Journal of Computers Vol. 29 No. 4, 2018, pp. 122-134
doi:10.3966/199115992018082904010

Research on Unequal Time Series Clustering for Hot Topics

Fu-Lian Yin[1], Bei-Bei Zhang[2*], Jing-Chun Liao[3], Jian-Bo Liu[4]

[1,2,3,4] College of Information Engineering, Communication University of China,
Beijing, 100024 China
yinfulian@cuc.edu.cn; cuc_zhangbeibei@qq.com; 1613056763@qq.com; ljb@cuc.edu.cn

Received 2 March 2017; Revised 2 July 2017; Accepted 9 August 2017

Abstract. In the traditional research on time series clustering for hot topics, the granularity of time series with day as the unit is coarse, which causing the bad timeliness. In this paper, a fine - grained hot topic time series acquisition scheme is proposed, which can make the time series of topics accurate to T_0 hour. The distance calculation of time series lost part of information to fit the unequal time series clustering in the traditional clustering algorithm. In this paper, the S-Euc distance (Segmented Euclidean distance) and S-DTW distance (Segmented Dynamic Time Warping distance) are introduced to segment the time series and calculate the total distance. The two method significantly raise in computational speed, silhouette coefficient and cluster compactness, compared with the traditional DTW distance and Euclidean. When the cluster number is bigger, the clustering silhouette coefficient of S-Euc based algorithm is about 8% higher than the S-DTW based algorithm. In the case of small number of clusters, the silhouette coefficient of S-DTW is about 65% higher than the S-Euc, but the computational complexity is higher.

Keywords: data acquisition, DTW distance, hierarchical clustering, time series clustering

1 Introduction

With the booming development of social media and the wide use of search engines and social networks, a wide range of information is spreading across the global at an alarming rate, and the rapidly changing topics appearing one after another on the Internet have greatly influenced people's life and all aspects of society. These topics evolve over time, forming typical time series data. The academia has carried out extensive researches on how to use these data to analyze the time evolution rule of hot issues, forecast their development, grasp the trend in advance and detect public opinion online.

There are many kinds of topic in the Internet, but the development of their time series is similar in general. By classification or clustering, topics can be classified into several categories. Due to similarity, time series prediction can be applied to topics in the same category. Generally, the algorithm flow of time series clustering can be divided into three stages: the representation for time series, the calculation of time series distance, and the clustering of time series [1]. The representation for time series is used to extract features from time series in order to descend dimension and reduce algorithm complexity. Feature-based clustering algorithm for time series refers to the morphological features, structural features and model feature data of time series instead of time series observations themselves, so as to create clustering methods focus on those typical features [2]. Frequently-used feature extraction methods are Discrete Fourier Transform (DFT) [3], Discrete Wavelet Transform (DWT) [4], Symbolic Aggregate Approximation (SAX) [5], Piecewise Linear Representation (PLR) [6], etc. To overcome the deficiencies in describing the overall sequence change trend, researchers put forward Piecewise Aggregate Approximation (PAA) [7] describing time series in a simple and intuitive way, getting better reduction and compression effect. But the mean approximation process may cause important data lost. Hung N Q V [8] proposed PLAA (Piecewise Linear Aggregate Approximation), solving the problem to a certain extent. But these time series representation methods compress the original time series data by remaining

[*] Corresponding Author

the key points data, removing part of points with noise or less information. This kind of time series processing will bring additional problems. Same length time series may become unequal length. It will lead to the difficulty for the series similarity measure.

Table 1. Comparison of different representation for time series

Representation for time series	Advantage		Disadvantage	
Discrete Fourier Transform (DFT)	Compress data.		Deficiency in describing the overall sequence change trend.	Same length time series may become unequal length.
Discrete Wavelet Transform (DWT)				
Symbolic Aggregate Approximation (SAX)				
Piecewise Linear Representation(PLR)				
Piecewise Aggregate Approximation (PAA)	/	1. Better compression. 2. Describing the overall sequence change trend well.	Mean approximation method may cause important data lost	
Piecewise Linear Aggregate Approximation(PLAA)	Avoiding the possible loss of important data.		/	

The classic series similarity measure today are Euclidean distance and DTW (Dynamic Time Warping) distance. The widely used Euclidean distance is simple formulated and easy to calculate, but it can only handle same length series. DTW distance was firstly proposed for speech recognition to solve template matching problems of the different lengths in pronunciation [9-10]. DTW distance supports the deformation on the time axis, so it can effectively overcome the shortcomings of the Euclidean distance. But its time cost is too large, which limited its application scope [11-12]. To overcome the large time cost, researchers proposed LEP (Locally Extreme Point) of time series [13]. In which, the original time series is described by extracting the locally extreme points from time series, reflecting the main features of the time series effectively and achieving the compression of time series. But it isn't suitable for the characteristic of time series for hot topics, because the hot topics don't always have so many extreme point, which can't reduce the time complexity obviously. Yang et al. [14] introduced a K_SC clustering algorithm for topic time series which can effectively describe the center curve of each cluster and the trend of topic development in clustering results with high accuracy. But the K_SC clustering algorithm is sensitive to the initial cluster center and time-cost. Han et al. [15] optimized the K-SC clustering algorithm, and a WKSC (Wavelet-based K-spectral centroid) clustering algorithm with lower computational complexity and better clustering effect is obtained. However, these clustering algorithms all assume that the topic time series have same length. The similarity measure method neglects the peaks of hot topics, simply aligning the peaks of the different series on the time axis. As a result, they ignore the different stages duration of the topic. Liu et al. [16] designed STS (short time series) distance of unequal time series based on sliding window and a center curve calculation method with a similar to K-means clustering algorithm to solve the problem of unequal time series. But it considers that the local features of time series are more valuable than global ones, and the two curves with unequal length but similar local information are clustered into one class, regardless of whether the rest of the information in the two sequences is similar. Which means it cannot effectively identify the existence of derived public opinion or detect the similarity between different opinions.

