News Events Clustering Method Based on Staging Incremental Single-Pass Technique

LI Yongyi [1,a] *, Gao Yin [2]

1 School of Electronics and Information Engineering
QinZhou University
535099 Guangxi, China
2 School of Computer Science and Engineering
South China University of Technology
510640 Guangzhou, China
a* corresponding author, e-mail:34268371@qq.com

***Abstract* — According to the characteristics of news events, the clustering classification method is suitable such news events. However, traditional Single-Pass methods have two major drawbacks: low-level time efficiency and accuracy. In this paper, we focus on the clustering characteristics of news events, and propose a new method based on staging incremental Single-Pass. Firstly, the Vector Space Model (VSM) is used to represent the text vectors, and regression analysis is performed by using the Term Frequency and Inverse Documentation Frequency TFIDF weighting equation to establish the weighted calculation model. Meanwhile, the centroid vectors are used to represent news events models. Secondly, segmented lists are applied to news events models. Lastly, the cosine values of the angles are used to calculate the levels of similarity, and a similarity-based approach is used to realize the clustering of news events, and to selectively and dynamically update the segmented lists and centroid vectors so that the timeliness and accuracy of clustering could be significantly improved. Empirical experiments have demonstrated that the proposed method is effective, and could be used to improve clustering accuracy and reduce event complexity.**

Keywords -- news events; text classification; segmented; incremental clustering

I. INTRODUCTION

We live in the information age where science and technology have been well-developed, with networks functioning as the major platform for information dissemination. Networks are manifested as a wide range of dynamic information that flows in a variety of forms and with rapid updating speed. Text represents a major expression form of information dissemination due to its simplicity and effectiveness. However, huge amount of text information has imposed increasing pressure on networks, and the content of such text information tends to become more complex over time. Pouring news events, as well as highly-focused topics, could spawn over networks within a short timeframe. In order to effectively and timely capture and analyze breaking news or headline news from complicated network information, the monitoring and analysis of network public opinions in a timely manner becomes a new research topic[1]. Obviously, how to effectively identify the text information that satisfies user demands from a huge and complicated volume of text information is the basis of the monitoring and analysis of network public opinions. Before trend analysis is applied to these breaking news or headline news, all relevant news events should be categorized, and such process is also known as text mining. Text clustering method could be used to automatically categorize the text information, and thus represents a key way to solve the problem as demonstrated in the following:

Text categorization is an effective way of processing massive text information. It could help provide good structures of text set, greatly simplify the access and processing of text information, and improve the processing efficiency of text information.

The search results can be effectively organized. The method of text categorization can be used to reasonably organize the search results by grouping them in accordance to the level of similarity between each other. Each group should have a clear topic, and thus users could rapidly scan over all groups and select those groups that are most relevant to their objectives.

The search process could be accelerated. It is possible to categorize the raw text in advance, and extract implied concepts from large-scale text sets, which could then be represented by projections in concept spaces. The texts with high level of similarity are categorized as the same group, with one center established for each group. All centers are compared to one another in order to dramatically accelerate the search process.

Text clustering methods are categorized into five classes, namely, partitioning methods, level-based methods, density-based methods, grid-based methods and model-based methods. Partitioning methods firstly set a data base with n objects or elements. A partitioning method establishes k partitions of data, with each partition representing a cluster. Then a partitioning criterion (such as similarity) is selected to complete the process text partitioning, with time complexity as O(nk). Level-based methods perform

hierarchical decomposition of the data object sets. The drawbacks of level-based methods lie in that the convergence performance is unsatisfactory, with time complexity as O(n2) [2]. Density-based methods mainly deal with arbitrary shaped clustering problems. For these methods, however, it is unable to reach convergence once the density of the area adjacent to the set exceeds a threshold. These methods are mainly used for the clustering of outlier data. Model-based methods assume a model for each cluster, and then find the best fitting for the defined models. These methods mainly deal with the clustering of spatial distributions. Grid-based methods adopt a multi-resolution grid data structure to vectorize object spaces into units with limited numbers so that a grid structure could be formed. The main advantage of these methods is that their processing speed is very fast, which is independent of the number of data objects. Instead, it depends on the number of units for every dimensional vector space.

