An improved Agglomerative levels K-means clustering algorithm

Yu jiankun
School of Information
Yunnan University of Finance and Economics
Kunming, China
Yjk1102@163.com

Guo jun
School of Information
Yunnan University of Finance and Economics
Kunming, China
564066191@qq.com

***Abstract*—The paper proposed a method which combines an improved hierarchical aggregation and K-means clustering algorithm, overcoming the selection problem of initial cluster centers and selection problem of termination condition. Application this method to cluster sina weibo topic and compare with tradition hierarchical aggregation and K-means clustering algorithm, finding the method can reduce false positives and missed rate.**

Keywords- K-means, Agglomerative hierarchical clustering, initial cancroids, termination condition

I. INTRODUCTION

Cluster analysis is an unsupervised learning method, there are partitioning method, the density method, level-based approach, the grid-based and model-based approach, which is based on their characteristics and their relationships to describe data objects, according to the cluster maximize the correlation between the object and the cluster correlation data minimization principle, data collection packet [1,2]. Now, cluster analysis has been in pattern recognition, machine learning, text mining, information retrieval and other research fields widely used, as one of a very active field of data mining research.

In each clustering algorithm, K-means algorithm is the classic division method, and Agglomerative hierarchical clustering algorithm represents the hierarchical clustering algorithm. K-means algorithm according to some measure of approximation strategy to divide each data object to its most similar clusters, and each cluster center is calculated and updated, iterative updating assignments and repeat steps until all clusters in the data object is no longer changed so far [3,4], K-means algorithm is simple, high efficiency, but known in advance of the initial cluster centers and the number of clusters K cluster random selection will cause the clustering of clusters of uncertainty; Agglomerative hierarchical clustering algorithm is based on the formation of the way to achieve data clustering tree from the bottom up, each data object will be seen as a cluster, merging with optimal proximity clusters according to a certain criterion, until the termination condition to meet or all data objects are gathered in a cluster [5,6], hierarchical clustering algorithm has higher accuracy, but the time complexity and space complexity is high, resulting in its not suitable for large-scale data sets in the calculation. Given the advantages and disadvantages this paper presents an improved combination method, and gives experiments show that the method is feasible.

This research is funded by Business Intelligence Technology Innovation Team Foundation of Yunnan University and Natural Science Foundation of Yunnan Province (2009CD076).

II. RELATED RESEARCH

In the hierarchical clustering algorithm involved selection of the termination condition, the condition is usually the number of class clusters and distance threshold, but in reality, select these two parameters requires a lot of Related knowledge, and the result will be selected related with the selected human subjectivity, resulting clustering inaccuracies, by observing the change of the outlier number to achieve termination condition judgment threshold.

Outliers Detection is another branch of data mining, achieved by detecting the dissimilarity of data object. Variety descriptions of outliers, and widely accepted is given by Hawkins [7], that outliers deviate from the other observation object, the data object is believed by different mechanisms. The basic method of outlier detection is calculating the distance between objects, and compared with a threshold value set in advance, if the distance is greater than the threshold value, then the point is outlier [8-9]. This Outliers Detection method based on the distance you need to set in advance the appropriate threshold, and ignores the positional distribution data characteristics.

TABLE I. SYMBOL MEANING TABLE

symbol	*Description*
N	The total number of data objects
d	Data dimension
Dnn（i）	Data I distance between it's nearest neighbor
Nout	Outliers numbers

Algorithm ideas: first using the formula (1) to calculate the distance between data objects, select two data values with the smallest distance as the nearest neighbor; followed step β values issue and the process is shown as in Figure 1, Finally, to judge whether the peak V, if yes, end of the operation, get Agglomerative hierarchical clustering algorithm termination condition, if not, continue outlier detection operation until V reaches a peak. The algorithm implementation as Figure 2 shown.

$$dist(p,q)=\sqrt{\sum_{i=1}^{d}(p_i-q_i)^2} \tag{1}$$

$$\overline{D_{NN}}=\sqrt[M]{\sum_{i=1}^{M}\left(\left(a_{\max}^{(i)}-a_{\min}^{(i)}\right)/\sqrt[M]{N}\right)^2} \tag{2}$$

978-1-4799-6543-4/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMeCG.2014.53

Wherein, M is the dimension of the vector data, $a_{\max}^{(i)}$ and $a_{\min}^{(i)}$ represents the maximum and minimum data dimension's i, N is the number of data objects.