Above all, in the traditional research on time series clustering for hot topics, the time series with day as the unit is coarse, which causing the bad timeliness. In view of the bad timeliness problem, this paper proposed a fine-grained hot topic time series acquisition scheme, which can make the time series of topics accurate to T_0 hour. When the analysis unit changes from day to T_0 hour, the traditional time series similarity measure method isn't practicable. Combined with the characteristic of time series for hot topics, this paper put forward S-DTW distance and S-Euc distance as a time series distance calculation method. Compared with the traditional Euclidean and DTW distance, the proposed method can handle the unequal time series cluster problem and greatly improve the clustering effect with less information distortion and lower time cost.

Table 2. Comparison of different similarity measure method for time series

<table>
<tr><th>Similarity measure method for time series</th><th colspan="2">Advantage</th><th colspan="2">Disadvantage</th></tr>
<tr><td>Euclidean distance</td><td colspan="2">Simple formulated and easy to calculate.</td><td colspan="2">Can’t handle unequal time series.</td></tr>
<tr><td>DTW (Dynamic Time Warping)</td><td colspan="2">Applicable in non-equal time series.</td><td colspan="2">Large time cost.</td></tr>
<tr><td>LEP (Locally Extreme Point)</td><td colspan="2">Lower time cost than DTW.</td><td colspan="2">Unsuitable for the characteristic of time series for hot topics.</td></tr>
<tr><td>K-SC (K-spectral centroid)</td><td>/</td><td rowspan="2">Describe the center curve of each cluster effectively.</td><td>Sensitive to the initial cluster center and large time cost.</td><td rowspan="2">1. Can’t handle unequal time series.
2. Ignore the different stages duration of the topic.</td></tr>
<tr><td>WKSC (Wavelet-based K-spectral centroid)</td><td>Lower time cost.</td><td>/</td></tr>
<tr><td>STS (Short time-series)</td><td colspan="2">Applicable in unequal time series.</td><td colspan="2">Sensitive to scaling.
Capture temporal information, regardless of the absolute values.</td></tr>
</table>

2 Fine-Grain Time Series Data Acquisition Scheme

In traditional time series researches, the time series are researched in terms of individual days, consequently the timeliness is far from enough whether they are used in time series clustering or forecasting. Moreover, the starting point and end of time series data are mostly artificial, which is not smart and relatively subjective.

The fine-grained time series acquisition method proposed in this paper can automatically acquire the viewings of a topic from the beginning to the end in the complete life cycle, and make the time series accurate to T_0 hours, which greatly improves the timeliness. Clustering based on fine-grained time-series data can also refine the classification of topic development, and predict the future time series more accurate. This article assumes that reading trends of the hot topic on micro blog platform can represent the overall trend of the topic. Micro blog provides a website where hot topic views is updated in real-time. Fine-grained topic time series of hot time collection program is as follows:

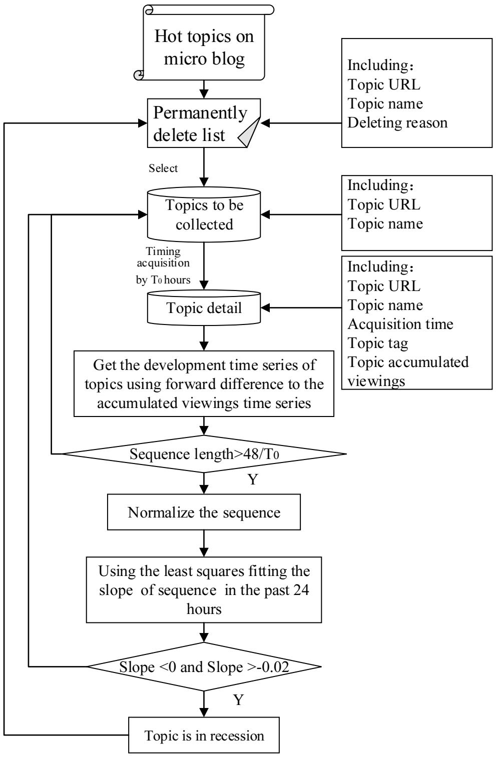


Fig. 1. Fine-Grain Time Series Data Acquisition Scheme

The permanently delete list stores the delete topics and why topics are deleted according to some conditions such as the topics in recession or some other conditions artificially.

3 Time Series Distance

3.1 Traditional Distance

Euclidean distance. Euclidean distance is the simplest common method of measuring the distance between two sequences. For example, the Euclidean distance of time series x and y is

$$D_{euc} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \quad \textbf{(1)}$$

This distance cannot be scaled on the time axis, and the distance between unequal time series cannot be calculated. For example, for {a, a, b, c} and {a, b, c, c}, although the sequence similarity is very high, but its Euclidean distance is relatively large.

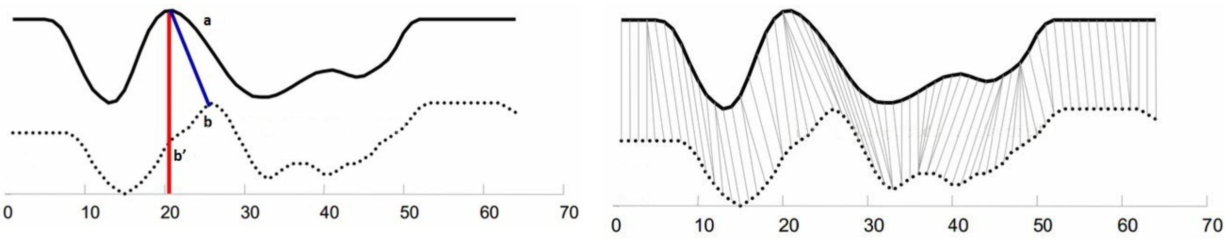
Dynamic time warping distance. Compared with the Euclidean, dynamic time warping distance (DTW) supports the bending on the time axis, so it's easily applied to the unequal time series similarity measure.

