A News Text Clustering Method Based on Similarity of Text Labels

Yuqiang Tong(✉) and Lize Gu

School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100091, China
ll72183723@qq.com, glzisc@bupt.edu.cn

Abstract. As an important text type, news texts have great research value in data mining, Such as hotspot tracking, public opinion analysis and other fields. News text clustering is a common method for studying the trend of news and hotspot tracking. Most of the existing clustering methods are based on the vector space model, with calculating the TF-IDF of words in the news text as feature items of the text. To improve the performance of clustering in the news texts, this paper presents a new clustering algorithm, this algorithm expresses the news text as a series of Text labels, which effectively solves the problem that the data latitude is too high, and the clusters is too hard to express. At the same time, by using a conceptual clustering algorithm, this method effectively reduces the number of comparisons. The experimental results show that the algorithm based on similarity of text labels improves the quality of clustering compared to traditional clustering methods.

Keywords: Data clustering · MinHash · Hierarchical clustering

1 Introduction

With the help of efficient information transmission on the Internet, it becomes faster and faster for a news from generation to public opinion. Through text clustering, news media and government departments can effectively find hot news and measure the development trend of news popularity. Therefore, clustering for news texts has become an important research topic in text clustering.

The traditional text clustering method is mainly to transform the text into a text vector model to deal with. By cutting the text into individual words, the doc information will be analyzed, TFIDF is a universal used text information [1], then comparing the similarity of the text vectors of the two articles and cluster the similar text. There are many ways to cluster text, such as K-means algorithm. Recently many text clustering algorithms based on K-means algorithms has been proposed, in paper [2] an algorithm based on the combination of nearest neighbor algorithm and K-means algorithm is proposed, which can lead to results with steady and high clustering quality. Except TFIDF, there are many other ways to measure the similarity. Hamming distance, Cosine Similarity and Euclidean Distance are common used.

Because the length of the news text is uncertain (a few hundred words to a thousand words), most of the texts after transformation vectors are high-dimensional sparse

S. Liu and G. Yang (Eds.): ADHIP 2018, LNICST 279, pp. 496–503, 2019.
https://doi.org/10.1007/978-3-030-19086-6_55

matrices, and the clustering effect of the traditional algorithms in high-dimensional space have a bad result.

In addition, in dealing with long texts, many targeted algorithms are proposed. Sim-Hash [3] is a common method to deal with it, which converts a text into a Hash Code by transforming the words after hashing, calculating the Distance between the Hash Code for clustering. The original purpose of Sim-Hash is to remove repeat texts from Massive texts. Since the text is converted into Hash Code, the amount of information is greatly compressed, so the effect of processing short texts (within 200 words) is not good enough. In paper [4] an algorithm based on Sim-Hash have a good performance in short message, but it is hard to process the news dataset with a large number of texts with different length.

In view of the shortcomings of the existing news text clustering methods, this paper proposes a news text clustering method by calculating the similarity of text labels. Compared with other methods, this method greatly reduces the dimensions of documents and speeds up clustering with a good clustering quality.

2 Related Theoretical Explanations

2.1 Document Dimension Reduction

The Jaccard Similarity. The Jaccard similarity between set A and set B is defined as the ratio of the intersection of A and B to the size of the union [5], as follow:

$$\text{Jaccard}(\text{A}, \text{B}) = |A \cap B| / |A \cup B| \tag{1}$$

K-shingle. A relatively common practice when comparing text similarity is to use the K-shingle set to represent a document [6]. Indicates the document, K-shingle is defined as all substrings of length K in the document. Suppose a document is a string: “abcde”, then the 2-shingle collection of the document is: “ab”, “bc”, “cd”, “de”. The similarity between two texts can be measured by calculating the Jaccard similarity of the K-shingle set of two texts. When selecting a K value, the following condition should be satisfied: Select a sufficiently large K to ensure that the probability of substrings in any given K-shingle set appearing in other sets is low enough.

The MinHash Algorithm. In computer science and data mining, the MinHash (or the min-wise independent permutations locality sensitive hashing scheme) is a technique for quickly estimating how similar two sets are. The scheme was invented by Broder [7], and initially used in the AltaVista search engine to detect duplicate web pages and eliminate them from search results [8]. It has also been applied in large-scale clustering problems, such as clustering documents by the similarity of their sets of words [7].

Guess a minimum hash, which is a hash vector of a text, and each bit of the vector indicates whether a certain K-string of K-shingle corresponding to the bit appears in the text. Since this minimum hash is a sparse 0, 1 vector. Therefore, the MinHash

corresponding to the minimum hash is the number of the first k nonzero bit of the smallest hash, as follow:

$$min_{h,k}(v) = \arg\min_k \{h(v[i])\} \tag{2}$$

where v is a K-shingle vector, $h(v[i])$ is a 0, 1 vector of the K-shingle vector.

