ISSN: 2414-1895 DOI: 10.6919/ICJE.202212_8(12).0014

An Improved Information Clustering Algorithm for Network Marketing System

Hankun Ye

School of International Trade and Economics, Jiangxi University of Finance and Economics, Nanchang, 330013, China

Abstract

Similar information clustering is one of the difficult and hot research fields in the internet search engine research in network marketing system. Using the clustering and analyzing techniques of data mining, the paper takes text clustering for example and presents a new information clustering algorithm based on density-isoline. Firstly, texts are preprocessed to satisfy succeed process. Then, the paper introduces density-isoline clustering algorithm and improves the specific definition of density function, algorithm flow and selection of neighborhood size and density threshold value when used in text clustering. The experimental results indicate that the improved algorithm has a higher accuracy compared with the original algorithm, and has a better stability.

Keywords

Search Engine Research; Information Clustering; Density-isoline Algorithm; Network Marketing System.

1. Introduction

With the popularization and application of electronic business and network marketing system, network has become an important part of the people's working and living, and various search engines have been an indispensable tool to retrieve the necessary resources for the people. However, the Internet search engine can often find thousands of search results. Even if some useful information is obtained, it is often mixed with a lot of "noises" to waste the users' time and money. Therefore, in order to efficiently and economically retrieve the resource subset relevant to the given search request and with the appropriate number, the similar information clustering ,hear taking text clustering for example, is performed and becomes one of important and hot research fields in data mining[1].

Text clustering is different from Text classification. The latter has them for each category while Text clustering has no category annotates in advance. The Text clustering is to divide the Text sets into several clusters according to the Text contents, and requires the similarity of the Text contents in the clusters as great as possible and that of different clusters as small as possible. It can organize the Web Text effectively, but also form a classification template to guide the classification of the Web Text. Therefore, the Text clustering can do the online information clustering based on the contents to facilitate the retrieval and reading[2].

Data mining was originally a term in the statistics and meant a process to explore the data rules and characteristics without prior hypothesis verification. In recent years, the data mining technology is considered to have the exciting research background. If this technology obtains the rapid development and perfection, it will be able to be applied widely. While the data mining technology begins to be applied in many respects abroad, it is now in the phase of theoretical exploration and applicable experiment generally at home, and only has the preliminary application in the computer network and the management decision-making. In addition, the reports of the application in the medical data field clustering are rarely seen[2].

Mathematically, the so-called clustering is that massive d-dimension data samples (n pieces) gather into k classes ($k \prec\prec n$) to maximize the similarity of the samples in the same classes and minimize it in different classes. The clustering process is to classify the data objects with many attributes constantly, the clustering algorithm carries out the classification automatically, and the data is cut into several classes through the recognition of data characteristic. Therefore, it is considered that the clustering rule can be used completely to mine the algorithm, find the clustering basis of each target, and then carry on the recognition and clustering according to this basis. The next key question is to seek a clustering mining algorithm to derive this clustering basis.

Although there are many text clustering algorithms advanced at present, many of them have carried on the certain hypotheses[3] to the sample data analysis like spherical distribution, linear distribution, plane distribution, etc.. The flaws of the traditional algorithms result in great difficulty increase of the data mining technology in the clustering of large-scale network text, of which, the practicability is also reduced greatly.

The presented clustering algorithm based on density iso_line starts from the concept of the contour line, produces the density iso_line graph on the basis of the sample distribution density and finds the slightly centralized region of sample distribution from the density iso_line graph again so as to obtain a better clustering results.

2. Text Preprocessing

Text clustering can be described as: a given Text set $D = \{d_1, d_2, ..., d_n\}$ eventually gets a cluster's set $C = \{C_1, C_2, ... C_n\}$, $\bigcup_{i=1}^{k} C_i = D$ derives $\forall d_i (d_i \in D)$, $\exists C_j (C_j \in C)$ and $d_i \in C_j$) , and also makes the objective function $Q(C)$ reach the minimum or maximum value, of which, n is total Text number, k is final clustering number, and $C_j \cap C_i \neq \phi, j \neq i$.