Definition 1. The dynamic time warping distance between time series *x* and *y* is defined as:

$$\begin{gathered} D_{tw}(<>,<>) = 0, \\ D_{tw}(\boldsymbol{x},<>) = D_{tw}(<>,\boldsymbol{y}) = \infty, \\ D_{tw}(\boldsymbol{x},\boldsymbol{y}) = \delta(head(\boldsymbol{x}), head(\boldsymbol{y})) + \min \left\{ \begin{matrix} D_{tw}(\boldsymbol{x}, rest(\boldsymbol{y}), \\ D_{tw}(rest, (\boldsymbol{x}), \boldsymbol{y}, \\ D_{tw}(rest(\boldsymbol{x}), rest(\boldsymbol{y})) \end{matrix} \right\} \end{gathered} \quad \textbf{(2)}$$

The base distance in this paper takes $\delta(x_i, y_i) = (x_i - y_i)^2$.

The DTW distance actually depends on the match relation between each point on the sequence *x* and *y*. In Fig. 2(a), the two curves on the overall shape of the waveform is very similar, but not aligned on the time axis. For example, at t_{20}, Point a of the solid line corresponds to Point b' of the dotted line. The traditional Euclidean cannot reflect the sequence similarity, and the Point a in the solid line aligning to the Point b in the dotted line can improve the similarity. In Fig. 2(b), the DTW distance aligns the two waveforms. That is their best matching path, making the two curves have the highest similarity.



(a)

(b)

Fig. 2

Table 1 is the algorithm to calculate the DTW distance [13].

Table 1. Calculations of the DTW distance between time series $\boldsymbol{x}$ and **y**

```
Given:
    $\boldsymbol{x} = x_1, x_2, \dots, x_n$, the first time series with length n;
    $\boldsymbol{y} = y_1, y_2, \dots, y_m$, the second time series with length m;
Output:
    cost: a matrix of size $n \times m$ including the cost values. $\boldsymbol{cost}_{n,m}$ is the DTW distance between x and y;
    path: a matrix of size $n \times m$ including a warping path.
----------------------------------------------------------------
    $D_{tw}(\boldsymbol{x}, \boldsymbol{y})$:
    Let $\delta$ be a distance between coordinates of sequences;
    $cost_{1,1} = \delta(x_1, y_1)$;
    $path_{1,1} = (0, 0)$;
    for $i \leftarrow 2$ to n do
        $cost_{i,1} = cost_{i-1,1} + \delta(x_i, y_1)$;
    end
    for $j \leftarrow 2$ to m do
        $cost_{i,j} = cost_{1,j-1} + \delta(x_1, y_j)$;
    end
    for $i \leftarrow 2$ to n do
      for $j \leftarrow 2$ to m do
        $cost_{i,j} = \min(cost_{i-1}, cost_{i,j-1}, cost_{i-1,j-1}) + \delta(x_i, y_j)$;
        $path_{i,j} = \text{min\_index}((i-1, j), (i, j-1), (i-1, j-1))$;
    end
end
Return cost, path.
```

3.2 Optimized Segmentation Distance

Segmented Euclidean distance. Compared with the traditional Euclidean distance, the segmented Euclidean distance is firstly divided by the "natural day", then calculate the segmented Euclidean distance, and finally the segmented distance is integrated. This avoids increasing of distance when matching distances across days.

Table 2 is the calculation procedure of S-Euc distance.

Table 2. Calculations of the S-Euc distance between time series x and $\boldsymbol{y}$

```
Given:
    $\boldsymbol{x} = \left\{x_{11}, \dots, x_{1a}, x_{21}, \dots, x_{2\frac{24}{T_0}}, \dots, x_{n1}, \dots, x_{nc}\right\}$, the first time series with n days;
    $\boldsymbol{y} = \left\{y_{11}, \dots, y_{1b}, y_{21}, \dots, y_{2\frac{24}{T_0}}, \dots, y_{m1}, \dots, y_{md}\right\}$, the second time series with m days;
    $1 \le a, b, c, d \le \frac{24}{T_0}$, $x_{ij}$ and $y_{ij}$ is the point j when the topic A and B happens for i days.
Output:
    $D_{seuc}(\boldsymbol{x}, \boldsymbol{y})$.
----------------------------------------------------------------
    SegmentByDay$(\boldsymbol{x})$:
    Get $\boldsymbol{x} = \left\{\boldsymbol{x_1} = \{x_{11}, \dots, x_{1a}\}, \boldsymbol{x_2} = \{x_{21}, \dots, x_{2\frac{24}{T_0}}\}, \dots, \boldsymbol{x_n} = \{x_{n1}, \dots, x_{nc}\}\right\}$;
    SegmentByDay$(\boldsymbol{y})$:
    Get $\boldsymbol{y} = \left\{\boldsymbol{y_1} = \{y_{11}, \dots, y_{1b}\}, \boldsymbol{y_2} = \{y_{21}, \dots, y_{2\frac{24}{T_0}}\}, \dots, \boldsymbol{y_m} = \{y_{m1}, \dots, y_{md}\}\right\}$;
```

```
if $a \geq b$ do
    $\boldsymbol{y}_1 = \{\text{repete}(0, a-b), \boldsymbol{y}_1\}$;
  else do
    $\boldsymbol{x}_1 = \{\text{repete}(0, b-a), \boldsymbol{x}_1\}$;
if $c \geq d$ do
    $\boldsymbol{y}_{\mathrm{m}} = \{\text{repete}(0, c-d), \boldsymbol{y}_m\}$;
  else do
    $\boldsymbol{x}_n = \{\text{repete}(0, d-c), \boldsymbol{x}_n\}$;
if $n > m$ do
    $dist1 = \sum_{i=1}^{m} D_{euc}(\boldsymbol{x}_i, \boldsymbol{y}_i)$;
    $dist2 = \sum_{i=m+1}^{m} \boldsymbol{x}_i^2$ ;
    $D_{seuc}(\boldsymbol{x}, \boldsymbol{y}) = \text{sqrt}(\text{sum}(dist\,1, dist\,2))$;
if $n = m$ do
    $D_{seuc}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} D_{euc}(\boldsymbol{x}_i, \boldsymbol{y}_i)$;
if $n < m$ do
    $dist\ 1 = \sum_{i=1}^{n} D_{euc}(\boldsymbol{x}_i, \boldsymbol{y}_i)$;
    $dist\ 2 = \sum_{i=m+1}^{m} \boldsymbol{y}_i^2$;
    $D_{seuc}(\boldsymbol{x}, \boldsymbol{y}) = \text{sqrt}(\text{sum}(dist\,1, dist\,2))$;
Return $D_{seuc}(\boldsymbol{x}, \boldsymbol{y})$.
```

