Send Orders for Reprints to reprints@benthamscience.ae

Open Access

Improved K-Means Algorithm in Text Semantic Clustering

Junhong Ma*

Xi'an International University, Shaanxi, 710000, China

Abstract: Text clustering is a very important technology in the area of text data mining. The semantic calculation method can greatly improve the computational. The aim of this paper is to improve the existing text clustering algorithms, for Chinese text and used semantic clustering method. First, in similarity calculation module of the clustering, used a staged and integrated semantic similarity algorithm, the text semantic and context factors was blended in the each computation stage. Then improved the traditional K-means algorithm, which used the priority strategy to divide up the relative centralize data at first, reduced the randomness of initial center, ensured that each cluster partition of the sample points have high similarity. Finally, the experiments have proved that the proposed algorithm not only can improve the accuracy of clustering, but also has the very high stability.

Keywords: K-means, semantic clustering, semantic similarity, similarity calculation, text clustering.

1. INTRODUCTION

Clustering divides data into meaningful or useful groups (clusters) without any prior knowledge [1, 2]. Text clustering is a fundamental and important work in information processing, which also is the key technology in data mining, machine learning, statistics, and biology, and etc. important issue in many fields [3, 4]. With the amount of all kinds of information and data in the world increasing and the study problems becoming more and more complex, the existing clustering techniques are also facing increasing challenges. The traditional approaches are sensitive to initializations and easy to be trapped in local optimal solutions [5]. In addition, understanding Chinese language from the view of semantic is more appropriate than from the statistical method. Therefore, the text clustering method based on the semantic similarity is growing concerned.

Semantic understanding algorithm focuses more on the texts of deeper meaning relationship, combined of context to explore deeper semantic representation, can get more accurate similar values of two texts, and thus get more accurate clustering calculation results [6]. At present, most of the clustering algorithms are based on statistical theory. For example, the clustering method based on VSM use the vector of Euclidean distance or cosine distance to calculate the relationship between texts [7], the basic idea is according to such as word frequency statistics information to get the feature term weights, and formatted vectors [8]. This approach ignores the semantic correlation between words and words, documents and documents thus reduced the clustering accuracy [9]. Text clustering principle is: the similarity between associated documents is a greater than the similarity between documents that are not associated. It is the most basic requirement is the more similarity of texts to a category or cluster, *i.e.* the lower similarity of texts into the other cluster [10]. Therefore, during cluster analysis, Similarity calculation is the foundation to achieve text clustering, different text similarity calculation method will produce different clustering results.

The key of text clustering are an effective solution to text similarity calculation and method to determine the cluster centers [11]. In the similarity calculation of text clustering, established a model based on semantic understanding, further improve the k-means algorithm, which considered with semantic factors when calculating, strive to make the text clustering results more accurate.

2. TEXT CLUSTERING METHOD BASED ON SEMANTIC MODEL

2.1. The General Processes of Text Clustering

The large-scale text clustering was an effective way to solve data understanding and information mining in the massive text [12]. Clustering which was general use as the front-end application, included the steps of text representation, clustered (and implementation of clustering algorithm selection), the effect of assessment. In the process of semantic clustering, the similarity calculation was one of the key modules. Before clustering it need to use the text similarity calculation method to establish the similarity matrix, and then appropriate clustering algorithm for clustering. So, good similarity calculation method can greatly improve the efficiency of the clustering [13].

The specific processes are shown in Fig. (**1**):

2.2. Semantic Clustering Based on Text Similarity Calculation

As it can be seen from the Fig, selecting the appropriate clustering algorithm and text similarity calculation is vital which will exert great influence on the result of the final clustering [14]. Calculation of text similarity need to go through