2.1 Characteristic Selection and Expression of Text

Vector space model (VSM) is commonly adopted to express each Text. In this model, each Text d is considered as a vector in a vector space. $tfidf$ is used as a measure of characteristic vector in this paper, and this measure gives the weight of each word t . See Formula 1 for the calculation of the weight.

$$tfidf(d,t) = tf(dt) * \log_2 \frac{N}{df(t)} \tag{1}$$

In formula 1, $tf(d,t)$ is the word frequency of word t in the Text d , $df(t)$ is all the Text numbers of word t contained in the Text set D , and N is total Text number. After the characteristic selection, Text $d \in D$ is the form of the vector, and the value of each dimension is the corresponding $tfidf(d,t)$ weight value, so the Text can be expressed as Formula 2.

$$d = \{(t_i, tfidf(d,t_i)) \big| 1 \leq i \leq m\} \tag{2}$$

Of which, t_i is the lexical entry, and m is the dimension of the characteristic vector. However, after the characteristic selection, m is still very large, thousands of dimensions at least and tens of thousands of dimensions at most while non-zero word frequency of each corresponding Text vector is very few, which makes Text VSM show the high-dimension and sparsity of the model.

2.2. Definition of similarity

In this paper, cosine distance is used to measure the similarity between the texts and defines the similarity of two texts d_1 and d_2 in Formula 3.

$$Sim(d_1,d_2) = \cos(d_1,d_2) = d_1 \cdot d_2 \|d_1\| \cdot \|d_2\| \quad (3)$$

In order to reduce the impact of different length of the Texts on calculating the Text similarity, each Text vector has been integrated to the unit length. See Formula 4.

$$d = d / \|d\| = \frac{\{tfidf(d,t_1), tfidf(d,t_2), ...tfidf(d,t_m)\}}{\sqrt{\{tfidf(d,t_1)^2, tfidf(d,t_2)^2, ...tfidf(d,t_m)^2\}}} \quad (4)$$

Thus, $\|d\| = 1$, and the similarity of the cosine is the dot product of two Text vectors, that is, $Sim(d_1,d_2) = d_1 \cdot d_2$.

3. Isopycnic Clustering Algorithm

From a new angle, the clustering is to find the quite intensive parts in the samples, and each intensive part is a class. Starting from this angle, it can design a density function and calculate the density nearby each sample so as to find the quite centralized regions in those samples according to the density value nearby each sample. These regions are the classes that we have to find. According to the definition of the clustering, the effect should be the best for the clustering herein. The isopycnic clustering algorithm starts from the clustering algorithm based on the density, combines the concept of the contour line map to present a new clustering algorithm. In the contour line map, it not only can determine which the mountains are according to the contour line, but also can find the mountain peaks higher than the certain height as required. In a similar way, the sample clustering can be the fact that the density of the sample distribution is firstly calculated to draw the contour line map (i.e. density iso_line graph (referred as iso-line graph)) for the sample distribution density, and then choose the appropriate density iso_line from the graph, so the parts that these density iso_lines surrounded are the quite intensive parts in the sample and the parts to be segmented. Moreover, it can find the classes with the different density degree correspondingly according to the different requests[5, 6,7].

According to the concept of the contour line map, the distances among the samples and the neighborhood radius size RT, the isopycnic clustering algorithm calculates the density value of each sample. And then according to the given density threshold value DT, it finds the samples that all the densities are larger than DT as well as the sample sets that the distances in these samples are smaller than the threshold distance RT, and combines the overlapping sample sets to obtain a group of clustering of original samples. However, the samples that did not belong to any classes and had extremely few classes are called the noise because their distributions are extremely sparse. In the isopycnic algorithm, all sample data are required to be pretreated, the value range of each attribute in the sample is [0,1] so as to calculate the distance matrix and simultaneously use the reasonable formula to count the size of the threshold distance.