Modified Segmented dynamic time warping distance. Because the computational complexity of traditional DTW distance is too large, not suitable for large data set mining, and cross-day to carry out the bending of the time axis will make the topic after different days of information disorder. Due to these problems, in this paper, an S-DTW distance (Modified Segmented dynamic time warping distance) is designed. The time series will be divided into "natural day", then the segmented DTW distance is calculated, and finally the segmented distance is integrated. On the one hand, by the "natural day" section effectively avoids the disorder caused by information alignment across different days. On the other hand, in the same day at different moments, the development of the topic is similar, so the data on the same day can be scaled appropriately on the time axis so that the sequences match to the minimum distance.

Table 3 is the calculation procedure of S-DTW distance.

Table 3. Calculations of the S-DTW distance between time series *x* and *y*

```
Given:
  $\boldsymbol{x} = \left\{x_{11}, \ldots, x_{1a}, x_{21}, \ldots, x_{2\frac{24}{T_0}}, \ldots, x_{n1}, \ldots, x_{nc}\right\}$, the first time series with n days;
  $\boldsymbol{y} = \left\{y_{11}, \ldots, y_{1b}, y_{21}, \ldots, y_{2\frac{24}{T_0}}, \ldots, y_{m1}, \ldots, y_{md}\right\}$, the second time series with m days;
  $1 \leq a, b, c, d \leq \frac{24}{T_0}$, $x_{ij}$ and $y_{ij}$ is the $j_{st}$ viewings when the topic A and B happens for i days.
Output:
  $D_{stdw}(\boldsymbol{x}, \boldsymbol{y})$.
  SegmentByDay$(\boldsymbol{x})$:
  Get $\boldsymbol{x} = \left\{\boldsymbol{x_1} = \{x_{11}, \ldots, x_{1a}\}, \boldsymbol{x_2} = \{x_{21}, \ldots, x_{2\frac{24}{T_0}}\}, \ldots, \boldsymbol{x_n} = \{x_{n1}, \ldots, x_{nc}\}\right\}$;
```

```
$\text{SegmentByDay}(\boldsymbol{y})$:
**Get** $\boldsymbol{y} = \left\{ \boldsymbol{y_1} = \{y_{11}, \ldots, y_{1b}\}, \boldsymbol{y_2} = \{y_{21}, \ldots, y_{2\frac{24}{T_0}}\}, \ldots, \boldsymbol{y_m} = \{y_{m1}, \ldots, y_{md}\} \right\}$;
**if** $n \geq m$ **do**
    $dist\ 1 = \sum_{i=1}^{m} D_{tw}(\boldsymbol{x_i}, \boldsymbol{y_i})$;
    $dist\ 2 = \sum_{i=m+1}^{n} \boldsymbol{x_i}^2$;
    $D_{stdw}(\boldsymbol{x}, \boldsymbol{y}) = \text{sqrt}(\text{sum}(dist\ 1, dist\ 2))$;
**if** $n = m$ **do**
    $D_{stdw}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} D_{tw}(\boldsymbol{x_i}, \boldsymbol{y_i})$;
**if** $n < m$ **do**
    $dist\ 1 = \sum_{i=1}^{n} D_{tw}(\boldsymbol{x_i}, \boldsymbol{y_i})$;
    $dist\ 2 = \sum_{i=n+1}^{m} \boldsymbol{y_i}^2$;
    $D_{stdw}(\boldsymbol{x}, \boldsymbol{y}) = \text{sqrt}(\text{sum}(dist\ 1, dist\ 2))$;
**Return** $D_{stdw}(\boldsymbol{x}, \boldsymbol{y})$.
```

4 Clustering Method

The hierarchical clustering algorithm is selected to cluster the time series in this paper, because the DTW and the S-DTW distance are scaled and transformed on the time axis, and the each cluster center cannot be effectively described by iterative algorithm such as K-means and FCM method [11]. Furthermore, hierarchical clustering does not require multiple iterations, which means the algorithm complexity is low. Given the N time series and distance matrix of N * N, hierarchical clustering algorithm process is as follows:

Step 1. Initially, each object is set one class, getting up to N classes, and each class contains only one object. The distance between the classes adopts the maximum distance between classes.

Step 2. Find the closest two classes and merge into one class, the total number of classes minus one.

Step 3. Calculate the distance between new classes and all the old classes.

Step 4. Repeat step 2 and 3, until all classes merge into the set cluster number.

Step 5. Calculate number of samples in each class, and remove the classes whose number of samples is less than the 2%-5% total number of samples N.

Step 6. Calculating silhouette coefficient and take it as a clustering effect evaluation index.

