary, and the count said the number of times the term appears in the document. So a document can be seen as a collection of several DWC structures such as the above. Combined with the cosine similarity formula, the relative distance of the two documents can be easily obtained:

$$Directio\ (A,B) = CAiCBiCAj\ \ 2 \times CBj\,2 \tag{1}$$

Max cluster radius R: If the cluster space as a sphere, which means that the user can accept the biggest difference between the same object is 2R.

Minimum number of clusters N: If a cluster has fewer than N objects, then ignore this clustering result, and this cluster core is labeled as "illegal cluster core".

2.2 Random cluster core heuristic rule

In the clustering process, a heuristic rule is used to simplify the whole clustering process:

Transitivity of adjacent properties, that is, for the same cluster core A, If B and C all are close neighbors with A, then B and C are also neighbors (i.e., In the case of not calculating the distance between B and C, the B and C fall into the same class), this clustering algorithm is very efficient, when the cost of computation of distance between objects is very heavy.

3 Clustering Algorithm Based on Random Cluster Core

3.1. Characteristics of K-means and DBSCAN clustering algorithms

Among the numerous clustering algorithms, K-MEANS and DBSCAN are commonly used. In the process of repeated iteration, K-MEANS is constantly assigned each point to the nearest neighbor cluster core, and then through the average position of all the objects in a cluster to update the location of the cluster core, and finally get the clustering results. After scanning the all of the distance between the objects, the DBSCAN method excludes the interference of outliers, starting from the density of clusters, the high density of objects cluster into a class[6].

3.2 Clustering algorithm based on random cluster core

The method used in this paper is based on the two classical algorithms, which can ensure the clustering results to meet the requirements, as far as possible to retain the advantages of these two classic algorithms:

1) Requires the user to specify the two input parameters:

a.Max cluster radius R.

b.Minimum number of clusters N.

By using the random cluster core heuristic rule, we can know that the distance between any two objects and another object in the R can be considered as a class of the three objects.

2) Accept the user's input data, select a certain object as the first cluster core of cluster:

Like K-means this object can be specified by the user, can also be randomly selected[7].

3) Then imitate the DBSCAN based on the cluster core density algorithm,all objects are scanned once,and if any one of the distance between the target and the cluster core in R, it can be argued that the objects belonging to this cluster.

4) After the scan is complete, the cluster core is checked once, if the number of objects larger than N, it is confirmed that the formation of new clusters, if less than N ,it will be labeled as "illegal".

5) Then select an object that has not been labeled as a new cluster core, and repeat step (2) -(4) to generate a new cluster. When all objects are attributed to a cluster or labeled as "illegal", the algorithm ends.

In addition, you need to consider a situation: when the new cluster core scan an object, and the distance between this object and the current cluster core is less than R, it is found that this object has been classified as a certain cluster[8]. At this point to the precise classification, the object is assigned to a more recent cluster core. This makes the clustering results more accurate, the entire algorithm pseudo-code is described as follows:

Random Cluster Core Algorithm:

```
function randomCluster(List obj_list, int K,int R){
  While  There are no labeled objects {
    If  Random object not assigned{
        Generate new clusters and the object as the
                                              cluster core
        While   New objects are found at random and
                calculate the   distance to the cluster core.
            If   Within the distance K
              If   The object has been assigned{
                  Assign the object to the nearest cluster
              }else{
                  Assign the object to the current cluster
                  }
            }
          }
      }
}
```

Examples of iterative clustering results:

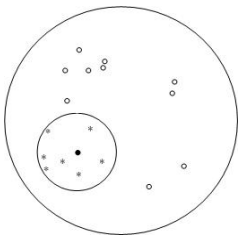

Figure 1. Clustering[A]

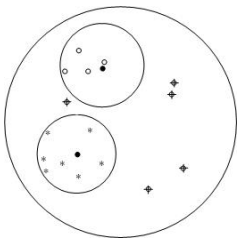

Figure 2. Clustering[B]

In the above two clustering figures, A figure shows the result of the once clustering iteration, and the B figure is the final clustering result. Clearly the object set is clustered into two classes, and the points outside the two classes are identified as "illegal cluster core" (N=3), which are not assigned to any of the classes. The clustering results meet two conditions: 1. the distance between any two objects in each class is less than 2R. 2. the number of objects in each class is not less than N.