Since news events are dynamic information flows that arrive chronologically, density-based, level-based and grid-based methods could not categorize the coming news events according to relevant topics in a timely manner. By contrast, only the partitioning methods are suitable for handling such daunting tasks. Since the essence of the clustering method of news events is based on the level of text similarity, the choice of Single-Pass, a commonly applied method, is reasonable. Timeliness and accuracy represent key elements for the clustering of news events.

First, the timeliness could be improved by two ways. (1) The improvement of the process of data acquisition and preprocessing related to news events. However, the performance depends on devices and data structures. (2) The improvement of the speed of comparing texts and clusters. Obviously, how to determine cluster models more efficiently is the key. Numerous researchers have developed incremental algorithms. For example, Guha et al. proposed a local clustering method [3]. Gupta et al. suggested partitioning a data flow into different data blocks, and completing the clustering process by assessing cluster centers. However, these algorithms depend on multiple times of chronological clustering, but could not increase clustering efficiency fundamentally [4].

Second, the characteristic items of clusters are the main factors that affect clustering accuracy of news events. Ordinary incremental algorithms use the average value or vector superimposing method. However, long-term streaming data would reduce the accuracy of such cluster models, and thus affect clustering performance. In the present study, we theoretically proposed ways to improve the timeliness and accuracy of Single-Pass algorithm, and also carry out experimental assessments.

II. TEXT CLUSTERING

The basic principle of text clustering is to partition the text set into several classes through calculation. Also, texts from the same category should be as similar as possible, whereas texts from different classes should be as different as possible. Text clustering has been widely applied to a variety range of fields, such as data mining, information retrieval and theme detection.

A. Single-Pass algorithm.

The Single-Pass algorithm is also known as single-channel algorithm, or single run algorithm. It is a classical heuristic method for data stream clustering. For data streams that arrive sequentially, the first text data stream is used as the clustering basis, and the following data are processed according to the input order. Meanwhile, the rest text data are compared with the first datum in an orderly manner. If the level of similarity meets the defined requirement, the latter datum is categorized into the same class as the first one. Otherwise, the current datum is used to create a new data cluster according to the matching degree between this datum and existing clusters. Meanwhile, this text is used as the clustering basis of the new cluster. All texts are clustered according to this method [5], with corresponding clustering flowchart shown in Figure. 1. The computation complexity of Single-Pass clustering is O(nk), where k is the number of clusters.

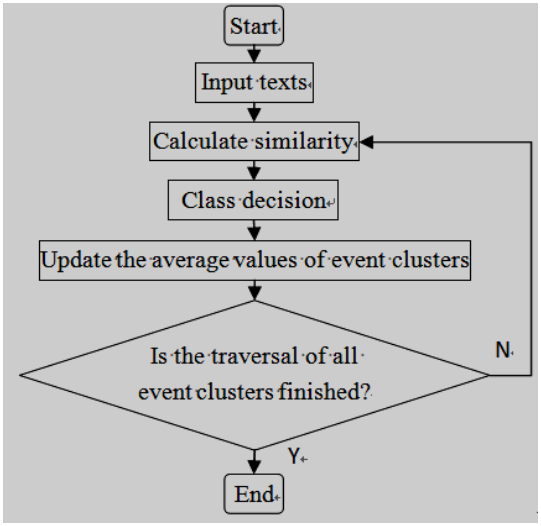

Figure. 1 The clustering process of the general Single-Pass algorithm.

The commonly used improved Single-Pass algorithms are incremental algorithms, which are suitable for various data stream mining tasks but with low level of efficiency. The two drawbacks of these algorithms lie in that the methods depend on the orderliness of input, and the update of clustering algorithms might result in low clustering accuracy. Therefore, when general incremental algorithms are applied to the clustering of news events with massive amount of information, the two drawbacks could become more prominent.

III. NEWS EVENTS CLUSTERING METHOD BASED ON STAGING INCREMENTAL SINGLE-PASS

During the clustering process of news events, the Single-Pass method, represented by a vector space model (VSM), is still widely adopted. To improve the timeliness and accuracy of general incremental Single-Pass algorithm, staged

increment is a key element. Staging means that during the clustering process, text vectors are put into lists with fixed length, and this staging process is independent on the orderliness of arriving news events. Therefore, paralleling processing of information as a whole could be realized, and the timeliness could be improved. Increment is based on the staging process, and the feature vectors of clusters are updated locally and dynamically. Under the situation that the total length of the vectors remains unchanged, the commonly used average values or vector superimposing method are altered in order to improve the accuracy of clustering feature vectors.

