ZIBELINE INTERNATIONAL PUBLISHING

Malaysian Sports Journal (MSJ)

DOI : *http://doi.org/10.26480/msj.01.2019.14.16*

ISSN : 2710-5954 (Online)
CODEN: MSJABY

REVIEW ARTICLE

MINING OF CLUSTERING ALGORITHM IN THE FIELD OF HOT SPOT EVENTS IN SPORTS CULTURE

Yasha Shetty[1], Xianhou Chang[2*], Bo Chang[3]

[1]Institute of Sports Science & Technology, Pune, India
[2]Huainan Normal University, Anhui 232038, China
[3]Huainan United University, Anhui 232038, China
**Corresponding Author Email: xianhouchang4022010@21cn.com*

ARTICLE DETAILS

Article History:

Received 03 August 2019
Accepted 10 September 2019
Available online 08 October 2019

ABSTRACT

In order to explore the application of clustering algorithm in the sports culture with focus on public opinion, based on the analysis of experimental results data from artificial corpus and web crawler corpus, more reasonable parameters were determined for clustering, and the artificial test result of the execution of the algorithm is also checked. The results show that the CEW algorithm can achieve good results in mining the core keywords of online hot events. It also proves that the clustering algorithm performs well in accuracy rate, but it needs to improve in its recall rate.

KEYWORDS

Clustering algorithm, sports culture, public opinion hot spot issues, mining.

1. INTRODUCTION

With the continuous development of information technology in China, following the principle of "Moore's law", the highly open Internet brings media a disruptive change in the spread of public opinion [1] [2]. More and more traditional media audiences are involved in the process of information dissemination through the Internet, and they are becoming more active in using Internet to express their attitudes, opinions and requests more quickly and directly, more sharply and realistically. These expressions have contributed to the complex and changing public opinion on hotly debated issues in today's network and they can also reflect timely and truly the social public opinion [3-4].

In recent years, more and more attention has been paid to the public opinion of hot spot issues in the field of sports culture, and all of them have raised public interest on hot topics on the Internet. Different from traditional media in spreading public opinions, the spread of public opinion based on Internet is faster, and it is not easy to find bursts, nor is it easy to control it after the outbreak. This makes the rapid and effective detection and monitoring of public opinion in Internet very important [5-6].

In this paper, through the analysis of the conclusion of the corpus experiment and the crawler acquisition corpus experiment in the hot spot issues of sports culture, it is hoped to identify more reasonable clustering parameters and lay a good foundation for the future application of large data sets.

2. ALGORITHM DESIGN

2.1 Density algorithm theory

Different from other traditional clustering algorithms, density-based clustering algorithm, in theory, takes constraints of algorithm search based on the number of elements in the space unit area, so as to find clusters of arbitrary shape elements.

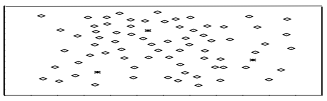

Figure 1: State description diagram of element distribution in space

As shown in Figure 1, it is assumed that the elements are distributed in a certain space, and each element has its own fixed position in the space and the distance between each element. As long as the density of elements near a region exceeds a threshold, it is clustered at the core of this element. According to the classical algorithm based on density clustering, such clustering algorithm can ignore the noise "noise" data and is suitable for finding arbitrary shape data clusters. The specific clustering process is shown in the following diagram.

Figure 2: The description of the near radius R and the threshold K

Based on the above description, the key of clustering lies in the following elements:

1. The determination of the adjacent area.

2. Threshold determination. The determination of the position of the element.

2.2 Data source clustering algorithm in this subject

The clustering algorithm completed two tasks: cluster nodes associated together and find word set with possible events of public opinion core word.

The key point of the data source clustering algorithm is how to construct a reasonable feature string association function D_{data}. Before constructing the D_{data}, first the decomposed word association function D_{word} is constructed.

By analyzing a large number of data corpuses, it is concluded that if two decomposed words often appear in the same feature string, the two decomposed words are more closely related. Therefore, it is possible to deduce that the number of two decomposed words, W1 and W2, appearing at the same time in all the feature strings, is x.

When the monotone of x rises, D_{word} (w1, w2, x) should be monotonically decreasing and infinitely approaching 0.

When x is a natural number, D_{word} (w1, w2, x)

When x=0, D_{word} (w1, w2, x) should be approaching to infinitely large number (relative to R).