1874-110X/14 2014 Bentham Open

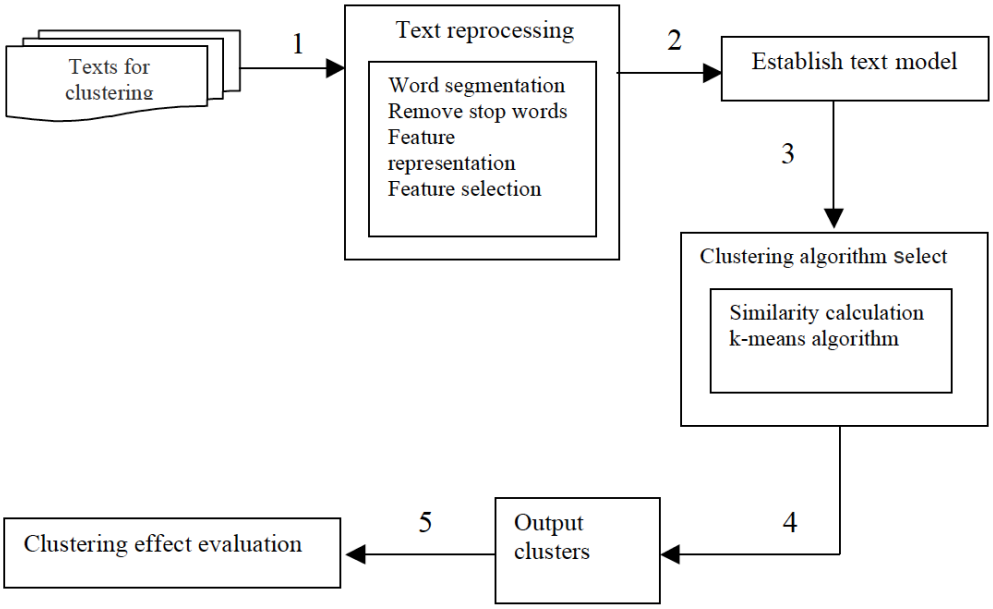

Fig. (1). General process of text clustering.

the word segmentation, removing stop words, feature selection and preprocessing steps, then can create text representation model [15]. Next, we used the semantic similarity calculation method to compute similarity, established a semantic model, generated similarity matrix, getting ready for further clustering.

The basic idea of this algorithm was: first of all for segmenting down text, completed the main classified work for the texts were divided into paragraphs, paragraphs into sentences, sentences into words; secondly for text feature selection; then calculated the similarity phased from words, sentences, paragraphs, and Fused calculated the result of the text similarity [16]. At each stage incorporated the factor of semantic similarity computation, completed the combination of partial to the whole.

Combined with the specific application of the similarity, improved semantic clustering based on text similarity calculation model is available to the following Fig. (**2**).

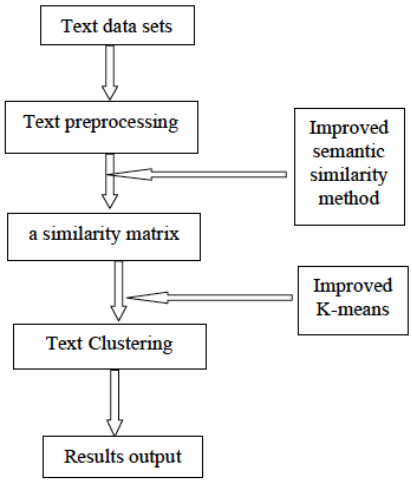

Fig. (2). Semantic clustering based on text similarity calculation.

3. IMPROVED K-MEANS ALGORITHM

3.1. Determined the Initial Centers

This paper used Improved K-means algorithm to implement clustering. Original K-Means clustering algorithm was randomly selected K numbers of sample points as the initial cluster centers sets, each of which represented a cluster center [17]. But K-Means clustering algorithm for these random selections of the initial cluster centers was very sensitive, not the same as the initial cluster centers tended to get very different clustering results [18]. Because K-Means algorithm is an iterative updated method, so when the initial cluster centers fell near the local value of the optimal range, the entire clustering algorithm is easy to fall into local optima. Therefore, in order to obtain better clustering results, should be improved from the selection of the initial cluster centers by reducing the randomness of the initial centers to achieve the purpose of optimizing the clustering algorithm.