3.1 Improvement of Definition of Density Function

This algorithm mainly calculates the density function, and further derives the density iso_line from the density of each sample. The isopycnic algorithm uses the concept of the neighborhood distribution density. While a sample neighborhood refers the region that this sample distance is smaller than a fixed value, the sample number contained in this neighborhood is its neighborhood distribution density[4,5,6].

Calculate the neighborhood sample distribution density as Formula 4.

$$Den(A) = size(\{B \mid f(B,A) \le T\}) \quad (5)$$

Among them, A and B are the input samples; Den (A) is the sample distribution density nearby sample A, and size (X) is the sample number in the set X, namely the size of set X. f (B, A) is a measure function of similarity between Sample A and Sample B, and T is a given threshold value. The measure function f is showed by the euclidean distance. Take the sample number that the distance in a sample is smaller than the fixed value T as the density of this sample showed by Formula 6.

$$Den(A) = size(\{B \mid Dist(B, A) < T\}) \tag{6}$$

In Formula 5, there are many methods to get the distance function. Generally use Minkovski distance as Formula 7, among which, λ is a positive integer. The most common distance measure takes $\lambda = 2$, namely euclidean distance.

$$Dist_{\lambda}(X, Y) = \left[\sum_{i=1}^{d} |x_i - y_i|^{\lambda}\right]^{1/\lambda} \tag{7}$$

3.2 Improvement of Algorithm Flow

Calculate the distances between two arbitrary samples to get the distance matrix $Dist$ as formula 8.

$$Dist(i, j) = D(X(i), X(j)) \tag{8}$$

According to the distance matrix, decide the neighborhood size as formula 9.

$$RT = mean(Dist)/(n\hat{}\,coefRT) \tag{9}$$

Calculate the density matrix Den: Namely give the number of the sample point within radius RT range. Each line in the matrix has to find the number that its distance is smaller than the threshold distance RT, and this number is the neighborhood sample distribution density that this line corresponds to the sample. According to the distribution density obtained, the sample iso-line graph can be drew (this iso-line graph does not need to be drew in actual clustering process, and is concealed in the density matrix.).

According to the density matrix, decide the density threshold value as formula 10, Sign “[]” in formula 9 indicates the rounding operation.

$$DT = mean(Den)/(coefDT) \tag{10}$$

The merge directs towards the sample A that each density is larger than DT. If the distances of sample B and A are smaller than RT, and B’s density is also larger than DT, A and B’s classes are merged into a class, and the merge of such classes can obtain the clustering result finally.

If the clustering result is not too ideal, the density threshold value DT can be adjusted through coefDT so as to carry on the optimization of the clustering result. Among them, X is the sample set, n is the sample number, and coefDT and coefRT are the adjustable factors.

3.3 Improvement of Neighborhood Size and Density Threshold Value Selection

It is simple to obtain each class from the iso-line graph, so the algorithm of the iso-line graph is mainly how to obtain the best iso-line graph, in fact, how to confirm the neighborhood size RT. If the neighborhood is excessively small, the neighborhood distribution density of each sample will be very small, the clustering result will have many classes, and each class will only contain the very few samples. The extreme situation is that the neighborhood is smaller than the minimum value of the distance among all the samples, and then each sample density is 1 and belongs to one class respectively. The number of the classes is equal to the sample number, so such clustering result is meaningless. On the contrary, if the neighborhood is oversize, each sample neighborhood distribution density is very large, and the density value is quite close, from which, the density iso_line drew is very difficult to reflect the true distribution condition of the sample and cannot distinguish two classes with closer distance clearly. The class with closer distance is fallen into a class generally in the clustering result. The extreme situation is the neighborhood is larger than the maximum value of the distances among all the samples, and then each sample density is n (n means the sample number) and all the samples are combined into one class. However, such clustering result is still meaningless.