When x=1, D_{word} (w1, w2, x) should be close to r (large or small)

If we simply quantify the relationship between the two words from the two words appearing at the same time, then the detection ability of the new public opinion outbreak event will be reduced as time goes on, so this is not reasonable. As a result, the effect of time on the value of the associated function is introduced. Based on the constraints of the above related functions, the following correlation function is constructed.

$$D_{word}(x) = \frac{\theta r}{x^t + 0.0001} \quad (1)$$

The r is the radius of the adjacent region, θ is the radius correction parameter, and the t is the time correction parameter. The longer the time the publication of an element is, the worse the public opinion goes. The relation between the specific parameters t and the optimization is shown in the following table:

Table 1: Prediction rules (decision tree) optimization

Time difference Δt	Inter(R)Core(TM)2 Duo P8700 2.53GHz (2 CPUs)
Δt<24hs	2
24hs<Δt<72hs	1
3 days<Δt<7 days	0.5
7 days<Δt<14 days	0.2
14days<Δt<30 days	0.1
30 days<Δt	0

After discussing $D_{word}(x)$, discuss the construction of the correlation function D_{data} (t_1, t_2) of the characteristic string $t_1$and t_2 is discussed. It is considered that the degree of association between the characteristic strings is determined by the degree of association between the effective words that make up the characteristic strings. Based on this idea, the following feature string association functions are constructed.

$$D_{data}(t_1, t_2, x) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} D_{word}(w_i, w_j, x_{ij})}{m \times n} \quad (2)$$

The implementation process of the data element clustering algorithm is shown in Figure 3:

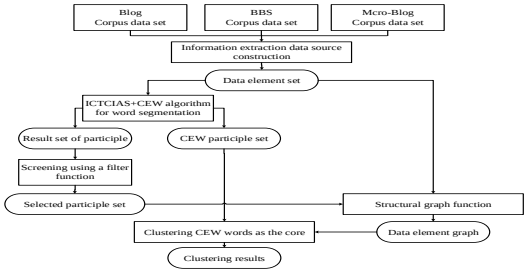

Figure 3: Data element clustering algorithm flow charts

From the flow chart introduced in Figure 3, we can see that after the completion of CEW word mining and undirected graph construction, the clustering results starting from a CEW word will not interfere with the clustering results starting from other CEW words. In other words, the distribution of clustering work can be deployed to different computers to improve the performance of the algorithm in processing large data sets.

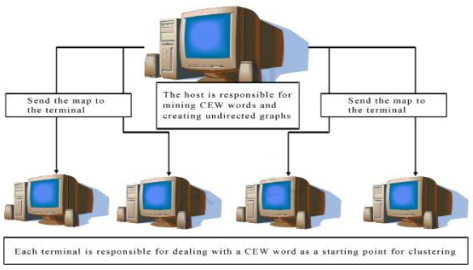

Figure 4: A schematic diagram on the calculation of clustering algorithm distribution

Because distributed computing technology is not within the scope of this topic, it will not be further discussed here. And the distributed computing method is not used in the cluster demonstration Demo used in this topic.

3. CLUSTERING ALGORITHM EXPERIMENT RESULTS

3.1 Clustering results for the corpus of artificial search organization

First, the data set of artificial organization and the results of CEW word mining are clustered. According to the theoretical design of the clustering algorithm, there are two important input parameters in the process of clustering: the clustering radius r, the clustering density threshold k. Among them, the clustering density k is no doubt related to the size of the set of data elements, so set the clustering density parameter as:

$$\text{Clustering density parameter} = \frac{\text{Clustering density threshold}}{\text{Total number of data sets}} \quad (3)$$

For clustering algorithms, there are two important metrics, accuracy rate and recall rate. The recall rate is the ratio of the number of related data elements and the number of all related data elements in the clustering result, which is the recall rate of the clustering algorithm. The accuracy rate is the ratio of the number of the correlation data element to the total data element in the clustering result, and it is the standard of precision of the clustering algorithm.

$$\text{Accuracy rate} = \frac{\text{Number of related data elements in clustering results}}{\text{Number of data elements in clustering results}} \quad (4)$$

$$\text{Recall} = \frac{\text{Number of related data elements in clustering results}}{\text{The number of all related data elements}} \quad (5)$$

Through analysis, it is found that the accuracy of clustering decreases monotonously, but the recall rate increases monotonously. The accuracy of clustering increases monotonously, but the recall rate decreases monotonously. This result is consistent with the theoretical conclusion of density-based clustering algorithm—when clustering radius increases, the clustering results will increase and consequently the result set will increase. As the denominator is invariant and the numerator increases, leading to the monotonically increasing recall rate; but for accuracy, increasing the numerator and denominator, its accuracy rate decreases. When the density of clustering increases monotonously, although the size of the result set will not change, the number of clusters will decrease, so the recall rate will decrease monotonously, and the accuracy will increase monotonously. For any clustering algorithm, the recall rate and accuracy rate are not the best. When the recall rate is high, the precision is low and the precision is high, the recall rate is low. It is often used to measure the accuracy of a retrieval system with the average value of the N accuracy rate under the N recall rate. The set of results of the cluster number is 1, as shown in Table 2

Table 2: Parameter table with core words in "Table tennis" data set in accordance with the requirements

Cluster radius	Cluster density parameter	Core words number	Cluster accuracy	Recall rate
1.2	2.2%	1	91%	80%
1.0	2.1%	1	94%	75%
0.8	1.4%	1	88%	52%
0.8	1.8%	1	97%	45%
0.8	2.1%	1	100%	31%

Considering the accuracy rate and recall rate of clustering, the cluster radius within 1.0~1.2 and the clustering density parameter within 2.0%~2.2% are selected as the benchmark for the next experiment.

