Hot Topic Clustering Based On Words Distances

Hongtao Liu[1,a)], Hongwei Guan[1,b)] , Jie Jian[2)], Xueyan Liu[2)]

[1]College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing400065, China.

[2]School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

[a)])liuht@cqupt.edu.cn, [b)]Corresponding author: 181518324@qq.com

Keywords: Clustering, Words distances

Abstract. In order to find the relevance of the key words in the hot topics effectively, we proposed a clustering method based on words-distances. We calculated the distances between the words firstly, then calculated the sectional density of each words. We regarded the words which have higher sectional density and far away from sectional density point as the center point in the clustering. After find the center point, we start to clustering. This method through decision diagram on estimating the number of clusters. At last, we can find the results on the evaluating indicator of accuracy rate and recall rate.

INTRODUCTION

Term co-occurrence model be used to quantify the correlation between words in a document. So it can be widely used in the field of information retrieval, text clustering and so on[1-3]. We can use this model to abtain the distance between words, which can be used for clustering analysis.

Clustering analysis is an important method in data mining. Its purpose is to assign the data set to multiple clusters, so that the similarity of the data points in the same cluster is as large as possible. The research of clustering algorithm has a long history. In 1975, Hartigan [4] put forward a “clustering algorithm”. On the basis of this, scholars have proposed many kinds of different clustering algorithms, mainly based on partition algorithm, hierarchical algorithm[5-6] and density based algorithm.

RELATED WORK

Alex and Alessandro etc. have proposed a method to search and find the peak density quickly for clustering[7]. The algorithm uses the distance between the data nodes, and can be applied to spherical data sets, furthermore, it can automatically discover the number of clusters.

First, the algorithm is based on the assumption: There are nodes with lower local density around cluster centers and the distance from these cluster centers to node having a higher density are large. For each data node *i*, compute its local density ρ_i and distance δ_i from i to a node with higher local density. The calculation of these two values depends on the distance d_{ij} between data points. Local density ρ_i of the node *i* is defined as follows:

$$\rho_i = \sum_j X（d_{ij} - d_c) \tag{1}$$

if x < 0, then X(x) = 1; otherwise X(x) = 0, dc is a distance cutoff. In general, it is better to select dc so that the average number of neighbor nodes is about 1% − 2% of the total amount of data set. Basically, the density ρ_i of the node *i* is equal to the number of nodes whose distance from i is less than cut-off distance. The algorithm is relatively sensitive to the density of nodes. And for large data sets, the algorithm has good robustness for the value of d_c. The distance δ_i is the minimum among the distances from the node *i* to the nodes with higher local density.

$$\delta_i = \min_{j:p_j>p_i}(d_{ij}) \tag{2}$$

However, for the densest nodes, $\delta_i=max_j(d_{ij})$. Thus, let the unusually large cluster nodes be

cluster centers. Then, after finding the centers, allocating nodes only needs a further step. That is to divide all nodes except class cluster centers. The cluster they are assigned is the same as its adjacent nodes with higher density.

In cluster analysis, to evaluate the reliability of the node assigning is a very important step. In this algorithm, it first finds the boundary area of each cluster. The border area is a set of nodes which although are assigned to a cluster but the distance to other classes nodes is less than dc. Then, for boundary region of each cluster, find the nodes with largest density, and use ρ_b to represent the maximum density. In clusters, if a node whose local density value is larger than ρ_b, it can be a core node, which means that it is more possible for nodes to be assigned to the cluster, while the other nodes are treated as noise ones.

Based on the idea of density peak algorithm and the theory of Term co-occurrence, this paper proposes a new clustering algorithm based on word distance.

THE NEW CLUSTERING METHOD BASED ON WORD DISTANCES

Algorithm

Clustering methods based on fast search and discovery of density peaks can be used for different clustering analysis. First, we must calculate the local density ρ_i of each node i and the shortest distance δ_i from i to nodes having higher local density .Then draw the decision diagram according to ρ_i and δ_i to find out the cluster centers from the decision diagram .After finding cluster centers, assign each node into a cluster center according to the type of its nearest node having high local density. Since the algorithm shows good applicability and high operational efficiency for multiple types of data, the authors wish to introduce it into other types of networks.