Step 7. Select different cluster number K and repeat steps 1-6, getting silhouette coefficient curve changing with the number of clustering. Observe whether there is extreme value point in the curve and pick out the clustering number corresponding to extreme value point. Compare the cluster compactness under several clustering number, choose the clustering number corresponding to best cluster compactness as the optimal clustering number under each distance to cluster the time series and analyze the cluster result.

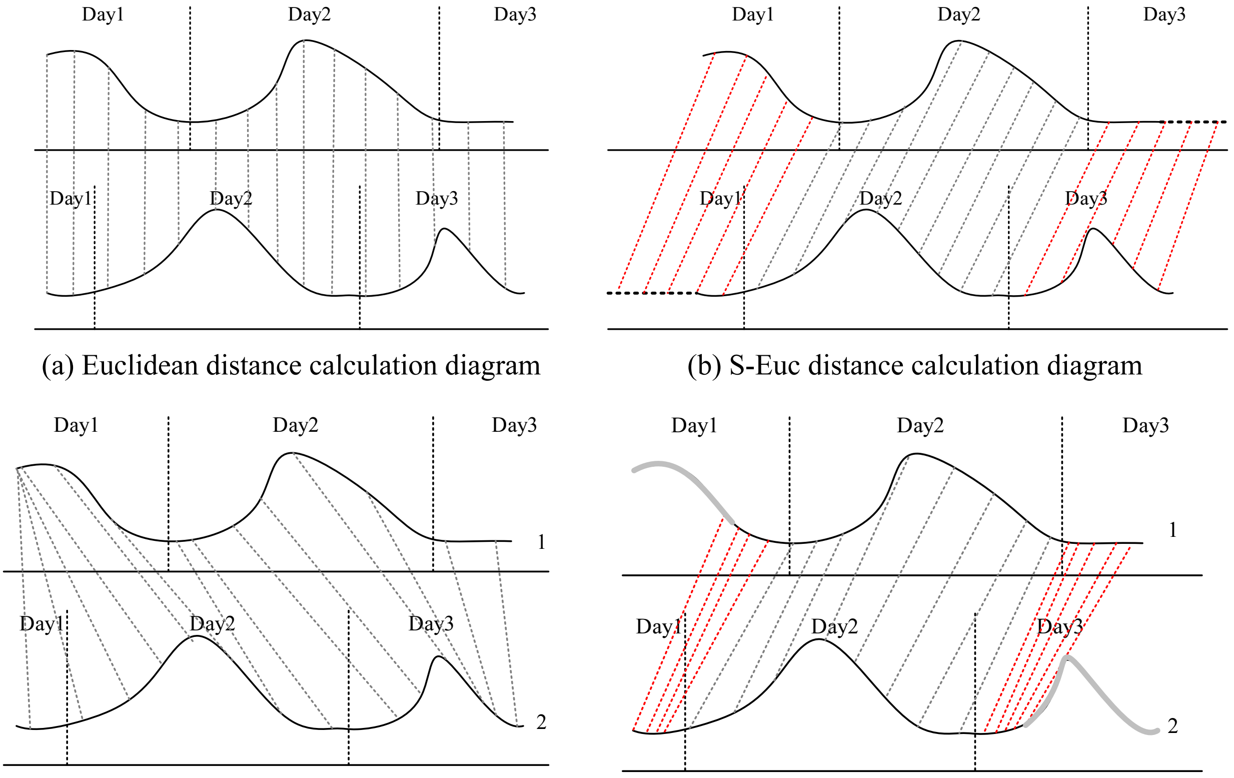
5 Experiment

This experiment mainly compares the clustering effect between the distances proposed in this paper and the traditional distances. The distances proposed in this paper include S-DTW and S-Euc. The traditional distances include DTW distance and Euclidean distance. The clustering effect is analyzed in theoretical performance, silhouette coefficient, clustering effect, etc. The concept of Silhouette coefficient was first proposed by Kaufman et al., Zhou et al. and Aranganayagi et al. [17-18] applied the Silhouette coefficient to the evaluation of clustering, which can effectively evaluate the clustering effect. The larger the Silhouette coefficient is, the better the clustering effect. The clustering effect is the subjective judgment of the compactness of the original data clustering in each cluster.

5.1 Theoretical Performance Analysis

Different distance calculation diagram is as follows, we can see that when the traditional DTW distance and Euclidean distance match distance across days, it is easy to increase the distance or confuse the information. DTW distance makes the whole time series match to the minimum distance. For the hot time series with different occurrence time point in the high tide period (Fig. 4), the two sequences are very similar, but the essence of the topic development is obviously different, the duration of the incubation period, the climactic period and the recession period are also completely different. The duration of the incubation period, the high tide period and the recession period can reflect the topic development law, which can be used as the distinguishing features of topics in different categories [19]. Therefore, the clustering effect of traditional DTW and Euclidean distance is not good theoretically.

There are two differences between the S-Euc and the S-DTW distance. At first, in the calculation process of S-DTW, the time axis will be scalable transformation, looking for the minimum matching distance of the two curves. S-Euc doesn't transform the time axis, so the recognition accuracy of the S-Euc will be higher and it will be able to identify the topic categories where the peak time point differs by T_0 hours. But in general, in the topic clustering, this high accuracy is not necessary, it's enough to identify the similar sequences. Otherwise those identified category are redundant. Secondly, the calculation of the distance between the first and last days of the topic is quite different. The S-DTW distance will give priority to finding a path that matches the shortest sequence of x_i and y_i to the shortest distance, so a portion of the data in the longer sequence will not be in the matching path, as the grey line in Fig. 3(d), some of the information is lost, but this part of the information only takes up a few hours, it is negligible for the overall distance calculation.



(a) Euclidean distance calculation diagram

(b) S-Euc distance calculation diagram

(c) DTW distance calculation diagram

(d) S-DTW distance calculation diagram

Fig. 3.

