ISSN: 2664-9640

Design of Topic Clustering Algorithm Based on K-Means

Longtian Fu[1] and Yumei Yu[2]

[1]FuZhou University of International Studies and Trade, Fuzhou 351010, China;

[2]Fujian Normal Univercity, Fuzhou 350011, China.

Abstract

With the rapid development of network technology and information technology, online public opinion has accumulated a lot of data. In order to better understand the trends of online public opinion, many scholars use a lot of artificial intelligence methods, but the complexity is high , The recognition rate is not ideal. This paper uses k-means to design an algorithm for detecting network public opinion topics. Simulation experiments show that the algorithm designed in this paper has low complexity, high recognition rate, and good stability.

Keywords

Topic detection,Massive data, K-Means.

1. Introduction

With the development of society, information technology and network technology have become more and more mature, and the amount of accumulated data worldwide has become larger and larger. Social platforms have also accumulated a lot of public opinion data. Public opinion data plays a very important role in the decision-making of government departments. Therefore, it is very important to understand the orientation of social public opinion by mass data. In 1996, the US Department of Defense initiated a new technology research on topic detection and tracking (TDT). The main goal is to detect hot topics from massive information [1]. In 1998, Allan scholars proposed a vector space model and experimented in news texts. Using cosine to represent distance to detect text similarity [2], the scholar's vector model has achieved great success and promoted the development of TDT technology Ponte and other scholars extended the vector model by adding feature information such as context on the basis of the vector model [3]; Masnizah Mohd and other scholars studied the topic structure in topic detection in 2011 and proposed a single-pass clustering method. Topics that have been detected are clustered [4]; in 2012, Jayashri and other scholars proposed a set of adaptive topic detection and tracking algorithms based on the text time of the samples to be inspected [5]. Based on previous studies, improvements were made and achieved good results; Yan scholar proposed the TF-IWF-IDF model in 2016. The scholar added position weights based on Heinrich scholars, and solved the problem of constant IDF value. Greatly reduced the false detection rate[6]; Naili scholars proposed a word vector model based on natural language tasks in 2017, and used semantic analysis to solve the topic clustering problem[7]. The research work of these scholars has greatly promoted the development of TDT technology. This article will design a public opinion topic clustering algorithm based on K-Means algorithm.

2. Topic Clustering Algorithm Based on K-Means

Clustering algorithms are a large class of algorithms for data mining, including Partitioning Methods, Hierarchical Methods, density-based methods, grid-based methods, model-based methods etc[8]. and each major category contains many famous Algorithm. This paper uses the K-means clustering algorithm (K-MEANS) in the partition method. This algorithm is an iterative

ISSN: 2664-9640

process. The basic principle is to first use K objects as the initial classification, and then calculate the distance between each object and other objects, Merge it into the nearest classification and iterate the process.

According to the principle of the K-MEANS algorithm, it is necessary to calculate the "distance" between topics, that is, the similarity between topic feature words [9]. Similarity calculation between topic feature words This paper uses the calculation method based on context word distribution similarity [10]. If two words have a high probability of appearing in the article at the same time, and the similarity of the distribution is high, then they should belong to the same class in this article. The process is then continuously iterated until the variance p of each classification is greater than the threshold α (pre-set).

Let P(xi|xj) represent the probability that xi appears when the word xj appears, and let P(xj|xi) mean the probability that xj appears when the word xi appears.

$$\mathrm{P}(x_i|x_j) = \frac{N(x_i,x_j)}{N(x_j)} \tag{1}$$

$$\mathrm{P}(x_j|x_i) = \frac{N(x_i,x_j)}{N(x_i)} \tag{2}$$

$$\mathrm{Sim}(x_i, x_j) = \frac{\mathrm{P}(x_i|x_j)}{\mathrm{P}(x_j|x_i)} \tag{3}$$