First, a text vector is acquired, and its weight is calculated. According to the weight, the angle cosine equation is adopted to calculate the level of similarity (T) between the text vector and the existing cluster. Then the similarity is compared with a threshold Tc. If T is within the threshold range, the text vector is compared with the next existing cluster. A staged list P with fixed length is set for each cluster. If no cluster is assigned to the text vector after the k-th cluster is compared, then the (k+1)-th event cluster is created. Otherwise, the text vector is added into the existing cluster. The staged lists play an important role during the clustering process. On the one hand, they temporarily save input text vectors. On the other hand, the cluster feature vector Z is updated according to the lists after the clustering process. The flowchart of news events clustering method based on staging incremental Single-Pass is shown in Figure. 2.

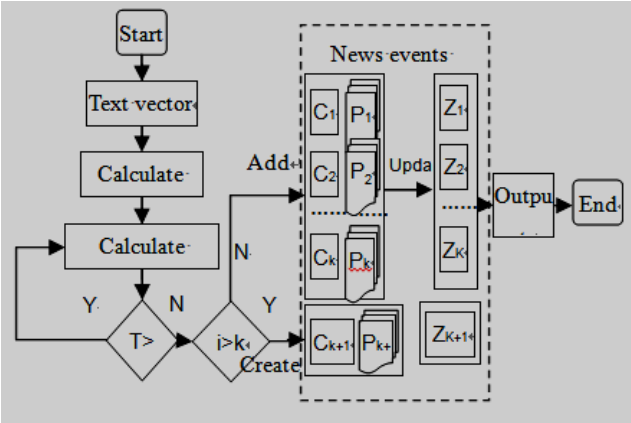

Figure. 2 The clustering process of news events clustering method based on subsection incremental Single-Pass.

A. Basic Model.

1) Text vectors

In this study, VSM is used to represent news texts [6], or, in other words, a piece of news text, d , could be expressed as follows:

$$d = (t_1, w_1, t_2, w_2, \ldots, t_n, w_n) \qquad (1)$$

where t1, t2, …, tn are feature terms representing text contents, i.e., the word segments of the text. w1，w2，…，wn are the weights of t1, t2, …, tn, respectively. Also, feature words are used to represent feature terms in the current study.

2) Weight calculation

The weights of feature terms are calculated according to TFIDF. The weight of word segment i in document k is calculated using the simple equation (2) [7].

$$w(i,k) = tf(i,k)\log_2 \frac{N}{n_i} \qquad (2)$$

Where tf(i,k) represents the frequency of word segment i in document k. N is the total number of documents, ni is the number of word segment i contained in the N documents, i.e., the inverse document frequency of word segment i. The logarithm operation in equation (2) means that if the feature term exists in the document, it is set as 1. Otherwise, it is set as 0. To avoid the situation where the weight is 0, the logarithm calculation can be modified as shown in equation (3).

$$w(i,k) = tf(i,k)\log_2\left(\frac{N}{n_i} + 0.01\right) \qquad (3)$$

It is shown in equation (3) that the higher the occurrence frequency of the feature term is, the less information entropy it would contain. If the occurrence of the feature term is limited to a small number of texts, more information entropy would be contained.

Considering the effect of text length on weights, the equation should be normalized so that the weight of each term could fall into the range of [0, 1]. The normalized equation is given by equation (4)

$$w(i,k) = \frac{tf(i,k)\log_2\left(\frac{N}{n_i} + 0.01\right)}{\sqrt{\sum_{i\in k}\left[tf(i,k)\log_2\left(\frac{N}{n_i} + 0.01\right)\right]^2}} \qquad (4)$$

The expressing of a cluster is similar to the expressing of a document vector. To query a document vector in clusters, the calculation of feature weights in clusters and that in document vectors are subtly distinguished. That is, the weight of word segment i in cluster k is calculated using equation (5)

$$w(i,k) = \frac{\left[1+\log_2 tf(i,k)\right]\log_2\left(\frac{N+1}{n_i+0.5}\right)}{\sqrt{\sum_{i\in k}\left\{\left[1+\log_2 tf(i,k)\right]\log_2\left(\frac{N+1}{n_i+0.5}\right)\right\}^2}} \qquad (5)$$