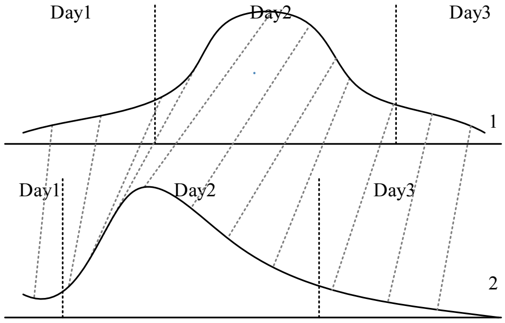


Fig. 4. DTW distance matching path of two sequences with different duration at different periods of topic

Table 4. Complexity of different distance.

Distance	Euclidean	DTW	S-Euc	S-DTW
Complexity	$O(\|\boldsymbol{x}\|)$	$O(\|\boldsymbol{x}\|*\|\boldsymbol{y}\|)$	$O(M*\|\boldsymbol{x}_i\|)$	$O(M*\|\boldsymbol{x}_i\|*\|\boldsymbol{y}_i\|)$

Set the time series are segmented by M at most, the complexity of different distance is as above table. Since the hot topic generally maintain 2-15 days, the length of x and y is generally between $\frac{2*24}{T_0}$ $\frac{15*24}{T_0}$. It can be seen that in the computational complexity, DTW > S-DTW > S-Euc > Euclidean. The traditional DTW distance calculation has the highest complexity. As the time series increases, the computational complexity grows at a square speed. The computational complexity of the other three distances is not significant in magnitude.

5.2 Silhouette Coefficient Analysis

In this paper, we select the hot topics of micro blog from August 1, 2016 to December 31, 2016 as data source and collect the data of topic readings, etc. Finally we use the topic time series acquisition method to screen out the topics in recession period (topics with a complete life cycle) with a total of about 840 topics. All of the procedures in this paper are performed on 64-bit Windows operating systems with Intel (R) Core (TM) i5-2400 CPU 3.1GHz 4GB RAM. It shows the Silhouette coefficients comparison of four distance in Fig. 5.

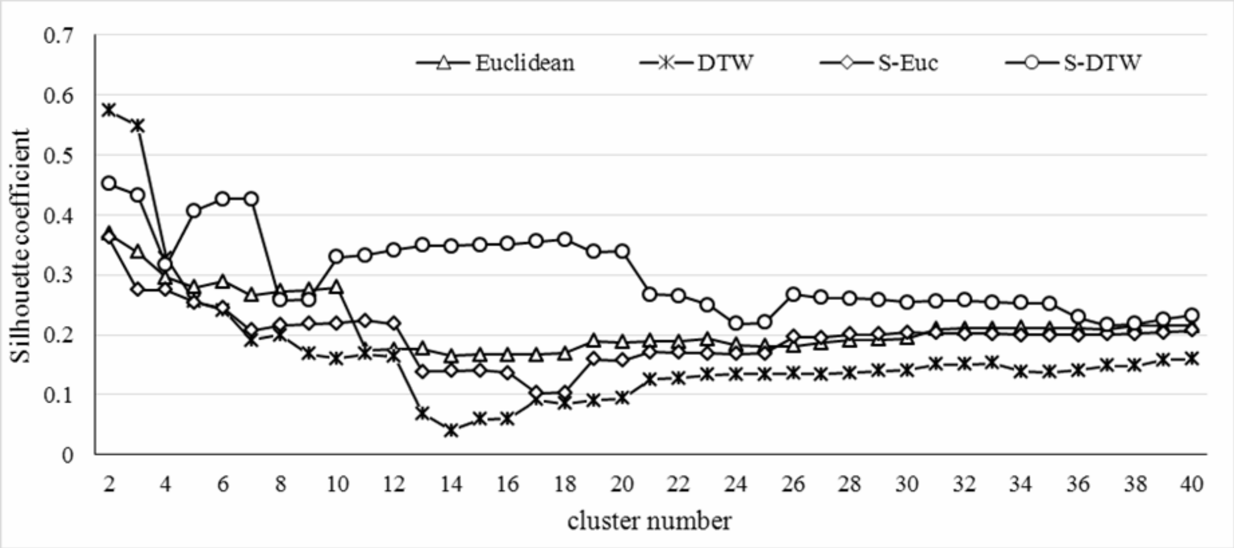


Fig. 5. Silhouette coefficients comparison of four distance

The figure shows that the silhouette coefficient of traditional Euclidean and DTW distance is lower than the two distance we present in this paper. It prove the S-Euc and S-DTW distance is better than the DTW and Euclidean in clustering effect. When the cluster number is bigger than 23, the clustering silhouette coefficient of S-Euc based algorithm is about 8% higher than the S-DTW based algorithm. When the cluster number is smaller than 23, the silhouette coefficient of S-DTW is about 65% higher than the S-Euc. It illustrate the importance and superiority of the time-segment method. S-Euc is more suitable for high classification accuracy. S-DTW is more suitable for general application because of the high silhouette coefficient and similar classification accuracy.

5.3 Cluster Compactness Analysis

For each distance, we'll pick out the clusters number corresponding to extreme point of coefficient curve, and draw the raw time series trend of each category according to the clustering results in Fig. 6. Then we compare the cluster compactness of several clustering numbers and choose the number under which it has the best cluster compactness as the best cluster number. In Fig. 6, each line in each class is colored by one color and the color density stands for the cluster compactness. The silhouette coefficient curve appeared two extreme points in S-DTW method, respectively, the cluster number is 7 and 18. So we draw the original data trend when cluster number is 7 and 18, and choose the best number of clusters in S-DTW method by comparing the cluster compactness.