Improved initial cluster centers thinking strategy identified herein was every time delineated relatively centralized data first [19], so that it can ensure that each divided cluster sample points had a higher similarity. Specific steps of the improved algorithm were the following.

INPUT: N data objects in data sets and k clusters

OUTPUT: set of k clusters

Algorithm:

For i=1: K-1

For i=1: K-1

1) Find the maximum distance point of the point and other points, denoted Mi1;

2) Find the farthest point Mi2 from points Mi1;

3) Let the distance from Mi2 point be less than or equal to point of N/K small elements (i.e. the distance of N / K points closer from Mi2 point) divided into clusters

4) Removed the data objects from the data set had been classified as cluster i, and found the center point of the cluster i;

End

The rest of the sample sets were classified into cluster K, and also determined the cluster centers of cluster K. Wherein, K was the number of clusters.

3.2. Improved K-means Algorithm Based on Initial Centers Strategy

According to 3.1, the improved algorithm implementation steps are as follows:

Step 1: Set i = 1, initialized k numbers cluster centers;

Step 2: Statistics each text vector and the k numbers of cluster center distance;

$$D\left(x_i, Z_m(i)\right), i = 1,2,...k; m = 1,2,...,k; \quad (1)$$

$$D\left(x_i, Z_m(i)\right) = \min\{D\left(x_i, Z_m(i)\right)\}, \quad m = 1,2,...,n; x_i \in W_k; \quad (2)$$

Step 3: According to the similarity of data objects between the K clusters, each object was assigned to the most similar clusters again;

Step 4: Calculated the mean of all the objects in each cluster to form a new cluster centers;

Step 5: According to the following target optimization function to calculate:

$$f(i) = \sum_{m=1}^{k} \sum_{k=1}^{n_m} \left\| x^{(i)} - Z_m(i) \right\|^2 \quad (3)$$

Above (3), $\left\| x^{(i)} - Z_m(i) \right\|^2$ represented the distance of vector xi to the cluster center;

Step 6: if $\left| f(i) - f(i-1) \right| < \varepsilon$, then end clustering.

Otherwise: use i=i+1, re-calculate the cluster center use the following formula 3:

$$Z_m(i) = \frac{1}{n} \sum_{i=1}^{n_m} x_i^{(m)} \quad (4)$$

Then it returned to step 2 to continue.

4. EXPERIMENTS AND RESULTS ANALYSIS

4.1. Data Sets of the Experiments

Data sets of the experiments were from Sina weibo, which were selected from four text categories, namely education, entertainment, food, politics, and selected 200 texts from each category respectively, these 800 texts reconstituted without a category labeled corpus, then used the 800 texts clustering tests. Common method of assessing the effect of clustering was: Selected the documents had been manual divided into many categories or labeled as a test set. At the end of the clustering; the clustering results were compared [20] with the existing manual classification results.

4.2. Text Clustering Testing Results

Experiments were carried out by the traditional k-means clustering method and improved k-means algorithm in semantic clustering, finally assessed with precision(p), recall(r) and F-measure value(F value) [21].

$$p = \frac{\text{number of correctly detected similar texts}}{\text{number of all the detected similar texts}} \quad (5)$$

$$r = \frac{\text{number of correctly detected similar texts}}{\text{number of actual existence of similar texts}} \quad (6)$$

$$\text{F} = \frac{2pr}{p+r} \quad (7)$$

The test results are shown in Table **1** and Table **2**.

As can be seen from the Table **1** and Table **2**, the precision and recall of the traditional K-means clustering are all lower than the improved K-means method in semantic clustering. According to Table **2** to calculate, the global precision and the global recall respectively of the traditional K-means are 81% and 92%, the improved semantic clustering are 90% and 95%.

After that, we also used the original algorithm and the improved algorithms with the other categories of datasets by Randomly test 10 times, the results in F value refers to the following Fig. (**3**).

