Short Text Clustering Algorithm Based on Frequent Closed Word Sets

Chunxia Jin and Qiuchan Bai

July 28, 2019

Short Text Clustering Algorithm Based on Frequent Closed Word Sets

Chunxia Jin
Faculty of Computer and Software Engineering
Huaiyin Institute of Technology
Huaian, China
jcxbzn@163.com

Qiuchan Bai
Faculty of Automation
Huaiyin Institute of Technology
Huaian, China
bqcbzn@163.com

***Abstract*—The text mining of microblog topic information can effectively obtain the attention degree of internet users for news events. It is of great significance in the field of public opinion monitoring and analysis. At the situation of the algorithm of traditional frequent word set is suitable for long text information clustering, this paper proposes to mine top-K frequent corpus in short text database and then to divide microblog topic texts covering the same frequent word sets into the same cluster. Combined with the largest frequent word-sets for similarity calculation, the overlapped document is re-divided to achieve microblog short text clustering. The experimental results of microblog topic dataset and the comparison with K-means clustering algorithm show that the proposed algorithm can effectively solve the sparseness and high-dimension problem of microblog topic short text clustering and greatly improve the microblog short text clustering effect.**

Keywords—frequent closed word sets, microblog topic, text clustering

I. Introduction

In rapid development period of information technology, social networking applications have replaced traditional media as the typical representative by microblog. It has occupied the dominant position of information dissemination. Microblog is easy to use, convenient and so on. It can fascinate a large number of internet user, the resulting online data has showed an explosive trend of growth [1]. Through the microblog platform, a message can be broadcasted in a short time and affect millions of users. Compared with traditional media, the microblog of dissemination time and breadth in the process of information dissemination have greatly enhanced. However, the dissemination of some false information and socially-emergent topics through social networking platforms such as microblog can also cause social panic and user property losses in a short period of time and have a huge social impact. Therefore, the detection technology of microblog topic has positive significance for the discovery of social hot topics, perception of internet users, public opinion detection and emergency response [2]. Microblog topics often display personal updates, topic sharing as 140 words or less text messages, pictures, audio and video and other multimedia content. These data are time-sensitive, sparse, singular, and redundant [3,4], many valuable information has been obscured in a large amount of redundant information. Microblog topic detection can not only filter invalid information, improve the quality of content, and improve the user experience, but also play the role of monitoring, public opinion control and viewpoint mining.

Traditional text mining algorithms generally use a fixed number of text features and feature weights. Microblog topic has its own unique characteristics. It is often a dynamic text. Its timing not only brings new revelations to microblog text mining, but also brings new challenges.

Frequent itemsets are the basic concepts in association rules mining. The frequent sets obtained in text mining are the frequent word feature sets. Since frequent word sets reflect the correlation between words, they contain more implicit semantic and contextual information than independent words. Thus frequent itemsets have been used in text mining for many applications [5-8]. Traditional frequent pattern mining algorithms usually require the user to specify a minimum support threshold. However, if the threshold is too large, sufficient frequent patterns may not be obtained. If the threshold is too small, a large number of redundant frequent patterns will be obtained, which is unfavorable for finding useful knowledge. Top-K frequent mode only requires the user to specify the K-number of patterns to be mined, the algorithm will automatically return to the most frequent K-patterns[9,10]. So it is a better solution. However frequent word-sets mining will produce the large number of word sets, and a exponential subset may be generated if the long pattern is mined. Therefore, this paper proposes a short word mining algorithm using frequent closed word-sets in order to reduce the sparseness and high dimensionality of the topic keywords of microblog short texts.

II. The Related Definitions of Frequent Word Sets

Definition 1: Frequent itemsets Given datasets D and itemsets X on D, $\min\_sup \epsilon (0,1)$, which is said as minmum support. If $\mathrm{sup(X)} = |D_x|/|D|$, sup(X) is regarded as the support of itemset X. When $\mathrm{sup(X)} \geq \min\_\mathrm{sup}$, the itemset X is said as a frequent itemset on D.

Definition 2: The maximum frequent itemset Given datasets D and itemsets X on D, if $\mathrm{sup(X)} \geq \min\_\mathrm{sup}$ and when it meets $\mathrm{sup(Y)} < \min\_\mathrm{sup}$ for $\forall(\mathrm{Y} \subseteq \mathrm{I} \wedge \mathrm{X} \subseteq \mathrm{Y})$, the itemset X is regarded as the maximum frequent itemset on D.