4 Algorithm Analysis

4.1. Complexity analysis of algorithm

As mentioned above, the algorithm discussed in this paper has a certain randomness. So now the time complexity of the algorithm, in order to facilitate the description, the assumption that the computation space in a plane, all objects furthest distance of two objects as R'. The computation space can be described as $\pi R'^2$, in the first iteration, the range of the cluster core chosen is the whole computing space[9].

Before the first iteration, the space that the cluster core can be chosen is entire circular area: $\pi R'^2$. Assuming that the user specified radius of the class is R, after the completion of the first iteration, the second iteration can choose the range of the cluster center S is $\pi R'^2 - \pi R^2$, this reduces the problem area. The scope that the third iteration can choose have two cases:

1) The area of clusters formed in the previous two iterations is not overlap: $S_2 = \pi R'^2 - 2\pi R^2$ (2)

2) The overlap area of the previous two iterations is s':
$S_2 = \pi R'^2 - 2\pi R^2 + s'$. (3)

Based on the rule of selecting the cluster core, when the cluster core that the second iteration selected is on the edge of the first cluster, s' reached the maximum value: $\max\ s' = 0.135\ \pi R'^2$. Comprehensive (2) (3) :

$$S_2 \le 1.135\ \pi R'^2 - 2\pi R^2 \quad (4)$$

3) As mentioned above, the conclusion can be drawn that each iteration can reduce the computation space by at least $0.32\ \pi R'^2$. In the iterative process of the algorithm, each iteration reduces the computation space. When the computation space is reduced to 0, the whole algorithm is over. Therefore, the number of iterations of the algorithm is up to $\pi R^2 / 0.32\ \pi R'^2$ times[10]. So from the above description, we can see that the complexity of the algorithm is related to the radius of the user's setting, the greater the radius of the user set, the less the number of iterations and the lower the complexity.

4.2. Advantages Compared to the Traditional Clustering Algorithm

As the same as K-means, the algorithm of this paper needs to set the random cluster core, and the clustering results are random[11]. But in K-means clustering, the clustering results are very easy to be disturbed by outliers. It lead to that K-means clustering to degree of standardization of data is very demanding[12]. The algorithm described in this paper is a good method to remove the outliers by the density characteristics of the cluster[13][14].

The same as the DBSCAN algorithm, the proposed algorithm is a kind of clustering algorithm based on density[15]. But on the performance of the proposed algorithm is better, because DBSCAN need to scan all the distance between two objects, which makes it a very large time cost of text clustering in the calculation of similarity. Text clustering algorithm based on random cluster core only calculate the distance from the object to the cluster core without calculating the similarity within the same cluster any two objects.

4.3. Agility of the algorithm

In addition, the clustering algorithm has the characteristics of agility, and the clustering results are not obtained after the whole algorithm is finished. Each cluster is produced one by one, so in the case of urgent need to get the clustering results may wish to get a part of the good class (these results are still reliable) for analysis.

4.4. Algorithm needs to be improved.

This algorithm also has some places to be improved, such as random cluster core selection may enable the clustering effect is not stable, the clustering radius R user can accept is difficult to grasp, all of these are need to be addressed in the following study.

5 Experimental Design And Evaluation

In this paper, the experimental data are taken from the Encyclopedia of Tourism's 35270 tourist attractions introduction. All the attractions of the tourist site and its attractions introduction are described as attractions entities, according to these entities set clustering experiment, to prove the point of view in this paper.

In the course of the experiment, more than 30000 tourism texts are divided into groups and do the clustering, first, use the Chinese Academy of Sciences of the open source algorithm on all the attractions of word segmentation to do the calculation of document similarity.

Clustering method is mainly used in this paper random cluster core algorithm, K-means algorithm and DBSCAN algorithm, and their execution time and the implementation of the results were analyzed and summarized. The execution time of the algorithm is used to evaluate the time efficiency, and the average distance from cluster core to each points is taken as the index to measure the accuracy of clustering after the completion of the cluster. In this paper, we introduce an effective clustering time index to evaluate the agility of the algorithm, and at the same time, we use the cosine angle to measure the distance between the text. Finally get the following results table:

Table 1. Comparison of Algorithms.

	execution time(hour)	average distance(cosine)	effective clustering time (hour)
Random cluster core	3.5	0.55432	1.4
K-MEANS	7.2	0.63221	7
DBSCAN	12	0.735	10

The experimental results show that these three kinds of clustering methods can achieve the classification of tourist attractions into the humanities, landscape, architecture, landscape and other clustering. The random cluster core algorithm is superior to the two classical algorithms in time efficiency, but in the clustering effect, random cluster algorithm has a lot to be improved. That is to say, the random cluster core clustering algorithm is a relatively fast, agile but rough clustering method.