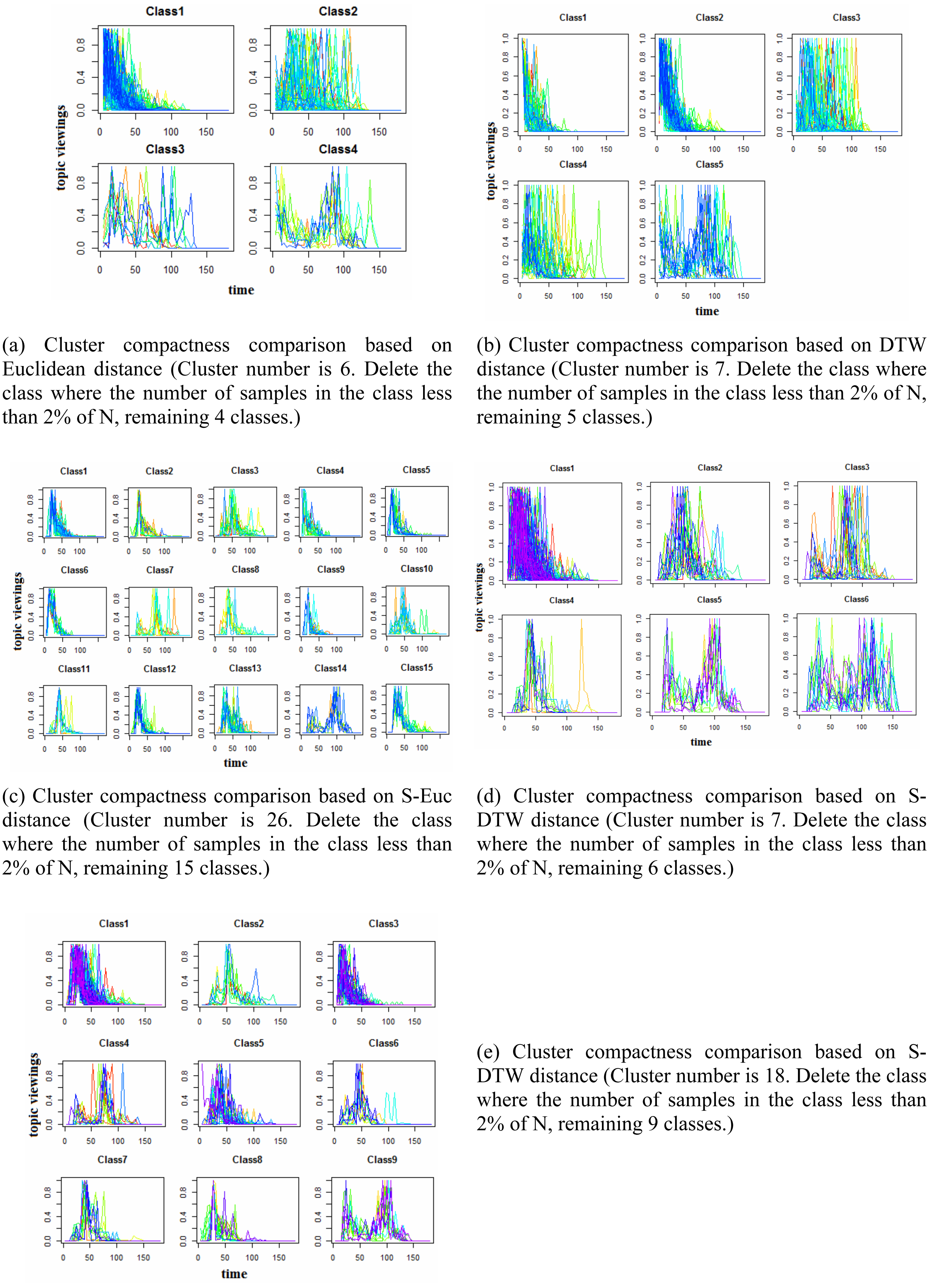
Fig. 6(a) and Fig. 6(b) shows that the clustering effect of traditional Euclidean and DTW distance is very bad. There are many classes with awful cluster compactness, such as class 2, 3 in Fig. 6(a), and class 3, 4, 5 in Fig. 6(b). The computational complexity of the DTW is too high, so the Euclidean and DTW distance is not suitable for time series clustering.

Fig. 6(c) shows that the clustering effect of the S-Euc distance proposed in this paper is very obvious when the cluster number is more than 23. The S-Euc distance can classify the time series more finely, but the overall trend of some categories is extremely similar, such as class 1, 5, 6, 9, 11, 12. It is redundant when they are identified as multiple categories. Fig. 6(d) and Fig. 6(e) shows that when the cluster number is less than 23, the proposed S-DTW distance has obvious cluster effect, which can characterize the trend of the topic well with high accuracy, and it can also identify classes with different peak arrival time and different topic development detail. When comparing the cluster compactness under cluster number 7 and 18, the effect is better when clusters is 18, so the best cluster number of the S-DTW method is 18.

The comprehensive evaluation of clustering performance of 4 distances shows that the traditional DTW distance and Euclidean distance are easy to increase distance or cause information disorder when matching the path by day. Therefore they are not suitable for distance measurement of topics' time series. When the number of clusters is less than 23, the silhouette coefficient of S-DTW distance is higher and the clustering effect is better. When the number of clusters is larger than 23, the silhouette coefficient of S-Euc distance is higher and the clustering effect is obvious. It means that the S-Euc distance can classify the time series more finely when the number of clusters is bigger. But there are too many classes whose trend is very similar, it will be identified as different categories which is redundant. The S-DTW distance is used to transform the time series in the same days after the occurrence of the topic, and the redundancy class is reduced when the accuracy of topic recognition is not changed greatly.