Definition 3: Closed frequent itemset Given datasets D and itemsets X on D, if $\mathrm{sup(X)} \geq \min\_\mathrm{sup}$ and when it meets $\mathrm{sup(Y)} < \min\_\mathrm{sup}$ for $\forall(\mathrm{Y} \subseteq \mathrm{I} \wedge \mathrm{sup(X)} = \mathrm{sup(Y)})$, the itemset X is regarded as the closed frequent itemset on D.

Definition 4: Boundary support After all frequent patterns that have been mined are sorted in descending order by support, if θ is exactly the support of the K-pattern, so it is called the boundary support (border_sup).

Definition 5: Top-K frequent word-sets Let Y be a frequent word-set, if the number of all frequent word-sets that those support degree is greater or equal to Sup(Y) is not exceed K, Y is the Top-K frequent word-set. If Y is not less than min_length that is the minnum length, then Y is called Top-K frequent word-set with minimum length.

Fund project: Research and application of intelligent city data sharing in jiangsu provincial construction system science and technology project.

III. Microblog Topic Short Text Clustering Algorithm Based on Frequent Closed Word Sets

This paper proposes short text mining of microblog topic by frequent closed word sets. Firstly, the algorithm uses short text frequent itemset to express each microblog text as a frequent word set. Then microblog documents that cover the same frequent word sets are grouped into the same cluster. And these overlapping documents belonging to multiple clusters are sorted by word frequency. Finally, combined with the similarity calculation based on the largest frequent word set, these overlapped documents are re-divided, then realizes microblog short text clustering.

The algorithm based on frequent closed word sets is done on the following three steps.

- Mining out all the frequent closed word sets.
- Each frequent closed word set is initialized and partitioned in order to create a cluster that contains all the documents covered by this frequent closed word set.
- The overlapped documents are divided into the most suitable clusters by adjusting the cluster and calculating the similarity.

A. The mining algorithm of Top-K frequent closed word set based on FP-growth

Top-K frequent word sets are mined using frequent word sets to represent text. In mining, Top-K frequent word set do not have to frequently try min_sup, merely transforms the work of mining frequent word sets into the work of mining K-most frequent word sets. So just set the K value to get the K-number of most frequent patterns you want. In the mining process, we may also throw those patterns of its length less than min_length, and use closed frequent word sets to replace frequent word sets, so it can avoid the problem that mining the number of words are too large.

The algorithm: The mining algorithm of Top-K frequent closed word set based on FP-growth.

Input: Business database D, the K-number of frequent word sets, the length of shortest corpus.

Output: A set of frequent closed word sets that satisfy the conditions.

Steps:

1) Initialize and set the min_sup of minimum support as 0.

2) Scan the transaction database of D, count each of the items in the transaction and these items are sort by decreasing order of frequency. These form a list of first frequent items ordered and FP-tree header node.

3) Rescanning database and updating the count of FP-tree header node, then using the count array of closed node to increase the min_sup of support and using this support to prune the FP-tree.

4) If first frequent item meets $|C_1| < K$, then set the $border_{sup}$of border support as 0.

5) Mining FP-tree from bottom up according to the table of header node, the condition FP-tree is constructed from the first node whose count is not 0 until the count is larger than the current min_up. Then each closed pattern that has been mined will be inserted into the pattern tree.

6) Output each pattern of the pattern tree until the number of patterns reaches to K.

Since the algorithm only needs to traverse the transaction database twice by constructing FP-tree, the mining efficiency of the algorithm is greatly improved.

B. Text clustering based on frequent closed word sets

In microblog corpus, the set of all documents is defined as $\mathrm{D} = \{T_1, T_2, \cdots, T_n\}$ the set of words in these documents is defined as $\mathrm{I} = \{i_1, i_2, \cdots, i_m\}$, the set of all frequent closed itemsets mined from all documents is defined as $\mathrm{MS} = \{M_1, M_2, \cdots, M_n\}$. The document set covered by a frequent closed itemset of M is referred to as P, which meets $\mathrm{P} \subseteq \mathrm{D}$.

The clustering process is that the document set is divided into several cluster sets as $\mathrm{CS} = \{C_1, C_2, \cdots, C_I\}$. The set of documents contained in a cluster is CP, which meets $\mathrm{CP} \subseteq \mathrm{D}$. The frequent closed itemset contained in the cluster is CM, which meets $\mathrm{CM} \subseteq \mathrm{MS}$. And the set of frequent items contained in a cluster is referred to as CI, which meets $\mathrm{CI} \subseteq \mathrm{I}$. The set of frequent items contained in a cluster is the union of the frequent items contained in all frequent closed itemsets.