3) News events model

Generally speaking, a news events model has either centroid vectors or central vectors. Central vector method uses one document of the news events as the center. However, this method needs large-scale search in order to localize central vectors, and thus could be easily affected by inaccurate selection of central vectors, resulting in subsequent wrong clustering. Centroid vector method uses the average value of all document vectors in a cluster to represent obtained news events information. Although it is

sensitive to noise and outliers, its application could help to avoid the problem of error propagation. Considering that news events clustering method requires high levels of efficiency, the centroid vector method is adopted for the news events model in this study as shown in equation (6), i.e., the average value of news text vectors from all obtained news events information [8].

$$C_N = \frac{1}{N}\sum_{k=1}^{N} d_k \tag{6}$$

Where N is the number of documents found from the news and event clusters. The document vector is $d_k = (t_1, w_1, ..., t_k, w_k)$.

4) Similarity calculation

Text similarity is a statistical technique to measure the similarity degree between texts. The similarity measurements of texts include cosine method, inner product method, distance function method and so on. The angle cosine equation is adopted to calculate the similarity between two vectors as defined in equation (7)

$$\mathrm{sim}(\mathrm{d_i}, \mathrm{d_j}) = \frac{\vec{d}_i \vec{d}_j}{|\vec{d}_i||\vec{d}_j|} = \frac{\sum_{k=1}^{m} d_i d_j}{\sqrt{\left(\sum_{k=1}^{m} d_i^{\,2}\right)\left(\sum_{k=1}^{m} d_j^{\,2}\right)}} \tag{7}$$

where d_i is the feature vector of new text, d_j is the centroid vector of the j-th cluster, and m is the dimensionality of a feature vector.

B. Staged Incremental News Event Clustering Model.

1) The calculation of the similarity between text vectors and centroid vectors

The similarity calculation between news text vector di and news event centroid vector zj is based on angle cosine between these vectors. By substituting the weights of two vectors into equation (7), we could obtain equation (8).

$$\mathrm{sim}(\mathrm{d_i}, \mathrm{z_j}) = \frac{\vec{d}_i \vec{z}_j}{|\vec{d}_i||\vec{z}_j|} = \frac{\sum_{k=1}^{m} w(\mathrm{i},\mathrm{k})\, w(\mathrm{j},\mathrm{k})}{\sqrt{\left(\sum_{k=1}^{m} w(i,k)^2\right)\left(\sum_{k=1}^{m} w(j,k)^2\right)}} \tag{8}$$

where z_j is the centroid vector of cluster C_i. After the calculation, the level of similarity is compared with threshold Tc, which is in the range of [0.4, 0.7]. If the similarity is not within the range, the text vector is compared with the next centroid vector. Otherwise, the text vector is added into cluster Cj.

2) The definition of staged lists

A staged list P(M+1) is defined for each cluster. Each node of the list saves a news events vector. The (M+1)-th node is used to save unclassified text vectors. Due to algorithmic complexity, the length of the lists should not be too long. The number of news event vectors in cluster Ci is set as m, and the length of the lists is M<m/100.

3) The update of the staged lists

Since the content of the lists determines the centroid vector, the lists could affect the variation of similarity, and thus fundamentally affect the clustering accuracy. Therefore, it is necessary to reasonably update the contents of the lists. A replacement method is adopted in the current study to update list P. That is, if the similarity index $\mathrm{sim}(\mathrm{d_i}, \mathrm{z_j}) \in [0.5, 0.7]$, then replace vector P[1] that first enters the list with di. Otherwise, list P remains unchanged.