6 Conclusion

In this paper, a fine-grained hot topic time series acquisition method is proposed, which can make the topics' time series accurate to hour, greatly improving the timeliness. We also design the S-Euc distance and S-DTW distance, and compare their clustering performance with the traditional Euclidean distance and the DTW distance from three aspects, theoretical performance, silhouette coefficient and clustering compactness. It is found that the traditional DTW distance and Euclidean distance make the clustering compactness bad and silhouette coefficient very low. Moreover, the computation complexity of DTW distance is higher, which is not suitable for time series clustering. The proposed S-Euc and S-DTW distance has higher silhouette coefficient, better clustering compactness and lower computation complexity. When the number of clusters is bigger (more than 23), the silhouette coefficient of S-Euc is about 8% higher than S-DTW, but the clustering effect is a little redundant subjectively. When the



(a) Cluster compactness comparison based on Euclidean distance (Cluster number is 6. Delete the class where the number of samples in the class less than 2% of N, remaining 4 classes.)

(b) Cluster compactness comparison based on DTW distance (Cluster number is 7. Delete the class where the number of samples in the class less than 2% of N, remaining 5 classes.)

(c) Cluster compactness comparison based on S-Euc distance (Cluster number is 26. Delete the class where the number of samples in the class less than 2% of N, remaining 15 classes.)

(d) Cluster compactness comparison based on S-DTW distance (Cluster number is 7. Delete the class where the number of samples in the class less than 2% of N, remaining 6 classes.)

(e) Cluster compactness comparison based on S-DTW distance (Cluster number is 18. Delete the class where the number of samples in the class less than 2% of N, remaining 9 classes.)

Fig. 6.

number of clusters is smaller(less than 23), the silhouette coefficient of S-DTW distance is about 65% higher than S-Euc. In comprehensive evaluation, for the strict request of clustering accuracy of topic clustering model, which can distinguish the classes with different peak arrival time in an hour, S-Euc distance is the best distance measurement of time series, or the S-DTW distance is the best similarity measure of time series with general application because of the high silhouette coefficient and similar cluster effect.

However, our work may improve the clustering performance by changing the clustering algorithm. As we mentioned in section 4, iterative cluster algorithm such as K-means and FCM can't describe the each cluster center in every iteration, because the DTW and the S-DTW distance are scaled and transformed on the time axis. So we can do some work to apply the iterative cluster algorithm into unequal time series cluster, and compare the clustering performance between different clustering algorithms.

Acknowledgements

Foundation Items: Social Science Project of Ministry of Radio, Film and Television in the State Administration of Press, Publication, Radio, Film and Television of China (GD1739).

References

[1] T. W. Liao, Clustering of time series data: a survey, Pattern Recognition 38(2005) 1857-1874.

[2] Y. Ma, X. Gao, B. Pan, Trend feature-based clustering for research funding time series data, in: Proc. Logistics, Informatics and Service Sciences, 2015.

[3] D. Muruga, D. Radha, V. Maheswari, P. Thambidurai, Similarity search in recent biased time series databases using Vari-DWT and polar wavelets, in: Proc. Emerging Trends in Robotics and Communication Technologies (INTERACT), 2010.

[4] D. Castro-Hernandez, R. Paranjape, Classification of user trajectories in LTE het nets using unsupervised-shapelets and multi-resolution wavelet decomposition, Vehicular Technology PP(99)(2017) 1-1.

[5] S. Giovanni, V. Paola, Time makes sense: event discovery in twitter using temporal similarity, Web Intelligence (WI) and Intelligent Agent Technologies (IAT) 2(2014) 186-193.

[6] N. Subhani, L. Rueda, A. Ngom, A. Ngom, C. Burden, New approaches to clustering microarray time-series data using multiple expression profile alignment, in: Proc. Computational Intelligence in Bioinformatics & Computational Biology, 2010.

[7] A.G. Li, Z. Qin, Dimensionality reduction and similarity search in large time series databases, Chinese Journal of Computers (9)(2005)1467-1475.

[8] N.Q.V. Hung, D.T. Anh, An improvement of PAA for dimensionality reduction in large time series databases, in: Proc. Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence, 2008.

[9] H. Sakoe, S. Chiba, A dynamic programming approach to continuous speech recognition, in: Proc. the Seventh International Congress on Acoustics, 1971.

[10] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 26(1)(1978) 43-49.

[11] C.M. Salgado, M.C. Ferreira, S.M. Vieira, Mixed fuzzy clustering for misaligned time series, in: Proc. IEEE Transactions on Fuzzy Systems, 2016.

[12] A. Mueen, N. Chavoshi, N. Abu-El-Rub, H. Hamooni, A. Minnich, AWarp: fast warping distance for sparse time series, in: Proc. 2016 IEEE International Conference on Data Mining, 2016.

[13] Y. Sun, Z. Li, Clustering algorithm for time series based on locally extreme point, Computer Engineering 41(5)(2015) 33-37.

[14] J. Yang, J. Leskovec, Patterns of temporal variation in online media, in: Proc. the 4th ACM International Conference on Web Search and Data Mining, 2011.

[15] Z. Han, N. Chen, J. Le, D. Duan, J. Sun, An efficient and effective clustering algorithm for time series of hot topics, Chinese Journal of Computers 35(11)(2012) 2337-2347.

[16] Q. Liu, K. Wang, W. Rao, Non-equal time series clustering algorithm with sliding window STS distance, Journal of Frontiers of Computer Science and Technology 9(2015) 1301-1313.

[17] H. Zhou, J. Gao, Automatic method for determining cluster number based on silhouette coefficient, Advanced Research on Intelligent System, Mechanical Design Engineering and Information Engineering 951(2014) 227-230.

[18] S. Aranganayagi, K. Thangavel, Clustering categorical data using silhouette coefficient as a relocating measure, in: Proc. Conference on Computational Intelligence and Multimedia Applications, 2007.

[19] P. Lin, L. Liu, P. Nie, X. Zhu, Research on network public opinion warning index system based on feature analysis of the public opinion, Information Technology Journal 12(2013) 5326-5330.