It can be seen from the above analysis that the neighborhood size should be between minimum and maximum values of the distances in all the samples, namely $\min(Dist) \le RT \le \max(Dist)$. In general, except that the value adopting of the neighborhood satisfies the above conditions, the neighborhood density distribution derived is even as far as possible and the distribution scope is broad as far as possible, so the obtained density iso_line can reflect the sample distribution of each density level to find each class that hides among it. In the isopycnic clustering algorithm, the methods of the neighborhood size and the density threshold value can be confirmed according to the sample number and the intensive degree of sample distribution. The value adoption of the neighborhood size is calculated by Formula 11.

$$RT=\frac{mean(dist}{n^{coefRT}} \tag{11}$$

Among them, mean(Dist) means the mean value of the distance among all the samples. n is the sample number, coefRT is the adjustment factor of the neighborhood radius, and the value adoption is from 0 to 1. Several experiments indicate that good clustering effect can be obtained under many situations when coefRT value adoption is 0.3.

The size of the density threshold value DT will decide the final result of the clustering. If the density threshold value is small excessively, it will be able to cause the combination of the classes in nearer distance; if the density threshold value is oversized, it will be able to divide one class into several classes or large parts of the samples into the noise so as not to be able to use the information carried in the sample. In the isopycnic algorithm, the value adoption of the density threshold value is calculated by Formula 12.

$$DT=\begin{cases} 2 & n<1000 \\ \dfrac{mean(Den)\times coefDT}{\log_{10}(n)} & n\geq 1000 \end{cases} \tag{12}$$

Among them, coefDT is the adjustment factor of the density, and the value adoption is from 0.7 to 1. Several experiments indicate that good clustering effect can be obtained under many situations when DT is 0.95.

Regarding the given sample set, the neighborhood size and the density threshold value are derived automatically from the algorithm according to the above formula. In addition, the size of the density threshold value DT can also be adjusted according to the clustering result so as to adjust the clustering result. In the event of knowing the number of the class in the sample set beforehand or determining to obtain several classes in advance, it is possible to adjust the adjustable parameter in the formula of the neighborhood size and the density threshold value to obtain the better clustering effect.

3.4 Algorithm Operation Analysis

In the algorithm, the reading-writing operation of each document takes one line in the distance matrix as a unit to reduce I/O operation times and improve the algorithm speed greatly. The time consumption of this algorithm is mainly used to calculate the distance matrix, and the time complexity is $O(n^2)$. Because this algorithm process needs to use the distance matrix between the samples, its size is n^2 which is not small number. In order to save the storage space, the integer carries on for the distance matrix, namely each distance is expressed by the integers of 2-byte length. In addition, the distance matrix can be also stored into the hard disk to save the memory space. The space of hard disk needed in all is $2\times n^2$ bytes.

Under the condition that d-dimension data is larger than 2, except the time to calculate the distance matrix Dist among the sample is added along with d increase, the calculations of other parts in the isopycnic clustering algorithm do not need any change, and time complexity and space complexity are not affected by d. In other words, other parts beyond the distance calculation in the isopycnic algorithm have nothing with d-dimension data and only are relevant to sample number n[6,7].

4. Experimental Verification

To test the effectiveness of the improved algorithm, the original algorithms[5] and the improved algorithms are compared. The experiment is made on the computer of the Celeron (R) 2.0G, 512M memory by VC++ and the experimental data is from www.China.com.cn and www.sohu.com. See Tab. 1 for the experimental results, of which, M_i is total Text number of category i; N_j is total Text number of cluster j; $M(n_{ij})$ is total Text number of category i included in cluster j when category j reaches the maximum F-measure value; $M(F(i,j))$ is the maximum value in category i and F-measure value of different clusters.

