Automatic retrieval of network hotspot information based on improved DBSCAN clustering algorithm

Dingsheng Deng1 *

1 Sichuan University Nationalities, Kangding, Sichuan, 626001, China
*Corresponding author’s e-mail: dds0904@scun.edu.cn

Abstract. With the continuous development of the Internet in the world today, the proliferation of network information makes it difficult for people to effectively screen out current hotspot information. In order to solve the problem of how to retrieve the current hot information quickly and accurately under the massive network information, an automatic network information retrieval method based on the improved DBSCAN clustering algorithm is proposed. The retrieved related keywords are combined to reduce the feature terms, which effectively solves the problem of repeated acquisition of public resource object neighbourhoods, greatly improves the accuracy and efficiency of the clustering algorithm, and realizes automatic retrieval of network information hot spots. The results show that the automatic network information retrieval method based on the improved DBSCAN clustering algorithm proposed in this paper can quickly and accurately find the current information hotspots on the Internet, help users to obtain the hotspot information of their own most interest, and promote the progress and development of the Internet.

1. Introduction
Since the 1990s, with the continuous growth of China’s economy and the continuous progress of society, Internet technology has also achieved rapid development, and large-scale Internet companies have continued to grow like bamboo shoots, playing an important role in promoting China’s economic development and social progress. However, when users face rapidly growing network information, it is very difficult to quickly obtain current network hotspot information. Obviously, the way to obtain network hot spots through manual retrieval is very inefficient for today's massive network information. Therefore, it is urgent to divide the massive network information into several meaningful clusters through efficient text clustering technology to realize the automatic discovery of network hotspots, so that users can accurately understand the current network hotspot information without searching and analyzing by themselves [1-2].

Aiming at the characteristics of network information, this paper proposes an improved method of network hotspot information retrieval based on improved DBSCAN. First, the vector space model is used to structure the network information text. Then, the improved DBSCAN clustering method is used to automatically retrieve network hotspot information. The combination of related keywords solves the problem of repeated acquisition of public resource object neighborhoods, and effectively realizes fast and accurate automatic retrieval of network hotspots.

2. Improve the DBSCAN algorithm
2.1 Text representation model
The vector space model (VSM) proposed by SALTON and others in the late 1960s effectively realized the conversion between natural language and mathematical models. The basic principle is: ignore the correlation between words and use vectors to represent text. The model only considers corpus statistics and does not consider semantic effects, which simplifies the relationship between keywords in the text, so that the distance between vectors can be used to determine whether the text is similar to the text. The vector space model can be applied in the fields of information retrieval and text classification. The basic process of establishing a vector space model is: the sequence of feature terms in the text and the relationship between them are not considered, so the text is represented as several independent feature term sets. Use this set to construct a high-dimensional space whose dimensions are each feature item, and the text can be represented by vectors in this space. The importance of the feature item to the text is the feature item weight, and the weight is the value of the text feature vector, and then whether the texts are similar is calculated by calculating the cosine similarity between the text vectors. The vector space model transforms text data into computable structured data, thereby solving the similarity problem between texts by calculating the similarity between vectors [4-6].

2.2DBSCAN clustering algorithm
DBSCAN (density based spatial clustering of applications with noise) is a classic density-based clustering algorithm. When the number of objects in a region with a given radius $R$ exceeds the density threshold $MinPts$, they are merged into similar classes to achieve clustering of objects with similar density. In the DBSCAN algorithm, the distance function between objects $x$ and $y$ in the data set $D$ is defined as $Dis(x, y)$. The purpose of the DBSCAN algorithm is to find the largest set of densely connected objects. The steps of the DBSCAN algorithm are: first scan the input object set, determine whether each input object is a core object, and find the direct density reachable set in the $R$ neighborhood $NR(p)$ of the core object; all direct density-reachable objects of the core object are merged to find the largest set of density-connected objects to achieve density-based object clustering.

2.3DBSCAN algorithm improvement based on time efficiency
It can be known from the DBSCAN algorithm that the clustering process needs to traverse the $R$ neighborhood of each object in the object set to ensure that each object is marked as a certain class. In practice, the $R$ neighborhoods of many objects cross each other, that is, an object exists in the $R$ neighborhoods of several objects at the same time, as shown in Figure 1.

![Figure 1. Core objects for neighborhood crossing](image.png)

In Figure 1, the objects $x$, $o$, and $y$ are all core objects, and the $R$ neighborhoods of these three objects must cross each other, and there are some common objects. This paper improves the DBSCAN clustering algorithm based on the problem of repeated acquisition of neighborhoods of public objects. The specific idea is: before obtaining the $R$ neighborhood of each object in the object set, first determine whether this object has been marked as an object in the $R$ neighborhood of an object, and if so, obtain the $R$ neighborhood of the object, otherwise judge the next object to avoid repeatedly acquiring its $R$ neighborhood for public objects, thereby reducing the number and time of acquisitions.
3. Network hotspot information retrieval based on improved DBSCAN algorithm

According to the improved DBSCAN algorithm, we need to represent the object text in section 23 as a vector space model, and then perform density clustering based on text similarity to obtain network hotspot information. The author clusters the information according to the domain category, which reduces the semantic interference to a certain extent, at the same time greatly reduces the size of the cluster, and improves the time efficiency of the cluster. The resource clustering process is shown below.