An important conclusion of MinHash is that the MinHash equal probability of two sets is equal to the Jaccard similarity between two sets. For a given set A and B, the Jaccard similarity of the two sets can be expressed as follow:

$$\text{Jaccard}(\text{A}, \text{B}) = \text{P}\left(min_{h,k}(v_A) = min_{h,k}(v_B)\right) \tag{3}$$

As the conclude of the MinHash algorithm, calculating the Jaccard similarity of two documents can be replaced to calculate the equal probability of the MinHash vector, so it is a good choice to dimensionality reduce by converting documents to MinHash vector.

2.2 The Clustering Algorithm

In general, clustering methods include hierarchical clustering and partitioning [9].

The most commonly used partitioning method is K-means clustering and related deformation methods [10]. By selecting a clustering center, the clustering center of the class and the set of classes are updated via iterative rules until the clustering center no longer changes.

There are two Shortcomings of partitioning in news texts clustering. First, it is hard to determine the number of clustering center, because to estimate how many news events will be reported each day is a difficult thing. Besides, the partitioning method such as K-means must calculate the distance between all the samples and the cluster center at once iterative, it will cost too much time.

Two types of hierarchical clustering are commonly used: Agglomerative and Divisive [9]. Agglomerative means that each the original sample has their own class, from the bottom to the up, according to similarity rules, merge different classes to reduce the number of classes until the stop condition is met. Divisive means all the samples share the same class, and from the top to the bottom, by using the similarity rules, splitting the classes until the stop condition is met.

Commonly used hierarchical clustering methods are Cobweb [11], CLUSETR/2 [12], and UPGMA [13].

Based on the above reasons, this paper adopts the idea of hierarchical clustering to implement news text clustering. When text clustering, different kinds of labels can be used to achieve clustering at different levels.

3 Proposed Labels Based Clustering Algorithm

3.1 News Text Labels Generate

Considering the influence of different paragraphs or sentence weights at different positions and the title influence in the news text, there are three kinds of labels to summarize, title label (generated by the title text), sentence label (generated by the start sentences of all the paragraph), text label (generated by the main text of the news).

Through the Eq. (2), the title label, sentence label, and text label can be expressed by the follow:

$$min_{h,k}(v_{title}) = \arg\min_k\{h(v_{title}[i])\} \tag{4}$$

$$min_{h,k}(v_{sentence}) = \arg\min_k\{h(v_{sentence}[i])\} \tag{5}$$

$$min_{h,k}(v_{text}) = \arg\min_k\{h(v_{text}[i])\} \tag{6}$$

For the text, it is much longer than the title or the sentences of the news, only using a MinHash vector to indicate the contents of the main text may cost a high bias. To avoid the problem, there will give several of different MinHash to group a set to indicate the text, as follow:

$$min_{H,k}(v_{text}) = \left\{min_{h_1,k}(v_{text}), min_{h_2,k}(v_{text}), \ldots min_{h_t,k}(v_{text})\right\} \tag{7}$$

Where H is a set of different kinds of MinHash, and $h_1, h_2, \ldots h_t \in H$.

According to the generate method of the labels, a label set of the news text can be expressed by the follow:

$$S_{H,k}(Doc) = \left\{min_{h,k}(v_{title}), min_{h,k}(v_{sentence}), min_{H,k}(v_{text})\right\} \tag{8}$$

Where $S_{H,k}(Doc)$ is the label set of a news text, $min_{h,k}(v_{title})$ is the label of title, $min_{h,k}(v_{sentence})$ is the label of sentence, $min_{H,k}(v_{text})$ is the label set of text. As for the size of the news text label set, guess the average text length is m, through the MinHash algorithm, it became to $(2 + \mathrm{t})\mathrm{k}$, which is small the original length.

By the conclusion of the MinHash algorithm, the similarity of two labels can be calculated by the follow:

Considering each bit in the label value, the probability p of two bits have the same number same to the Jaccard similarity of the two texts. Therefor for the K bit in the label value, the similarity can be expressed as the same number of bits divided by K:

$$\mathrm{Dist}(d_1, d_2) = \sum\nolimits_{i=1}^{k} x_i \oplus x_i \tag{9}$$

Where x_i and y_i are the values on the i th bit of the label d_1, d_2.

3.2 The Hierarchical Clustering of News Text

In the label generation method, the label set of each news text contains a title label, a sentence label, and t text labels (8). When clustering, different labels can be used to achieve clustering at different levels.

First consider the title label. For two title label, each on has a length K MinHash vector, by the similarity algorithm of labels (9), for two texts with Jaccard similarity p, the expectation of the same number of bits in the two title labels is $p \times k$, so it is an idea to cluster two texts into same group which have a same number of bits larger than $p \times k$.

The sentence label can be used same with the title label, through the hierarchical clustering by using the title label and the sentence label, the news texts which have a similarity title and similarity sentence.

For each group, the final step is clustering the texts by using text label set. Consider any two texts, each of them has t different text labels. If their Jaccard similarity is p, the probability of t text label pairs at least has one same label pair is follow:

$$\mathrm{P}(label1, label2) = 1 - \left(1 - p^k\right)^t \tag{10}$$

Where p is the Jaccard similarity of two labels, k is the length of MinHash vector, t is the length of text label set. The Table 1. shows the $\mathrm{P}(label1, label2)$ in different Jaccard similarity in the k = 5, t = 20 condition.