Fig. (**3**) presents the results of the experiment: the F value of traditional K-means clustering volatility fluctuates between 76% -90%, while the improved algorithm is always about 91%. The reason why the F value of traditional K-means algorithm is not high and unstable is due to the double impacts of the lack of semantic analysis in similarity calculation and the initial cluster centers randomness. The improved

Table 1. Test results of traditional K-means.

Category	Clustering Results	Manual Results	Same Texts	Precision	Recall	F Value
Education	220	200	181	82	91	86
Entertainment	235	200	190	81	95	87
Food	204	200	176	86	88	87
Politics	246	200	187	76	94	84

Table 2. Test results of Improved K-means in semantic clustering.

Category	Clustering results	Manual results	Same texts	Precision	Recall	F value
Education	215	200	195	91	98	94
Entertainment	196	200	179	91	90	90
Food	208	200	188	90	94	92
Politics	219	200	194	89	97	93

clustering method is based on semantics, and adopted strategies to improve initial center algorithm. Thus the results compared to the traditional K-Means clustering algorithm not only improved accuracy and algorithm but also run relatively stable.

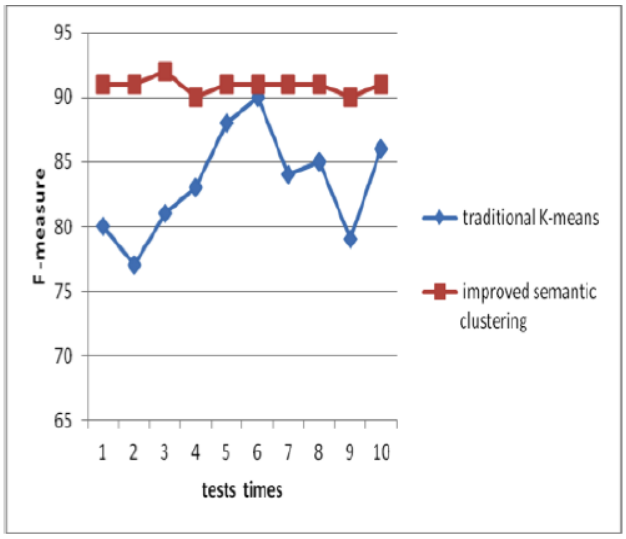

Fig. (3). Randomly test 10 times the F value comparison chart.

CONCLUSION

Text clustering is an unsupervised learning process [21]. It is based on some distance between samples to unsupervised clustering process. Using the clustering method can take a large amount of text be divided into the cluster which user can understand .The text within a cluster having a similarity, thereby enabling the user can quickly grasp the contents contained in the large amount of text, to accelerate the speed of analysis and decision support.

The choice of the clustering algorithm is often accompanied by the choice of a similarity calculation method [22]. Therefore, the appropriate calculation method is very important to text clustering. The text clustering based on semantics in this paper shows efficiency than the traditional method, the semantic strengthen model raised the computational efficiency. This shows that the semantic understanding method is better than statistical methods for Chinese text. The key areas of our future work include: continue to study on semantic clustering algorithm, applying this method to more specific fields, such as Internet public opinion analysis, question answering systems, enterprise competitive intelligence, etc.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is financially supported by the scientific research project fund of Shaanxi provincial education department of China (Grant NO.2013JK1146), and the natural science foundation research project fund of Shaanxi provincial science and technology department of China (Grant NO. 2014JM8323).

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Philosophical Transactions of the Royal Society of London*, vol. A247, pp. 529-551, April 1955.

[2] J. C. Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford, Clarendon, 1892, pp. 68-73.

[3] H.T. Zheng, B.Y. Kang, and H.G. Kim, "Exploiting noun phrases and semantic relationships for text document clustering", *Information Sciences,* vol. 13, pp. 2249-2262, 2009 (http://www. sciencedirect. com/science/article/pii/ S0020025509001108 - cor1).

[4] Y.T. Liu, and J. Xiong, "Research on semantic relevancy calculating model of natural language and its weight coefficients", *Application Research of Computers,* vol. 31, no. 6, pp.1661-1664, 2014.