4) The initialization of centroid vectors

Centroid vector Z is the feature vector of cluster Ci. The essence of centroid vectors is to calculate the average value of vectors. With an increase in the number of news events in clusters, computational complexity would also increase. Staged list P might help solve this problem. The initialization of centroid vectors is according to the change of staged list. If a list is not filled up, then the actual element number n is used to calculate the average. Otherwise, the predefined list length M is used, which is expressed in equation (9)

$$Z = \begin{cases} \dfrac{1}{n}\sum_{k=1}^{n} P_k & (n \le M) \\ \dfrac{1}{M}\sum_{k=1}^{M} P_k & (n > M) \end{cases} \tag{9}$$

5) The update of centroid vectors

The news events clusters are updated with the addition of more documents. Therefore, the centroid vectors should also be dynamically updated. Since a list is defined for each cluster, the relation between lists and centroid vectors should be considered. The incremental updating equation for centroid vectors after list updating is derived from equation (6), which is also a combination of the functions of staged lists as shown in equation (10)

$$Z' = \frac{1}{M+1}[Z + P_1] \tag{10}$$

where M is the length of staged lists, Z is the previous centroid vector, and P1 is the vector newly added to the list.

Since the dimensionality of cluster vectors is limited, the centroid vectors could not accurately represent cluster characteristics. To improve the accuracy of centroid vector Z, the centroid vector data need to be normalized. One method can be used where the average is calculated chronologically as shown in equation (11).

$$Z_t = \frac{1}{N}\sum_{k=1}^{N} C_{ik} \tag{11}$$

Where Zt is the centroid vector of Ci at time t, Cik is the vectors of cluster Ci, and N is the number of cluster Ci.

C. Algorithm Implementation Steps.

The implementation steps of the staged incremental Single-Pass news events clustering algorithm are as follows:

Input: text vectors of news events, d1,d2,d3,…,dn;

Output: news events clusters, C1,C2,C3,…,Ck;

for each moment, input text vector set di(d1,d2,d3,…,dn), do{

if no cluster exists, then {

create cluster Ci; initialize list Pi, and centroid vector Zi; go to step 9);}

else save di to the (M+1)-th node in the staged list Pi of existing news events cluster Ci;

calculate the level of similarity between di and centroid vector Ci, i.e., Ti;

if Ti>Tc, then classify di into cluster Ci;

if 0.5<=Ti<=0.7, then update list Pi; dynamically update centroid vector Zi;

else(Ti<=Tc) go to step 3);

i++;

} out put all news event clusters, C1, C2, C3, …, Ck;

End

In the algorithm, steps 2) to 9) complete the processing and clustering of a text vector di. Particularly, step 3) is the establishment of a cluster, and steps 4) to 7) add the text vector di into an existing cluster by checking the level of similarity, and dynamically update the list and centroid vector according to the threshold. n in the algorithm is the number of text vectors simultaneously processed, and k is the number of news events clusters be finally outputted. The computational complexity of the algorithm is O(nk). Therefore, the news events clustering method based on staged incremental Single-Pass has relatively high level of time efficiency.

IV. EXPERIMENTAL ANALYSIS

A. Algorithm Assessment Indices.

The assessment indices for the proposed clustering system are in accordance with the measurements of text retrieval, which are two commonly used performance assessment indices, namely, recall ("r" for short) and precision ("p" for short) [9]. The significance of Single-Pass algorithm when applied to deal with news events clustering lies in improved timeliness and accuracy. That is, the algorithm could help to accurately and effectively cluster news events within a limited timeframe. Therefore, these two mentioned indices are used as the assessment methods in our experiments.

Precision, also known as accuracy, is the ratio of the texts correctly categorized as a class by a classifier. Recall refers to the ratio between the texts categorized as a class by a classifier and the texts actually belonging to that class. The number of texts correctly categorized as a class by a classifier is referred to as a; the number of texts that are incorrectly categorized into that class is referred to as b; the number of misclassified texts by a classifier is referred to as c. The calculation algorithm of recall and precision is shown in equation (12) and (13), respectively.

$$r = \frac{a}{a+c} \tag{12}$$

$$p = \frac{a}{a+b} \tag{13}$$

B. Data Source.

The data of the experiments are acquired from NetEase classified news data sets and 10 sets of news data between Sep. 20th and Sep. 22nd , 2013, which are classical text classification data sets. The data preprocessing includes web purification, word frequency statistics and feature extraction word. Data are categorized into 10 classes, including finance, culture, education, recruitment, tourism, IT, health, military, automotive and sports. There are 1759 data samples, among which 959 are training samples and 800 are test samples. The experiment platform is developed under Visual C++ 6.0 environment. The test results are shown in TABLE I .