1) Resource text classification by field: After classifying the target resource text set by field, the corresponding resource clustering is performed in a certain discipline. This "classification before clustering" improves the accuracy and time efficiency of clustering to some extent.

2) Text feature extraction: Each resource has corresponding information such as title, keywords, and abstract. These information are carefully extracted by the author, which can reflect the main idea of the corresponding information. The keywords of each information are the most refined and important, so the keywords are selected as the feature items of the information.

3) Feature set reduction: the number of feature items selected according to the keywords will increase with the increase of the number of information, so it is necessary to reduce the feature set obtained, and take the first \( n \) feature sets with more frequent occurrence of feature items in all information texts. Then \( n \) value needs to be selected properly according to the number of target resources, because when \( n \) is too large, redundancy will be introduced to make the clustering accuracy insufficient and when \( n \) is too small, the feature information contained is not enough to distinguish different categories. At the same time, in addition to deleting meaningless keywords with low word frequency, it is necessary to merge keywords with obvious inclusion relationship, so as to improve the accuracy of clustering results to a certain extent.

4) Build feature space: build feature space according to the reduced feature set \( \{T_1, T_2, ..., T_n\} \).

5) Construct characteristic vector space model: calculate the importance (i.e. weight \( W_k(1 \leq k \leq n) \)) of each feature term \( T_k(1 \leq k \leq n) \) for each text in the feature term space, so as to establish the vector space model. The target information text is represented as \( D(T_1, T_2, ..., T_n) \) by vector space model. Where \( D \) is the target information text, \( T_k(1 \leq k \leq n) \) is the independent characteristic term of \( D \). Then, according to the TF-IDF method, the weight \( W_k \) of the characteristic term \( T_k \) is determined.

TF-IDF is a widely used weight calculation method that comprehensively considers the word frequency of feature terms in the author's text and all texts. That is, the more often a feature item appears in the author's text, the greater its importance, that is, proportional to the weight result; and the more the text contains a feature item, the lower its ability to distinguish text, that is, The less important a certain text is, it is inversely proportional to the weight result. The specific calculation formula of TF-IDF is:

\[
TF - IDF(T_k, D) = \frac{TF(T_k, D) \times lb\left(\frac{N}{DF(T_k)} + 0.01\right)}{\sqrt{\sum_{i=1}^{n}[TF(T_i, D) \times lb\left(\frac{N}{DF(T_i)} + 0.01\right)]}}
\]

In the formula: \( TF(T_i, D) \) is the word frequency of \( T_i \) in \( D \), that is, \( TF(T_k, D) = n_k / \sum_{i} n_i \) , \( n_k \) is the number of occurrences of \( T_k \) in \( D \), and \( \sum_{i} n_i \) is sum of occurrences of all feature terms in \( D \),

\( IDF(T_k, D) = lb\left(\frac{N}{DF(T_k)}\right) \) is the ability of feature word \( T_k \) to distinguish text, that is, the logarithmic of ratio of the total number of texts \( N \) to the total number of texts containing feature word \( T_k \), the purpose of adding 0.01 before taking logarithm is to prevent that when \( DF(T_k) = N \), \( lb(1) = 0 \) , and the sum under the root sign is 0, so as to avoid the denominator of the expression being 0.
In this paper, $W_k$ is used to represent the weight of $T_k$ ($1 \leq k \leq n$), then the vector of text $D$ is expressed as:

$$D = D(W_1, W_2, \cdots, W_n)$$

(2)

(6) Text similarity calculation: text similarity between text $D_1$ and text $D_2$ $\text{Sim}(D_1, D_2)$ is measured by the angle cosine between vectors:

$$\text{Sim}(D_1, D_2) = \cos \theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} W_{1k}^2\right)\left(\sum_{k=1}^{n} W_{2k}^2\right)}}$$

(3)

(7) Text clustering: the improved density based db-scan clustering method is used to achieve text clustering, and the distance between objects is measured by the similarity between text vectors.

(8) Description of clustering results: two eigenvalues with the largest word frequency in the resource texts grouped into one group are combined into clustering results, i.e. the name of the hot spot found.

4. experimental verification

Taking the Sina MicroBlog hot search information of a week as a test sample, the effectiveness of clustering algorithm is verified by comparing the results of artificial classification and experimental clustering. There are 492 test samples in total, which are divided into five hot spots artificially: star, movie, song, entertainment and party. The common indexes to evaluate the clustering effect are precision rate and recall rate. The former indicates the ratio of the objects in the clustering result that really belong to a certain category to the total number of objects in the clustering result of this category, and the latter indicates the proportion of all objects that really belong to a certain category, and the clustering results in the same judgment.