[5] S. S. Khana, and A. Ahmad,"Cluster center initialization algorithm for K-modes clustering", *Expert Systems with Applications,* vol. 40, no. 18, pp. 7444-7456, 2013.

[6] Y.G. Wang, L. Lin, and X.G. Liu, "Research on text clustering algorithm based on improved particle swarm optimization", *Computer Engineering,* vol. 40, no. 11, pp. 172-177, 2014.

[7] X. S. Hang, and J. Pan, "A novel k-means clustering based on the immune programming algorithm", *Chinese Journal of Computers,* vol. 26, no. 5, pp. 94-99, 2003.

[8] B. Yang, D.Y. Liu, and J.M. Liu, "Complex network clustering algorithms", *Journal of Software*, vol. 20, no. 1, pp. 54-66, 2009.

[9] Q. Geng, *Research and Implementation of the Text Cluster Based on Text Similarity Calculation*, Harbin Engineering University, China, 2010.

[10] J.C. Chen, G.W. Hu, and Z.H. Yang, "Text clustering based on global center-determination", *Computer Engineering and Applications,* vol. 47, pp. 147-150, 2011.

[11] S. Wen, J.Y. Zhao, and S.J. Zhu, "Hierarchical clustering based on a bayesian harmony measure", *Pattern Recognition and Artificial Intelligence (PR&AI),* vol. 26, no. 12, pp. 1161-1168, 2013.

[12] Y. P. Wen, Z. G. Chen, and J. X. Liu, "Automatic grouping chinese customer addresses by clustering analysis", *Journal of Chinese Computer Systems*, vol. 34, no. 9, pp. 2060-2063, 2013.

[13] J.A. Nasir, V. Iraklis, A. Karim, and T. George, "Semantic smoothing for text clustering", *Knowledge-Based Systems,* vol. 54, pp. 216-229, 2013.

[14] J. Ma, "A staged and integrated semantic similarity algorithm of text", *New Technology of Library and Information Service,* vol. 29, no. 10, pp. 20-26, 2013.

[15] H. Tao, *Search of Group Intelligent Text Clustering Methods Based on Semantic Similarity*, Jiangsu University of Science and Technology, China, 2012.

[16] W.M.B.W. Mohd, and A.H. Beg, "An improved parameter less data clustering technique based on maximum distance of data and lloyd k-means algorithm", *Procedia Technology,* vol. 1, pp. 367-371, 2012 (http://www.sciencedirect.com/science/article/pii/S2212017312000771 - cor0005, and Tutut. Herawan).

[17] Y. Zhang, J. Liu, and H. Li, "An outlier detection method based on probability", *Computer Engineering,* vol. 39, no. 3, pp. 46- 55, 2013.

[18] W. Liu, and B. Qiu, "A short text modeling method combining semantic and statistical information", *Information Sciences,* vol. 15, pp. 4031-4041, 2010 (http://www.sciencedirect.com/ science/article/pii/ S002002 5510002823 - cor1)

[19] C. Luo, Y. Li, and S.M. Chung, "Text document clustering based on neighbors", *Data,* vol. 68, no. 11, pp. 1271-1288, 2009.

[20] D. Sánchez and M. Batet, "A semantic similarity method based on information content exploiting multiple ontologies", *Expert Systems with Applications,* vol. 40, no. 4, pp. 1393-1399, 2013 (http://www.sciencedirect.com/science/article/pii/S0957417412010 10X - cor1).

[21] H.J. Sun, G.H. Shan, and Y.L. Gao, "Algorithm for high-dimensional categorical data weighted subspace clustering", *Computer Engineering and Applications,* vol. 50, no. 23, pp. 131-135, 2014.

[22] M.N. Uddina, T.H. Duonga, and N.T. Nguyenb, "Semantic similarity measures for enhancing information retrieval in folksonomies", *Expert Systems with Applications,* vol. 40, no. 5, pp. 1645-1653, 2013.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014

© Junhong Ma; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/-licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.