The cluster algorithm based on frequent closed itemsets is as follows:

Input: The set of MS satisfying the Top-K closed frequent itemsets of min_length.

Output: The set of CS.

Steps:

1) Frequent itemsets are sorted by length in MS.

2) Select M_i of frequent closed itemset in proper order.

3) If it meets $CP_i \subseteq D_c$, go to step2 to select the next frequent itemset, otherwise generate cluster C_i.

4) Update $D_c = D_c \cup CP_i$ of document set covered cluster, go to step2.

5) When all frequent closed itemsets have been processed, a collection of clusters has been generated.

C. The similarity document reprocessing using the largest frequent word sets

After the text clustering is implemented, the document classification covered by the frequent word set can be easily obtained from large-scale microblog information. However, for the overlapping part (multiple intersection), when searching a specified keyword, the document containing the keyword does not have a clear division. That is it cannot be determined which cluster the document belongs to, so resulting query results is inaccurate. Therefore, the document needs to be reprocessed in an appropriate way in order to the overlapping documents are divided into the most suitable clusters.

In this paper, the most frequent word sets are used to calculate similar documents, the document will be maximally divided and it enhances the clustering effect of microblog short texts. The algorithm idea is that firstly the maximum frequent word sets will be sorted in word frequency order, and selects these frequent word sets with the same number of word sets as feature vectors to do similar calculation in the current cluster. Assumed the document D_i is divided into two clusters of C_m and C_n, which C_m has frequent word sets as $Y_m = \{w_2, w_3\}$, C_n has frequent word sets as $Y_n = \{w_1, w_2\}$, D_i has the maximum frequent word sets as $Y_i = \{w_1, w_2, w_3\}$. In the process of calculation, we can get the word frequency order of the largest frequent word sets in D_i. Assuming that the sorting result is w_1, w_3, w_2, w_1, w_3 are used as feature vector to do similarity calculation according to the number of frequent word sets in the cluster that is contained. Then the similarity between feature vectors and two clusters is calculated respectively on the base of vector space model and feature vectors of frequent word set. Finally, the document of D_i is classified into appropriate clusters according to the similarity.

The similarity of two documents D_1, D_2 is calculated as follows:

$$\mathrm{Sim}(D_1, D_2) = \frac{\sum_{k=1}^{n} W_{Ik} \times W_{2k}}{\sqrt{(\sum_{k=1}^{n} W_{1k}^2)(\sum_{k=1}^{n} W_{2k}^2)}} \qquad (1)$$

In which W_{Ik} and W_{2k}respectively represent the weights of K-th features of documents D_1 and D_2 , and k meets $1 \le k \le n$.

IV. The Experimental Results and Data Analysis

A. The evaluation

This paper uses Precision and Recall to test the quality of clustering algorithm. Precision rate reflects the clustering degree of similar text and no similar text merged into the same class. The more the precision of clustering is higher, thte more the content of each class is concentrated. Recall reflects the clustering degree of all same topic texts merged into the same class. The more the Recall rate of clustering is higher, the more similar text is concentrated too. The clustering precision reflects the distinction between different themes text, but the recall rate of clustering reflects the ability of identifying the same theme text.

Precision, Recall of a clustering are defined as follows:

$$P = \frac{|TC|}{|TC| + |FC|} \times 100\% \qquad (2)$$

$$R = \frac{|TC|}{|MC|} \times 100\% \qquad (3)$$

Where TC represents the cluster that are consistent with manual labeling in clustering algorithm, and FC represents the cluster that are inconsistent with manual labeling. MC represents artificially annotated clusters.

B. Data analysis of experimental results

In this paper, we compared the clustering algorithm based on frequent corpus sets with K-means clustering algorithm. In K-means algorithm, the parameter K is set as twelve in advance according to the selection of the data set. In the clustering algorithm based on frequent word sets, we first need predict the choice of parameters. Experiments show that when the K of the number of frequent closed itemsets is set as 45 and the minimum length of frequent closed itemsets is five, the experimental results will be more accurate and clustering effect will be better.

For example, three clustering results are cited the following, the experimental results are shown in Table 1. Three cluster topics are respectively microblog discussion of house price, smog and civil servant.