N(xi,xj)represents the number of simultaneous occurrences of the words xi and xj, and N(xi)and N(xj) represent the respective occurrences of the words xi and xj, respectively. The calculation of similarity is shown in formula (3). By finding the ratio of the two, a threshold δ is finally set. When the similarity is greater than δ, it can be considered to be merged, otherwise it does not belong to the same category. The specific algorithm implementation process is shown below.

```
Algorithm: Topic clustering based on K-MEANS
Input: K objects (one thesaurus)
Output: Optimize thesaurus
begin
    Set the classification variance threshold α;
    Enter the thesaurus term of this article;
    Set each word in the thesaurus as a category;
    while (p<=α) {
        Calculate the classification variance p;
         If p> α, then exit;
         else continue
         for (i <= number of objects) {
             for (j <= number of objects -1) {
                 Calculate sim (xi, xj);
                 if (sim> δ) then xi, xj merge into the same class
                 else do not merge and continue;
             }
         }
    }
end
```

ISSN: 2664-9640

3. Simulation

The experimental data comes from WeChat chat data of a cyber fraud case in a public security department in a certain place. The amount of data is 10G, and the 47 most frequent feature words are detected. According to the detection results provided by the public security department, a total of 49 high-frequency feature words. Therefore, the accuracy rate is pr = 100 * 47/49 = 95.9%, and the missed detection rate is Mr = 100 × (49-47) /49=4.1%. The evaluation results show that the topic detection algorithm designed in this paper performs well and has high accuracy.

4. Conclusion

The algorithm designed in this paper has been run in a grassroots public security department for three months. Practice has proved that the algorithm has a high detection rate, stable operation and good adaptability.

References

[1] TDT Homepage at the National Institute of Standards and Technology [EB/OL], http: // www_nist. gov/ TDT.

[2] Allan J,Carbonell J, Doddington G, etal.Topic Detection and Tracking Pilot Study[J].1998.

[3] Ponte J and Croft W B.Text segmentation by topic.in:Carol Peters,Costantino hanos eds.Proceedings of Research and Advanced Technology for Digital Libraries. Pisa Italy.1997. Europe: ECDL, 1997. 113-125.

[4] MasnizahMohd, Fabio Crestani, Ian Ruthven. Construction of Topics and ClustersinTopic Detection and Tracking Tasks.in: 2011 International Conference on SemanticTechnology and Information RetrievalMalaysia2011. US: IEEE Computer Society, 2011.28-29 .

[5] Jayashri M,Chitra P.Topic clustering and topic evolution based on temporal parameters. in: International Conference on Recent TrendsIn Information Technology(ICRTIT).India. 2012. USA: IEEE, 2012. 559-564.

[6] Yan D,Hua E,Hu B.An Improved Single-Pass Algorithm for Chinese Microblog Topic Detection and Tracking[C]// IEEE International Congress on Big Data. IEEE, 2016.

[7] Naili M,Chaibi A H,Ghezala H H B.Comparative study of word embedding methods in topic segmentation[J].Procedia Computer Science,2017,112:340-349

[8] Liang Haiping, Tian Chao, Wang Tieqiang, Cao Xin, Yang Xiaodong, Liu Yingpei. Research on Similarity Matching of Power System Operating Sections Based on Improved K-means Algorithm [J / OL]. Electric Power Automation Equipment, 2019 (07): 1-7 [2019-07-15] .https: // doi. org/10.16081 / j.issn.1006-6047.2019.07.018.

[9] Xia Weizhao, Teng Huan, Cao Min. Research on Evaluation Model of Three-phase Electric Energy Meter Based on K-means ++ Algorithm [J / OL] .Electrical Measurement and Instrumentation: 1-5 [2019-07-15] .http: // kns.cnki.net/kcms/detail/23.1202.TH. 20190709.1431.006.html.

[10] Xue Suzhi, Lu Ran, Ren Yuanyuan. Hot topic discovery of microblogs based on speed increase [J]. Application Research of Computers, 2013, 30 (09): 2598-2601.