According to the mentioned algorithm in time efficiency advantage. According to the mentioned algorithm in time efficiency advantage. Precise

experiments under different number of test samples, samples were divided into 5000, 10000, 20000 document, measured their time, finally get the following coordinate figure:

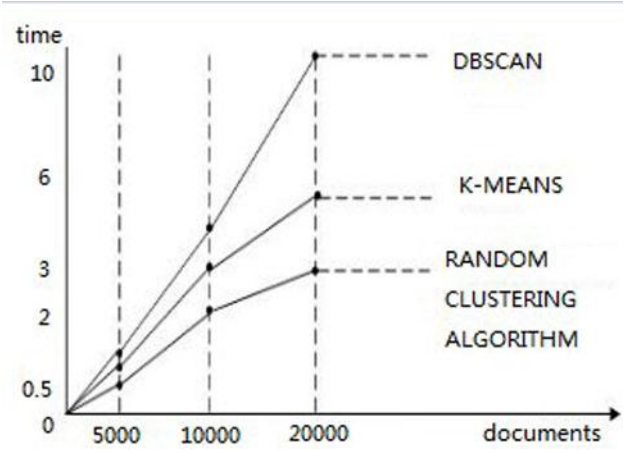

Figure 3. Sample Analysis Coordinate Chart

It can be seen that the time efficiency advantage of random cluster core algorithm is very obvious in the experiment of different number of documents. Of course, as with the K-means algorithm, random clustering algorithm also have the cluster core selection caused by unstable situation, the advantages of this algorithm compared with the traditional algorithm also needs to be defined in the future work.

6 Conclusion

This paper provides an effective solution for clustering tasks. In this paper, we analyze the two mature methods in clustering scheme: DBSCAN and K-means algorithm, and propose the improvement measures on the iterative scheme, and the outliers are discarded. From the experimental results, it is faster and more flexible. In the future research work, it is necessary to make further amendments to the clustering of the algorithm.

References

1. Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases[C] . Seattle: Proceedings of the ACM S IGMOD Conference, 1998. 73-84.
2. JA Hartigan, MA Wong. A k-means clustering algorithm, Applied statistics, 1979.
3. Pang-Ning Tan,Michael Steinbach ,Vipin Kumar. Data Mining:Introduction, 2005.
4. Jiawei Han,Micheline Kamber.Data Mining:Concepts and Techniques, 2006.
5. Gullo F, Domeniconi C, Tagarelli A. Enhancing single-objective projective clustering ensembles. Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010, 833-838.
6. RONG Qiu-sheng, YAN Jun-biao, GUOGuo-qiang. Research and Implementation of Clustering Algorithm Based on DBSCAN. Computer Applications, 2004, 04.
7. YANG Li, ZUO Chun, WANG Yu-Guo. K-Nearest Neighbor Classification Based on Semantic Distance . Journal of Software, 2005, 12.
8. K Hattori, Y Torii. Effective algorithms for the nearest neighbor method in the clustering problem Pattern Recognition, 1993.
9. PAN Li-fang, YANG Bing-ru. Study on KNN arithmetic based on cluster. Computer Engineering and Des, 2009, 18.
10. MENG Yujian, MA Jianghong. An improved K algorithm for the initial cluster core. Statistics & Decision, 2014,12.
11. WANG Yongcheng. Chinese information processing technology and its foundation[M]. Shanghai: Shanghai Jiao Tong University press, 1990.
12. CHU Yue-zhong, XU Bo. Research on optimization of dynamic nearest neighbor clustering algorithm, Computer Engineering and Design, 2011, 05.
13. WANG Qiong. An Improved K-means Optimization Approach for Text Clustering. Computer and Modernization, 2015, 03.
14. Ren Lifang. SPEEDING K-NN CLASSIFICATION METHOD BASED ON CLUSTE R ING. Computer Applications and Software, 2015, 10.
15. LI Shuangqing, MU Shengdi. Improved DBSCAN algorithm and its application. Computer Engineering and Applications, 2014, 50(8) :72-76.