TABLE I. Clustering results of two algorithms

cluster topic	Clustering based on frequent word sets	K-means
house price	house price, space, property market, money, fall, rise, situation, developer, trend, government, worry, people, market	property market, developer, trend, attention, dollar, house price, rise, lower, fall complaint, crash
smog	building, smog, wind, air, capital, dust, America, consistence, Beijing, mask, weather	space, situation, wind and sand, blow, environmental protection, dust, people, government, wind, three hundred thousand
civil servant	establishment, civil servant, rise, office, three hundred thousand, hour, China, income, person, test, officer, eliminate	China, villages and towns, office, civil servant, salary, establishment, pressure, two hundred thousand

From Table 1 can be seen that the clustering algorithm based on frequent word sets has a higher accuracy and is semantically easier to understand. However, the cross-talk between K-means clustering results is very serious. That is multiple topics are confused as a topic, the clustering result is very unsatisfactory. Using the accuracy and recall rate to evaluate the performance of clustering results, the comparison results of two algorithms are shown in Table 2.

TABLE II. The comparison results of two algorithms

Performance	The clustering algorithm based on Top-K frequent closed word sets	The clustering algorithm of K-means
Accuracy	78.23%	33.67%
Recall rate	67.62%	32.28%

From Table 2 can be seen that the accuracy and recall of clustering algorithms based on frequent word sets are significantly higher than K-means clustering algorithm. In K-means clustering algorithm, the algorithm uses feature vector representing text, it is only statistical information of feature word frequency, which lacks textual semantic information. Moreover, the feature vectors of microblog short texts are very sparse, which leads to the clustering result of K-means algorithm is not very satisfactory. However, in clustering algorithm based on frequent word sets, all documents are divided into different clustering according to frequent word sets. Those documents with same frequent closed word sets will be divided into a clustering. And the algorithm uses frequent word sets describing this clustering information, which can well express text semantic information. The algorithm can greatly improve the clustering effect.

V. Conclusion

This paper proposes to the algorithm of microblog short text clustering based on frequent closed word set. The algorithm mines Top-K frequent closed word sets of length which is not less than min_length. Using these frequent closed word sets represent feature vectors of microblog short text in a clustering algorithm, it can greatly reduce the dimensions of vectors. Secondly, microblog documents with same frequent word sets are classified into the same cluster.

And these overlapping documents belonging to multiple clusters are sorted by word frequency. Finally, combined with the similarity calculation based on the largest frequent word set, these overlapped documents are re-divided. The method is helpful for clustering the documents of the same topic into the same category and improving the accuracy of the microblog short text clustering algorithm. On the basis of the research, the future work will optimize the performance of the algorithm, such as improving the frequent closed itemsets mining algorithm and considering more semantic information in the similarity calculation of maximum frequent word set.

ACKNOWLEDGMENT

This work is supported by Jiangsu provincial construction system science and technology project (2018ZD269).

REFERENCES

[1] Y. Zhen, L. Wang, Y.X. Lai, "Online comment clustering based on an improved semantic distance," Journal of Software, vol. 25, pp. 2777-2789, Decmber 2014.

[2] Y.F. Qiu, L.Y. Wang, L.S. Shao, "User internet modeling approach based on short text of microblog," Computer Engineering, vol. 40, pp. 275-279, February 2014.

[3] M. Yuan, Y.X. Yang, X. Zhang, "Short text feature extension method based on frequent term sets," Journal of Southeast University, vol. 44, pp. 256-260, February 2014.

[4] Qin Q L, Yang X, Gu H, "Microblog users roles in topic diffusion based on individual attribute features," International Journal of Hybrid Information Technology, vol. 9, pp. 447-460, April 2016,.

[5] Ye Y, Du Y, Fu X, "Hot topic extraction based on Chinese Microblog's features topic model," IEEE International Conference on Cloud Computing and Big Data Analysis. IEEE, 2016, pp. 348-353.

[6] H.F. Ma, X.T. Zeng, X.H. Li, "Short text feature extension method of improved frequent term set," Computer Engineering, vol. 42, pp. 213-218, October 2016,.

[7] H.J. Huang, J.S. Tan, J.H. Qin, "The topic detection algorithm based on microblog model," Chinese Journal of Network and Information Security, vol. 2, pp. 30-38, May 2016.

[8] X.S. Zhang, C.Y. Jia, "A new documents clustering method based on frequent itemsets," Journal of Computer Research and Development, vol. 55, pp. 102-112, January 2018.

[9] Barouni Ebrahimi, Alireza, "Measuring productive collocational knowledge of the most frequent words," International Journal of Applied Linguistics, vol. 29, pp. 30-43, March 2019.

[10] Kozlowski, Marek; Rybinski, Henryk, "Word sense induction with closed frequent termsets," Computational Intelligence, vol. 33, pp. 335-367, August 2017.