But there are two problems to solve. First of all, in the original algorithm need to calculate the distance between nodes, but the specific distance calculation formula is not given and carry out a distance file directly. The second question is based on the connection of the node, divided into two kinds of weighted and unweighted. For nodes that are not connected, if they are considered to have a direct distance from infinity, there will be a lot of infinity in the distance matrix. One is not in line with the actual situation, the other is not conducive to get a good classification effect. For the first question, it is solved by the distance between the words and other words. For the second question, this paper use the passing of distances to calculates the nodes which distances were not given. The algorithm of WBC can be summarized as following four steps:

1. Calculate the distance between hot topics;
2. Calculating ρ_i and δ_i of all nodes;
3. Draw the decision graph and get the cluster center;
4. Clustering according to cluster center.

Calculation of the distance matrix

The distance between the network nodes and the similarity measure is different, and the model will be different. The inter-node distance calculation is based on the term co-occurrence model. The probability of words-occurence is the sum of the probabilities of the simultaneous occurrence of two words within the same subject. It is difficult to determine in the actual modeling that the text contains the subject and the scope covered by the subject. So we need to make a reasonable assumption that a window unit, such as a micro-blog, a comment. There are N window units in the document set, and the prior probabilities of the topics are $P(t)=1/n$. When the word w appears in the window unit, the conditional probability is $P(w|t)=1$, otherwise the conditional probability is expressed as $P(w|t)=0$. Assume that the number of windows in the document w_i and w_j is n. According to the formula, the joint probability of these two words is $P(w_i,w_j)=x/n$. So we can push the word co-occurrence probability formula:

$$P(w_i, w_j) = \frac{S(w_i, w_j)}{n(S)} \tag{4}$$

Where $S(w_i,w_j)$ represents the number of windows in the document space that contain both w_i

and w_j. $n(S)$ represents the total number of window units in the document collection. We can use the value of $S(w_i,w_j)$ to express the relationship between the keywords, the greater the value of the relationship between the words more intimate. The back of the clustering algorithm needs to use the distance between words, so we can draw the distance between the formula:

$$d_{ij} = \frac{1}{S(w_i, w_j)} \tag{5}$$

The smaller the distance, the greater the degree of correlation between the words, the greater the probability of belonging to the same topic community.

Clustering

We assume that m clustering centers are obtained, and each cluster center belongs to only one category. p_i represents the probability that node i belongs to different clusters. The center node of the jth cluster is n, then $p_{nj,j}=1$ and $p_{nj,k}=0(k\neq j)$, thus determining the probability vector of all cluster centers. For other nodes, the probability vector is calculated from the node with the highest local density. Let the node number be i, assuming that the number of nodes with the local density of node i is N_c, and the distance to the node is sorted from small to large, remember as $i_1,i_2,\ldots i_{Nc}$.Then the probability vector of the node is:

$$p_i = \sum_{j=1}^{\min\{N,N_c\}} w_{i_j} p_{i_j} \tag{6}$$

Where w_{ij} represents the weight,

$$w_{i_j} = \frac{1/d_{ij}}{\sum_{k=1}^{\min\{N,N_c\}} 1/d_{ij}} \tag{7}$$

When the probability vector of all nodes is obtained, for node i, if the node satisfies $r=\text{argmax}\{p_{i,s},s=1,2,\ldots,c\}$, then node i belongs to class r.

EXPERIMENT

Experimental data

Our experimental data were 5 micro-blog hot event. Dataset as shown in Table 1:

TABLE 1. Hot event information

Dataset ID	Dataset name	Records-number
R1	F-10 female pilots sacrifice	508
R2	Lin Yi quit addiction center	185
R3	US election	480
R4	South Korean scandal	620
R5	World Internet Conference	381
R6	Noise data	1200

Evaluation

Then we put the data into the improved algorithm to calculate, get the following experimental results:

We select the p and δ two values are larger points as a clustering center clustering. Then we use the algorithm to clustering. The deep blue dots represent R1 data. The deep red dots represent the

R2 data. The light blue dots represent the R3 data. The yellow dots represent the R4 data. The orange dots represent R5 data. The black dots represent noise data.