In DBSCAN algorithm, $\text{Minpts}=4$, and the value of $R$ is 0.8 after normalization, that is to say, when the similarity between texts is greater than or equal to 0.8, it can be classified into one category. The dimension of the feature space is set to 15. First, cluster the resource texts after the reduction of general feature words (the feature words with inclusion relations are not merged), and the specific results are shown in Table 1.

| Hotspot category | Number of hot searches | Precision rate (%) | Recall rate (%) |
|------------------|------------------------|--------------------|-----------------|
| Star             | 188                    | 100                | 94.87           |
| Movie            | 65                     | 100                | 99.02           |
| Song             | 47                     | 100                | 100             |
| Entertainment    | 112                    | 100                | 89.21           |
| Party            | 80                     | 100                | 98.35           |

It can be seen from Table 1 that the recall rate of some categories is relatively low. By analyzing the results of artificial classification and clustering, it can be seen that there is an inclusion relationship between some characteristic words, which can be classified into one category in fact, while the clustering algorithm treats them as independent characteristic words. Taking "neural network and genetic algorithm" as an example, some related keywords in some texts are "BP neural network", "human" The texts with these keywords, such as artificial neural network, hybrid genetic algorithm and improved genetic algorithm, should also be classified as "neural network and genetic algorithm".
Table 2. Results of hot spot information retrieval after merging words with relational features

| Hotspot category | Number of hot searches | Precision rate (%) | Recall rate (%) |
|------------------|------------------------|--------------------|-----------------|
| Star             | 188                    | 100                | 100             |
| Movie            | 65                     | 100                | 100             |
| Song             | 47                     | 100                | 100             |
| Entertainment    | 112                    | 96.54              | 100             |
| Party            | 80                     | 100                | 99.58           |

Therefore, this paper improves the feature reduction part of the method, that is, in addition to deleting meaningless keywords with low word frequency, it also merges keywords with obvious inclusion relationship, so as to improve the accuracy of clustering. As shown in Table 2, the results of hot spot information retrieval after merging the words with relation characteristics are combined.

It can be seen from Table 2 that the recall rate of hot spot categories has improved to some extent; it is also found that the precision rate of "entertainment" hot spots has decreased, because other effective keywords are inevitably introduced when merging keywords with inclusion relations, but the recall rate is greatly improved, so the method is still effective in general.

This paper validates the DBSCAN algorithm based on time efficiency improvement, that is, to solve the problem of repeated neighborhood acquisition of common objects in DBSCAN clustering algorithm. The clustering results of the algorithm before and after the improvement are the same, and the time comparison results of the algorithm before and after the improvement are shown in Figure 2. It can be seen from Figure 2 that, on the premise of ensuring the accuracy of clustering, the improved clustering time is shortened by avoiding the repeated acquisition of common object neighborhood, and becomes more and more obvious with the increase of the number of information texts. Therefore, the improvement of algorithm based on time efficiency is of great significance to the clustering of massive information texts.

Figure 2. Comparison of clustering time before and after DBSCAN algorithm improvement

Through the above clustering effect evaluation and clustering time results, we can see that the improved DBSCAN algorithm can quickly and effectively automatically discover network hot spot information, and through clustering based on entries, feature reduction based on merging with inclusion relation and solving the problem of neighborhood repeated acquisition of public objects in DBSCAN clustering algorithm, it can ensure clustering to a certain extent Class accuracy and time efficiency.
5. Conclusion
In this paper, a clustering method based on information keywords is proposed for Internet massive information resources, which realizes the automatic retrieval of network hot spot information, so that Internet users can quickly understand the corresponding network hot spots without self search and analysis, and improves the use value and friendliness of the Internet. However, the parameters in the DBSCAN algorithm are selected through many experiments, how to quickly determine the appropriate and effective parameters will be the next research problem.

Acknowledgments
This paper was financially supported by the Key Project of Natural Science of Sichuan Provincial Education Department (No. 17ZA0295), 2017 Applied Demonstration Course Project of Sichuan University for Nationalities (No. sfkc201705) and Key Project of Natural Science of Sichuan University for Nationalities (No. XYZB19001ZA).

References
[1] Wu, X., Huang, Z., Peng, X., et al. (2018) Building a Spatially-Embedded Network of Tourism Hotspots From Geotagged Social Media Data. IEEE Access., 6: 21945-21955.
[2] Soteriou, V., Theocharides, T., Kakoulli, E. (2016) A Holistic Approach Towards Intelligent Hotspot Prevention in Network-on-Chip-Based Multicores. IEEE Transactions on Computers., 65: 819-833.
[3] Bryant, A., Cios, K. (2018) RNN-DBSCAN: A Density-Based Clustering Algorithm Using Reverse Nearest Neighbor Density Estimates. IEEE Transactions on Knowledge and Data Engineering., 30: 1109-1121.
[4] Lai, W., Zhou, M., Hu, F., et al. (2019) A New DBSCAN Parameters Determination Method Based on Improved MVO. IEEE Access., 7: 104085-104095.
[5] Shen, J., Hao, X., Liang, Z.Y., et al. (2016) Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm. IEEE Transactions on Image Processing., 25: 5933-5942.
[6] Wang, K., Yu, X., Xiong, Q., et al. (2019) Learning to Improve WLAN Indoor Positioning Accuracy Based on DBSCAN-KRF Algorithm From RSS Fingerprint Data. IEEE Access., 7: 72308-72315.