1. When first detect outliers, $\beta = \beta_{step} = D_{NN}(\alpha_{first}) / \overline{D_{NN}}$.
$D_{NN}(\alpha_{first})$ maximum values for the data sets. $\overline{D_{NN}}$
Calculation formula as (2) shown.
2. making $v = \Delta n_{out} / \Delta \beta = \Delta n_{out} / \beta_{step}$,
Δn_{outs} number incremental of outliers. $\beta = L \times \beta_{step}$,
$L = \{1,2,\cdots,\}$ opration step numbers.
3. when v first achieve peak. $\beta = (L_i - 1) \times \beta_{step}$

Figure 1 β Values Implementation

To obtain Agglomerative hierarchical clustering algorithm termination condition.

Input: N data objects.
Output: clustering algorithm termination condition ξ.
Procedure:
1. for i=1 to N.
2. using Equation (1) calculates the distance between data I and any other data object distance: dist(i, j).
3. End for.
4. Determine β value.
5. $\xi = \sqrt[M]{\sum_{i=1}^{M}((a_{Max}^{(i)} - a_{min}^{(i)})/\sqrt[M]{N})^2} \Big/ \beta$.
6. for i=1 to N.
7. $D_{NN}(i) = \min\{dist(i,j), (\forall i,j, i \neq j)\}$.
8. if $D_{NN}(i) > \xi$.
9. data object I is outlier. $n_{out} = n_{out} + 1$.
10. If first calculating n_{out}.
11. yes. jumping to step 6.
12. no. calculating Δn_{out}, updating v.
13. updating β value.
14. until v reduced.
15. End for.
16. Using β value updating ξ, and β value before v's decrease.
17. Output ξ

Figure 2 To obtain Agglomerative hierarchical clustering algorithm termination condition

III. IMPROVED AGGREGATION HIERARCHICAL AND K-MEANS ALGORITHM

Improved Agglomerative hierarchical clustering algorithm does not need pre- termination conditions, but basing on Part 2 threshold value, by calculating aggregation hierarchical clustering algorithm clusters' distance between any two classes, and compared with a threshold value determine whether or not to terminate the algorithm. Although, to a certain, can overcome deficiencies of existing aggregation Hierarchical clustering algorithms, but its time complexity of higher demand is still not improved. K-means algorithm is simple, efficient and time complexity is not high, but need to pre-set the initial cluster centers and cluster number of clusters. Based on this, combining improved Agglomerative hierarchical clustering algorithm and K-means algorithm, both reduces hierarchical clustering algorithm's time complexity and overcomes the K-means algorithm's initial cluster centers selection sensitive issues.

The basic idea is to use hierarchical clustering algorithm to improve portion of the data clustering, calculating average of each cluster which is hierarchical clustering results, and using this average as the initial cluster centers of K-means algorithm, using K-means algorithm to achieve remaining data clustering. Algorithm specific implementation steps shown in Figure 3. Algorithm involved Calculation shown as (3).

$$\mu_k = \frac{1}{N_K} \sum_{j \in cluster_K} x_j \qquad (3)$$

Algorithm: Improved aggregation Hierarchical K-means clustering algorithm.

Input: Data Sets N and Clustering algorithm termination condition ξ.
Output: clustering results.
过程:
1. From the data collection N randomly selected portion of the data K.
2. Calculate data sets K, generate data matrix.
3. each data object as a separate cluster.
4. repeat.
5. Calculating distance value between any two data objects, two objects with the nearest value merger as a class.
6. Computing the new clusters and other clusters distance, update the distance matrix.
7. Until distance between two nearest classes greater than the threshold ξ, derived clusters number.
8. Using (3) to calculate mean of each cluster data objects as the initial cluster centers of K-means algorithm, K-means algorithm number of initial cluster and the number of class clusters generated by hierarchical clustering.
9. input (N-K) data sets.
10. repeat.
11. calculation distance between data objects and cluster centers, if the distance is less than the set threshold value, Classified as the same category; Otherwise, as a collection of new data.
12. update cluster centers and the number of class clusters.
13. until data division is completed.
Output clustering results, the algorithm ends.

Figure 3 Improved aggregation Hierarchical K-means clustering algorithm

IV. EXPERIMENT

In this paper, through data sets topic clustering, to verify method's feasibility and validity.

A. Experimental Data

Using sina's open application interface (Referred as API) grab sina micro-blog messages as data sources in this article. Sina Micro-blog's API is provided by sina service provider to obtain data/service access, need use OAUTH2.0 or Basic Auth to achieve authentication, obtaining appropriate App Key、APP Secret 、Acess Token to achieve sina application's data obtaining、Information posting and other operations. Using sina's API to obtain data sets, return XML or JSON format file. This section getting the file format is JSON, authorized authentication method is oauth2.0.