For hot even data, since its own node is tagged. We chose the accuracy rate and recall rate as evaluation index. The experimental results are shown in the following table 2:

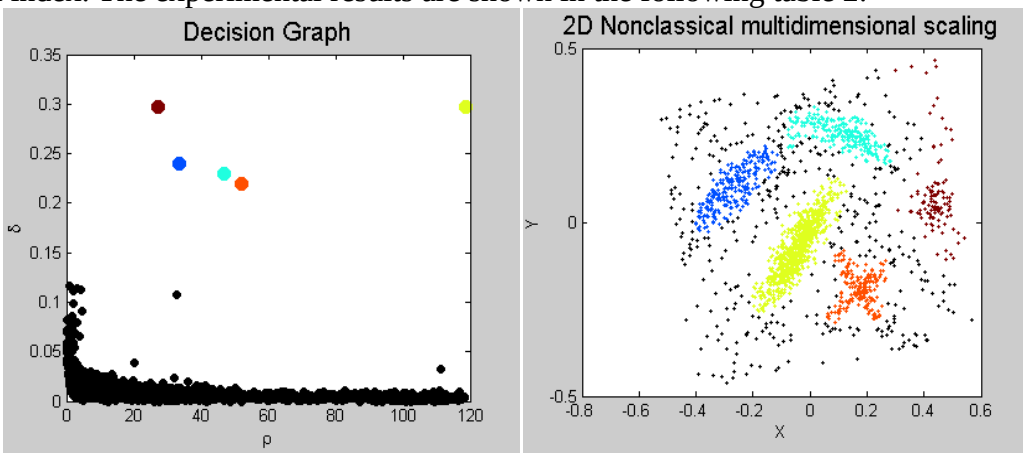

FIGURE 1. Decision Graph and Clustering Result

TABLE 2. The results of comparing clustering algorithm by Accuracy rate and Recall rate

Dataset	(Recall rate) Improved WBC	(Recall rate) Single-Pass	(Accuracy rate) Improved WBC	(Accuracy rate) Single-Pass
R1	0.7087	**0.7598**	0.7273	**0.7732**
R2	0.7027	**0.7297**	0.7471	**0.7337**
R3	0.6667	**0.7916**	0.6823	**0.7585**
R4	0.7580	**0.8065**	0.7655	**0.8076**
R5	0.6824	**0.7847**	0.7046	**0.7706**

CONCLUSION

It can be seen from the experimental results that the WBC algorithm used in this paper is superior to Single-Pass algorithm in accuracy and recall rate. When the data is large, the distance between all words must be calculated. The algorithm proposed in this paper takes longer time.

ACKNOWLEDGMENT

This research is supported by the following fundings or programs: the National Natural Science Foundation of China (61402309), the Fundamental Research Funds for the Central Universities (No. XDJK2014B012), the National Social Science Foundation of China (13CGL146), the National Social Science Foundation of China (15BGL2729), the Study on the Key Common Characteristics of Network Transaction Fraud (14SKF01).

REFERENCES

[1] Qiao Y N, Yong Q, Hui H. The Research on Term Field Based Term Co-Occurrence Model[C]// Semantics, Knowledge and Grid, Third International Conference on. IEEE, 2007:471-474.
[2] Zhang Y, Shi K, Qingpeng X U, et al. Spam Filter Based on Term Co-Occurrence Model[J]. Journal of Chinese Information Processing, 2009.
[3] Gao, Jianfeng, Zhou, et al. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations[J]. 2002.
[4] Hartigan J A. Clustering Algorithms [M]. New York: John Wiley & Sons Inc.1975.

[5] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.
[6] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.
[7] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.