TABLE I A STATISTICAL TABLE OF NEWS EVENTS CLUSTERING PERFORMANCE

Class	Total No.	Traditional Single-Pass			Staged incremental Single-Pass		
		Time	Recall	Precision	Time	Recall	Precision
automotive	42	0.1	0.74	0.94	0.1	0.74	0.89
recruitment	43	0.12	0.84	0.71	0.12	0.84	0.75
finance	48	0.14	0.54	0.67	0.15	0.73	0.76
education	58	0.2	0.64	0.86	0.18	0.78	0.88
tourism	67	0.22	0.64	0.54	0.2	0.67	0.62
military	85	0.25	0.84	0.95	0.22	0.89	0.96
health	100	0.5	0.97	0.92	0.4	0.95	0.93
culture	107	0.8	0.82	0.59	0.5	0.87	0.66
sports	199	1	0.93	0.97	0.6	0.93	0.98
IT	210	1.2	0.82	0.95	0.67	0.88	0.95

It can be seen clearly from the table 1 that the use of the staged incremental Single-Pass method could help improve the news events clustering performance when compared with traditional Single-Pass. The time efficiency, precision and recall are all significantly improved, with an average precision of 84% and an average recall of 83%.

C. Result Analysis.

It is shown in Figure. 3 that the accuracy of news events clustering is irrelevant to text numbers, but mainly depends on the convergence of centroid vectors. Traditional Single-Pass algorithms require that centroid vectors be updated each time when text categorization is done. Therefore, N (the number of texts belonging to this cluster) times of centroid vector updating are performed for a class[10-12]. The larger N is, the stronger the convergence of the centroid vectors is, and the updating results would affect the accuracy of clustering. Therefore, in order to improve the convergence of centroid vectors, selective measurements have to be taken. Or, in other words, the centroid vectors should be determined based on valid texts. The Single-Pass method based on staged incremental strategy could help solve this selection problem. On the one hand, staged lists could save representative texts locally. On the other hand, the centroid

vectors are updated incrementally in a timely manner by comparing the similarity value and threshold value.

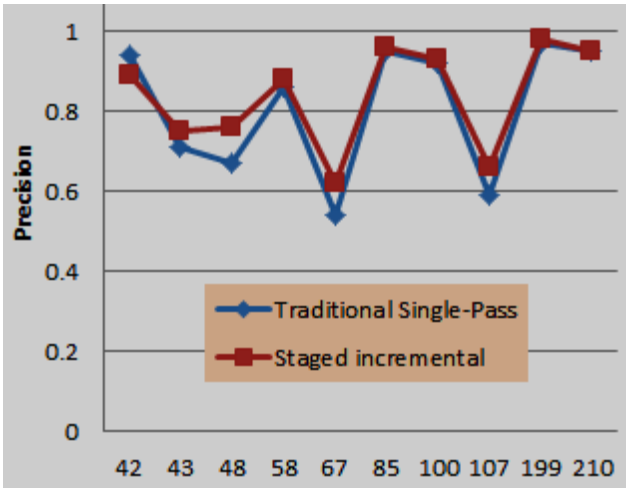

Figure. 3 Precision comparison of the two methods.

The time complexity of ordinary Single-Pass algorithms is O(nk), with spent time on processing each document as Ti=Tc+T(m), where

Tc is similarity calculating time;

T(m) is spent time on updating centroid vector updating;

m is the number of texts included in each cluster;

T(m) would increase with an increase of m;

Although the time complexity of staged incremental Single-Pass is also O(nk), spent time on processing each document is Ti=Tc+a*T(M), where

a is the probability of updating centroid vectors;

T(M) is spent time on updating centroid vectors;

M is the length of each staged list.

Since T(M) increases with an increase in M (M<m/100), T(M) is much less than T(m). Figure. 4 shows the comparison of the two algorithms in terms of time efficiency. When the number of documents is relatively small, spent time on the two algorithms is not significant. However, when the number of documents is larger than the length of the list (85), the traditional Single-Pass algorithm would require significantly more time as n increases. By contrast, the increase of spent time on staged incremental Single-Pass is not significant.