When using the API to obtain data, need to create application on sina mcro-blog open platform, getting APP Key and App Secret, and for the application setting the callback address and cancel the callback address.

Application specific implementation is based on the Java language to create SinaMicroData project in the Eclipse environment, first configure the project config.properties file. Acquiring App Key and APP Secret, and callback address (callback when creating applications Weibo address is set to http://127.0.0.1) were assigned to the corresponding variable, and specify the interfaces and URL which related data storage formats, in order to achieve Weibo authorization, data acquisition, etc. After authorization, will receive a variable of type string: Access Token, According to Access Token variables and statuses / public_timeline interface to obtain the latest public microblogging; as Comment Informations' parameter, using Comments interface to get message of comments and forward. Specific implementation steps shown as Figure 4.

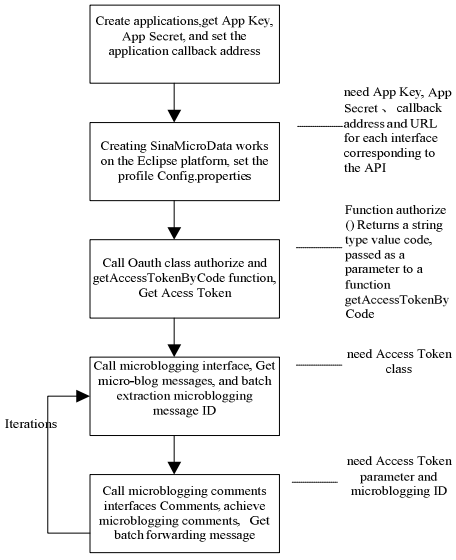

Figure 4 Microblogging Data Acquisition

Using Sina microblogging API to get the message format is JSON, which contains many such "gender, followers_count, friends_count" microblogging users labels,and etc. In addition to these label information, microblogging information obtained contains many irrelevant data, such as duplicate microblogging, microblogging advertising, so use regular expressions to parse out microblogging user ID, microblogging messages and and other attributes ,removing duplicate entries and advertising microblogging.

B. Preprocessing Experimental Data

Experimental data preprocessing including segmentation, stop word filtering, feature extraction and feature items weight calculation.

(1) Segmentation

In this paper, using Chinese Academy of Sciences ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) Chinese lexical analysis system to achieve segmentation and POS tagging. Data from the network, including a large network terminology, homophonic words and some other irregular characters need to be normalized and extended ICTCLAS dictionary.

(2) Stopwords filter

Building stopwords vocabulary, lexical items and disable vocabularies are contrasted, if appear in the table, election it, completing stopwords filtering.

(3)Extracting feature item

Using DF (Document Frequency) methods to achieve feature item extraction.

(4)Feature weight calculation

Calculate the feature weight using tf-idf method.

C. Experimental Design and Analysis

Verify the effectiveness of this method in two ways.

(1) By comparing the experimental validation the proposed method is feasible. Agglomerative hierarchical clustering algorithm and K-means algorithm as the benchmark experiments, with proposed method to implemented topic clustering, using missing rate and false positive rate as experiment evaluation criteria which used commonly in the field of Topic detection and tracking [11-12], verify the effectiveness of this method. Wherein, missing rate is undetected miscro-blog numbers and all micro-blog total numbers (both kind of them are related a certain micro-blog topic i) ratio; False detection rate is ratio between numbers micro-blog message which doesn't belong microblogging topic i but vesting mistake to I and all micro-blog topic i numbers that do not belong to it[13-14]. Calculating formula as (4) and (5).

$$p_{missi} = \frac{MB}{SMB} \times 100\% \quad (4)$$

$$p_{FAi} = \frac{FA_i}{NT} \times 100\% \quad (5)$$

During the experiment, first, preprocessing microblogging message's documents set, obtaining vector space which is consist of documents feature words, and then using Agglomerative traditional hierarchical clustering algorithm, K-means clustering algorithm, and

the method to achieve topic clustering. Results shown in Table 2.