Table 1. A set of comparison clustering results

Cluster	Original K_means Algorithm				Improved Algorithm		
	M_i	N_j	$M(n_{ij})$	$M(F(i,j))$	N_j	$M(n_{ij})$	$M(F(i,j))$
Arts	44	40	20	0.48	39	28	0.68
Politics	83	80	53	0.65	81	61	0.74
Health	41	39	31	0.71	39	34	0.85
Sports	65	63	38	0.59	64	57	0.88
News	89	74	59	0.73	81	77	0.91
Culture	104	95	63	0.64	94	88	0.89
Education	112	98	74	0.71	102	96	0.90
Military	132	104	89	0.77	123	119	0.93
Science	144	123	98	0.74	137	132	0.94
Average	$F = 0.67$				$F = 0.86$		

In Tab. 1, the data shows that adopting the improved clustering algorithm improves the accuracy of the clustering results. In order to verify the stability of the improved clustering algorithm result, multi-group data is used to perform the comparative experiment respectively by two algorithms to obtain

30 groups of experimental data. The F-measure value distribution in the experimental results is shown in Tab. 2.

Table 2. Comparison experimental clustering results

F-measure interval	F-measure typical value	Original K-means algorithm F-measure value falls into the experimental frequency of this interval	Improved K-means algorithm F-measure value falls into the experimental frequency of this interval
[0.15,0.25]	0.20	0	0
[0.25,0.35]	0.30	1	0
[0.35,0.45]	0.40	2	0
[0.45,0.55]	0.50	4	0
[0.55,0.65]	0.60	6	0
[0.65,0.75]	0.70	9	8
[0.75,0.85]	0.80	1	12
[0.85,0.95]	0.90	0	8
[0.95,1.00]	1.0	0	0

It can be seen from Tab. 2 that there is poor stability in the clustering results obtained by the ordinary algorithm and scattered F-measure value; but the improved clustering algorithm has better stability of the clustering results, more concentrated F-measure value and higher F-measure average value. The experiment shows that the improved clustering algorithm improves its accuracy and stability greatly. In the use of ordinary clustering algorithm, F-value of the clustering results scatters from 0.60 to 0.75; in the use of the improved algorithm, the stability of its value is from 0.75 to 0.85.

5. Conclusion

The density-isoline clustering algorithm starts from the isoline graph of the sample distribution to turn the regions an density-isoline segmented into one class respectively to find quite intensive parts in the sample distribution, which do satisfy the clustering requirement. In a qualified sense, the clustering result obtained by the density-isoline algorithm should be best. However, to get the best clustering result has to get a better density-isoline graph firstly, select the appropriate density threshold value herewith secondly and combine the points the selected density-isoline embraced finally to obtain each clustering class. It can be seen from the experimental result that: firstly, the density-isoline clustering algorithm not only can discover the intensive regions of various samples, but also can effectively remove the noise disturbance to obtain the better clustering result; next, in the realization efficiency of the algorithm, the processing time is reduced greatly and the practical application value of the density-isoline clustering algorithm is enhanced in the clustering of the 3D medical data field greatly because the data field pretreatment is presented and the algorithm is revised and improved according to actual characteristic of medical data field.

References

[1] Xu Sen, Lu Zhi mao, Gu Guo□chang . Two spectral algorithms for ensembling document clusters .Acta Automatica Sinica, 2019, 35(7) : 997-1002.

[2] Tao Li. Document clustering via Adaptive Subspace Iteration. Proceedings of the 12th ACM International Conference on Multimedia, New York: ACM Publisher, 2022. 364 -367.

[3] Halkidi M, Vazirgiannis M.NPClu: An approach for clustering spatially extended objects[J].Intelligent Data Analysis, 2021, 12(6):587-606.

[4] Zhang Y C, Xie F, " A Clustering Algorithm Based on Density-isoline", Journal of Beijing University of Posts and Telecommunications,2019,25(2):8-13.

[5] Ester M,riegel H P, " Density-based Algorithm for Discovering Cluster in Large Spatial Database with noise", proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Dregon 2019:124-128.

[6] Zhao Yanchang, Song Junde. AGRID: An efficient algorithm for clustering large high-dimensio naldatasets . Proc the 7th Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD-03). Seoul, Korea : 2020.

[7] Halkidi M, Vazirgiannis M. NPClu:An approach for clustering spatially extended objects. Intelligent Data Analysis,2020,12(6):587-606.