Table 1. The table of probability of different Jaccard similarity.

Jaccard similarity	P(label1, label2)
0.2	0.006
0.3	0.047
0.4	0.186
0.5	0.470
0.6	0.802
0.7	0.975
0.8	0.9996

Based on the above conclusion, when hierarchically clustering through the text labels, compare t pairs of text label for any two texts, if there is at least one pair of identical label pair which has a same label, grouping them into a same class.

4 Experiments of Label-Based Clustering

4.1 News Texts Datasets

By reptiles from more than 50 news websites, the datasets totally including more than 300,000 news items from October 2017 to November 2017. Each news item contains news URL, news authors, authoring date, publishing sources, title, and text.

4.2 Experimental Environment

The experimental environment is a distributed Hadoop computing platform. The platform contains a single Namenode server three Datanode server, the configuration of each node server is an Intel Core i7-7760 dual-core processor, 1 GB memory, and 1000 GB hard disk capacity. The Hadoop configuration using the default configuration.

4.3 Evaluation Standards for Experiments

This experiment compares the clustering effects of different algorithms. The evaluation standards contain the average clustering quality, the overall quality, and the time-consuming of the clustering.

The average clustering quality means the average number of quality of each class. The quality of one class is defined by the ratio of all samples in the class that same to the class center to the total number of samples in the class in which the class center means the mode news of the class.

The overall quality means the ratio of the number of samples have the right class to the number of all samples.

The Table 2 shows the average quality and the overall quality of the TF-IDF algorithm, SimHash algorithm, and the Label-based algorithm.

Table 2. Table of the comparison of 3 kinds of algorithm in clustering quality

Algorithm name	Average quality (%)	Overall quality (%)
Label-based	86.1	88.9
TF-IDF	64.7	67.7
SimHash	74.8	79.2

The Table 3 shows the time cost of clustering of 3 kinds of clustering algorithms.

Table 3. Table of the comparison of 3 kinds of algorithm in time cost

Algorithm name	Time consuming
Label-based	88.9
TF-IDF	67.7
SimHash	79.2

From the above experimental data, the overall quality and the average quality of the label-based algorithm is better than the SimHash algorithm and TF-IDF algorithm, From the time-consuming point of view, the TF-IDF has a better performance. The reason is that the time to generate the Label by using the MinHash will cost a long time. However, this problem can be solved by improving the parallel processing capability of the computing environment.

5 Summary

This paper proposes a new clustering method of news texts clustering method. This method is based on the characteristics of news texts, adopts the MinHash method, and proposes a news text labeling model. Through the labels of the news texts, this paper completes the clustering process based on the hierarchical clustering. Experiments show that the algorithm is of higher quality than traditional methods.

The next step of research:

1. Optimize the label set, and further improve the speed and efficiency of the algorithm.
2. Combining specific scenarios, applying the algorithm to web data mining tasks, such as news hotspot discovery, news public opinion analysis, and so on.

Acknowledgements. The authors wish to thank the Information Security Center at Beijing University of Posts and Telecommunications for providing the distributed computing environment. The authors acknowledge the financial support of The National Science and Technology Major Project (Grant No. 2017YFB080301).

References

1. Wu, H.C., Luk, R.W.P., Wong, K.F., et al.: Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inf. Syst. **26**(3), 55–59 (2008)
2. Zhang, W.M., Jiang, W.U., Yuan, X.J.: K-means text clustering algorithm based on density and nearest neighbor. J. Comput. Appl. **30**(7), 1933–1935 (2010)
3. Sadowski, C., Levi, G.: SimHash: hash-based similarity detection (2007)
4. Pi, B., Fu, S., Wang, W., et al.: SimHash-based effective and efficient detecting of near-duplicate short messages. In: Proceedings of International Symposium on Computerence & Computational Technology, vol. 4, pp. 20–25 (2014)
5. Real, R., Vargas, J.M.: The probabilistic basis of Jaccard's index of similarity. Syst. Biol. **45**(3), 380–385 (1996)
6. Manaa, M.E., Abdulameer, G.: Web Documents Similarity using K-Shingle tokens and MinHash technique. J. Eng. Appl. Sci. **13**, 1499–1505 (2018)
7. Broder, A.: On the resemblance and containment of documents, pp. 21–29 (1997)
8. Broder, A.Z., Charikar, M., Frieze, A.M., et al.: Min-wise independent permutations. J. Comput. Syst. Sci. **60**(3), 630–659 (2000)
9. Luxburg, U.: A Tutorial on Spectral Clustering. Kluwer Academic Publishers, Hingham (2007)
10. Hartigan, J.A.: A K-means clustering algorithm. Appl. Stat. **28**(1), 100–108 (1979)

11. Douglas, H.F.: Knowledge acquisition via incremental conceptual clustering. Mach. Learn. **2**(2), 139–172 (1987)
12. Michalski, R.S., Stepp, R.E.: Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **5**(4), 396–410 (1983)
13. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A K-means clustering algorithm. J. Roy. Stat. Soc. **28**(1), 100–108 (1979)