TABLE II CLUSTERING RESULTS

algorithm	Pmin	Pfa
Traditional Agglomerative hierarchical clustering algorithm	0.2623	0.04289
K-means clustering algorithm	0.2738	0.04373
Improved aggregation Hierarchical K-means clustering algorithm	0.2568	0.03901

As can be seen from Table 2, the proposed method compared with K-means algorithm and the traditional hierarchical clustering algorithm, in topic detection, reducing the false rate and missed rate. This is because the proposed method does not require people to determine the cluster number of clusters, avoiding aggregation hierarchical clustering algorithm terminates deviation, K-means algorithm's initial cluster centers selection is based on improved hierarchical clustering algorithms, to certain extent, avoiding sensitivity which is produced by randomly selected the cluster center, coupled with the validity and accuracy of the hierarchical clustering algorithm itself, making K-means clustering initial center selection compared to the random selection method is more fit, so that the algorithm has good performance in topic detection.

Also be seen from the table, although missing rate and false detection rate has decreased, but the effect is not obvious, it is because microblogging message text is shorter, grammar is not standardized, contains more network terminology, using VSM text representation model ,data sparsely more serious, affecting the results of clustering algorithms.

(2) From the actual situation to illustrate the effectiveness of this method. Using paper proposed method obtain result clusters and comparing with topics which come from Weibo feature list, Results shown in Table 3.

Table IV Comparison of experimental results

Topic List	*Proposed method*	*microblogging own features*
PubMed	PubMed sprint dream	PubMed 2014 Dream
People to face shame	Officials shameless people	people Officials To face
Playing basketball was punishment	Chinese education civilization play basketball	civilization basketball education
Price ladder	Water charge Saving Water using	Water charge Water price Saving

For comparative analysis of proposed method, the top three clusters and topic clustering feature list of keywords extracted, the results shown in the table above.

V. CONCLUSION

In this paper, basis aggregation Hierarchical Clustering algorithms, introduction of outlier detection algorithm ideas, automatically obtain termination condition, improved Agglomerative hierarchical clustering algorithm, finally, through the acquisition of Sina Weibo message set as an experimental data sources, through data source to detection topic and evaluation its results , show that the proposed method is feasible. Analysis of the experimental results can be found, this is because microblogging message text is usually limited to 140 characters, and the grammar is not standardized, network terminology more, for these problems, the next step is how to select the main research work or improve the text representation model to make it more suitable for the micro-blog message text clustering analysis. This is because microblogging message text is usually limited to 140 characters, and the grammar is not standardized, network terminology more, for these problems, the next step is how to select the main research work or improve the text representation model to make it more suitable for the micro-blog message text clustering analysis.

REFERENCES

[1] J.G. Sun, J. Liu, L.Zhao. Clustering Algorithm. Journal of Software, 2008, pp. 48~61

[2] R. Xu, D. Wunsch. Survey of clustering algorithms, Neural Networks, IEEE Transactions on, 2005, pp.645~678

[3] T.Su, J.Dy. A deterministic method for initializing K-means clustering. Tools with Artificial Intelligence, ICTAI 2004. 16th IEEE International Conference on. IEEE, 2004,pp:784-786

[4] T. Huang, S.Y. Liu and Y.N.Tan. Research on K-means clustering algorithm, Computer Technology and Development, vol. 21, no. 7, 2011, pp. 54~62

[5] A. K. Jain, R C. Dubes. Algorithms for clustering data, Prentice-Hall, Inc., 1988.

[6] X.Y.Ma,Y.Tang. Hierarchical clustering algorithm, Computer Science, vol. 34, no. 7, 2008, pp. 34~36.

[7] L davies, U Davies. The identification of multiple outliers, Journal of the American Statistical Association, 1993, pp. 782-792.

[8] E.M. Knorr, R.T.Ng. Finding intensional knowledge of distance-based outliers, VLDB, 1999, pp.211-222.

[9] S. Ramaswamy, R. Rastogi, K. Shim. Efficient algorithms for mining outliers from large data sets , ACM SIGMOD Record. ACM, 2000, pp. 427-438.

[10] T.Lv, T.Su, Z.Wang,et al. An auto-stopped hierarchical clustering algorithm integrating outlier detection algorithm.Advances in Web-Age Information Management. Springer Berlin Heidelberg, 2005,pp:464-474.

[11] Y.Hong,Y.Zhang,T.Liu. Summary of evaluation and research topic detection and tracking. Journal of Chinese Information,vol.21,no.6,2007,pp:71-87.

[12] H.W.Luo,Q.Liu,X.Q.Cheng. Development and research topic detection and tracking technology, National Joint Conference on Computational Linguistics (JSCL-2003) Proceedings, Beijing,Tsinghua University Press,2003,pp:560-566.