Table 3: Performance result table for clustering algorithm of artificial corpus data set

CPU	Inter(R)Core(TM)2 Duo P8700 2.53GHz (2 CPUs)
Memory	2G
Data element quantity	1000 pieces
Maximum memory consumption	84740KB
Computation time	6243mm (max)

3.2 Clustering results for collecting corpus of network crawler

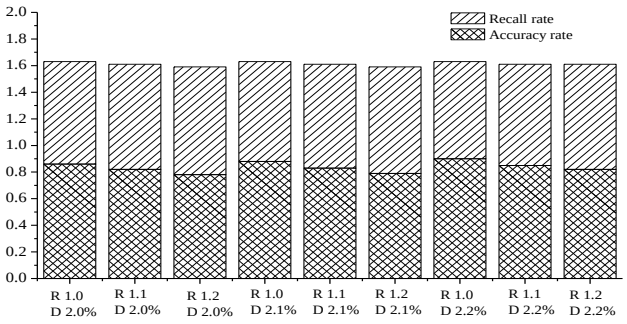

Figure 5: Statistical graph of clustering results for network crawler collection data set

Figure 5 reflects the accuracy rate and recall rate of clustering. Using clustering radius within 1.0~1.2 and clustering density parameter within 2.0%~2.2% as benchmark parameters, the corpus data set is clustered repeatedly. The results are shown in Figure 6: clustering within the range and the number of keywords obtained is 6.

3.3 Experiment result analysis

After analyzing the experimental results, it is found that the online hot spot event mining technology has the following characteristics:

First, the cluster radius parameter is within 1.0~1.2, cluster density parameter within 2.0~2.2, the number of keywords obtained by clustering algorithm (i.e. data cluster number) is stable. Compared with the artificial data sets results collected, and the result is correct. This proves the effectiveness of the algorithm.

Second, cluster analysis of corpus data collected are on multi carrier. The algorithm successfully excavated the core keywords of the hot spot events and achieved clustering, indicating that this topic is effective for the universal data acquisition method designed by different online public opinion carriers.

Third, compared with the experimental results in 3.1, the test results on real multi carrier corpus data coincide with the expected results, indicating that the clustering algorithm has the practical application value.

Fourth, compared with the accuracy rate and recall rate of clustering, it can be found that the accuracy of the algorithm is slightly better than the recall rate. This shows that the clustering results are less related to the data and cannot cluster all the relevant data items into the results.

Fifth, the meaning of core words and phrases that are partially excavated is not complete enough. Only by combining the semantics of related words can the core event content in the set of clustering results be understood.

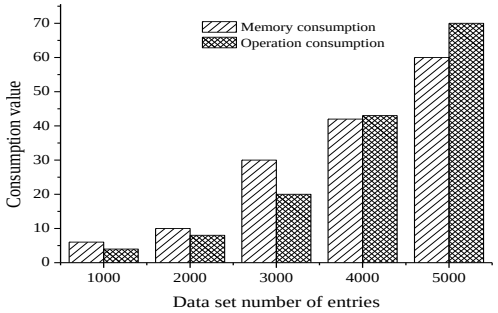

Figure 6: Comparison charts of the number of entries and algorithm implementation consumption in data clustering

From the chart, it is seen that the computational time grows faster than the rate of memory consumption, which is consistent with the analysis made before. Therefore, it can be gathered that if the algorithm is used for handling large data sets, it is necessary to mention the use of the distributed computing method to reduce the computational time.

4. CONCLUSION

Based on the experimental data and the artificially organized web crawler corpus results analysis, more reasonable parameters for clustering were determined, and the manual inspection of the implementation of the algorithm results carried out. It has been verified that the CEW algorithm in mining core keywords from hot spot social events can achieve good results and proved that it performed well in clustering algorithm when it comes to accuracy rate, but the performance needs to be improved in the recall rate. Moreover, by optimizing the program structure, the time complexity of the two algorithms is optimized to a linear level, which lays a good foundation for the future application of the algorithm on large data sets.

REFERENCES

[1] Yan, W. 2013. News on the internet: information and citizenship in the 21st century. Journalism. 14(4), 558-559.

[2] Chen, F., Wang, B., Wei, Z. 2014. The rise of the internet city in China: Production and consumption of internet information. Urban Studies. 52(13), 2313-2329.

[3] Brunzell, D.H., Chang, J.R., Schneider, B. 2013. China tightens media really do? —— Seek the Internet public opinion " normalization". Vol.184(3-4), 328-338.

[4] Luo, Y. 2014. The Internet and Agenda Setting in China: The Influence of Online Public Opinion on Media Coverage and Government Policy. International Journal of Communication. 8(2), 1289-1312.

[5] Tomz, M.R., Weeks, J. L. P. 2013. Public Opinion and the Democratic Peace. American Political Science Review. 107(4), 849-865.

[6] Gamsonand, W. A, Modigliani, A. 2015. Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. American Journal of Sociology. 95(1), 1-3.