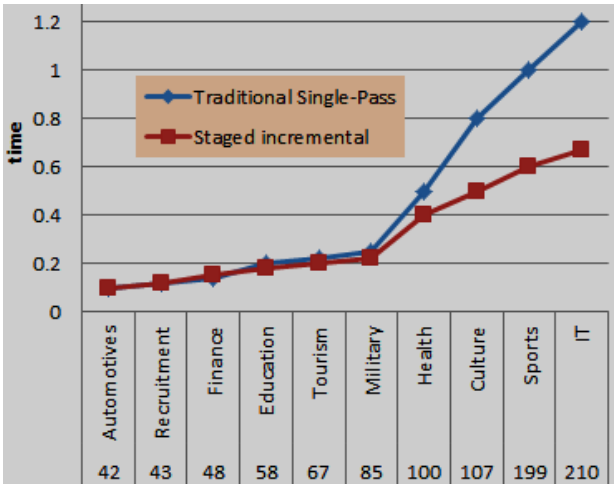

Figure. 4 A comparison of time efficiency when both methods are applied.

V. CONCLUSIONS

Experimental results show that the proposed news events clustering method based on staged incremental Single-Pass could help improve timeliness and accuracy when compared with traditional Single-Pass algorithm. The selective dynamic updating of staged lists and centroid vectors could not only reduce the computational burden commonly found in traditional incremental methods, but also increase the precision of news events clustering, with time complexity significantly reduced.

ACKNOWLEDGMENT

The research is supported by:

National Natural Science Foundation of China No. 51204071

Science and Technology Planning Project of Guangdong Province (2012B010100019).

Ministry of Education University-Industry-Research Project of Guangdong Province (2010B090400535).

The Fundamental Research Funds for the Central Universities(2013ZZ0047).

Scientific research project of guangxi colleges and universities (ZD2014137)

Guangxi education science "twelfth five-year" plan project: Research on the teaching resources informationization construction In the teaching mode of MOOC(2015C435)

REFERENCES

[1] Xiaori Xu. “Study on the Way to Solve the Paroxysmal Public Feelings on Internet”. Journal of North China Electric Power University (Social Sciences). Vol.1, pp. 89-93, 2007.

[2] Ronglu Li. “Research on Text Classification and Its Related Technologies.Shanghai”: Fu Dan University, 2005.

[3] Guha S, Mishra N, Motwani R, et al. ”Clustering Data Streams”. In:Proceedings of the Annual Symposium on Foundations of Computer Science. pp.359-366, 2000.

[4] Gupta C, Grossman R L.GenIc. “A Single Pass Generalized Incremental Algorithm for Clustering”. In: Proceedings of the 2004

SlAM International Conference on Data Mining, Philadelphia. Pp.137-153, 2004.

[5] Chan BAI, Chunxia JIN, Hui ZHANG. "Topic-text clustering algorithm based on word co-occurrence". Computer Engineering & Science. Vol.35, No.7, pp. 166-167, 2013.

[6] Yan Li, Yun Lou. "Applied research of text clustering algorithm in network monitoring public opinion". Electronic Design Engineering. Vol.21, No.1, pp. 70-71, 2013.

[7] Fan Li, Aiwu Lin , Guoshe Chen. "A Chinese text categorization system based on the improved VSM". Journal of Huazhong University of Science and Technology. Vol.33, No.3, pp.53-54, 2005.

[8] Yahong LI, Suge WANG, Deyu LI. "Exploiting multiple semantic features for comment text topic clustering".Computer Engineering and Applications. Vol.49, No.2, pp.190-191, 2013.

[9] Kuo ZHANG, JuanZi LI, Gang WU. "A New Event Detection Model Based on Term Reweighting". Journal of Software. Vol.19, No.4, pp.817-828, 2013.

[10] XiaoMing ZHANG, ZhouJun LI, WenHan CHAO. "Research of Automatic Topic Detection Based on Incremental Clustering". Journal of Software. Vol.23, No.6, pp.1578-1587, 2012.

[11] Kansheng SHI, Haitao LIU, Yincai BAI. "Text Clustering Method with Improved Fitness Function and Cosine Similarity Measure . Journal of University of Electronic Science and Technology of China". Vol.42, No.4, pp.621-622, 2013.

[12] Chunyan ZHANG, Zhiqing MENG, Pei YUAN. "Mining Algorithm for Temporal Text Association Rules in Text Mining". Computer Science. Vol.40, No.6, pp.220-221, 2013.