Hindawi
Complexity
Volume 2017, Article ID 8130961, 10 pages
https://doi.org/10.1155/2017/8130961

Research Article

An Improved Clustering Method for Detection System of Public Security Events Based on Genetic Algorithm and Semisupervised Learning

Heng Wang,[1] Zhenzhen Zhao,[2] Zhiwei Guo,[3] Zhenfeng Wang,[1] and Guangyin Xu[1]

[1]*Collaborative Innovation Center of Biomass Energy, Henan Agricultural University, Henan 450002, China*
[2]*College of Computer and Information Engineering, Henan University of Economics and Law, Henan 450002, China*
[3]*College of Communication Engineering, Chongqing University, Chongqing 400044, China*

Correspondence should be addressed to Guangyin Xu; xucan264001@126.com

Received 27 March 2017; Accepted 8 May 2017; Published 18 June 2017

Academic Editor: Junhu Ruan

The occurrence of series of events is always associated with the news report, social network, and Internet media. In this paper, a detecting system for public security events is designed, which carries out clustering operation to cluster relevant text data, in order to benefit relevant departments by evaluation and handling. Firstly, texts are mapped into three-dimensional space using the vector space model. Then, to overcome the shortcoming of the traditional clustering algorithm, an improved fuzzy c-means (FCM) algorithm based on adaptive genetic algorithm and semisupervised learning is proposed. In the proposed algorithm, adaptive genetic algorithm is employed to select optimal initial clustering centers. Meanwhile, motivated by semisupervised learning, guiding effect of prior knowledge is used to accelerate iterative process. Finally, simulation experiments are conducted from two aspects of qualitative analysis and quantitative analysis, which demonstrate that the proposed algorithm performs excellently in improving clustering centers, clustering results, and consuming time.

1. Introduction

With the increasing development of communication and transmission mode, a growing number of media reports and public opinion concerning public security events are transmitting through a variety of ways, which brings to these events a higher level of interest and transparency. Various kinds of public security events, such as bus bombing, violence caused by extreme nationalism, and chopping people in campus, create serious damage to the life and property of broad masses of the people and bring about extremely bad influence to social stability and national unity.

The occurrence of a series of events is not occasional. Some are probably fermented individual behaviors influenced by some social factors such as massive media report, as well as inspiration and emotional impact brought by propaganda of the Internet public opinion. Once events caused by individual behaviors form a pattern, dangers in it are no less than any group events.

It is undeniable that in contemporary society the widespread use of the Internet, mobile network, and social platforms, providing much convenience to living of people, yet becomes hotbed of spread and flooding of various kings of negative information to some extent. Timely finding events perhaps causing significant disruption that are spreading through the Internet and other channels is a significant research subject.

Whether trying to solve the problem of data rich and lack of knowledge, or monitoring network public opinion, we need the help of topic detection and tracking (TDT) technology [1]. The research of TDT that addresses event-based organization of broadcast news has always been an issue in the field of natural language processing or data mining from the beginning. The goal of TDT is to detect new

topics from the news media information flow automatically and track existing topics dynamically. Through clustering method, information flow is aggregated into several classes, leading high similarity in one class and low similarity in different classes. With the high update speed of Internet news and its huge amount, the clustering algorithms used in TDT are generally on incremental way. In this domain, two approaches have been proposed, called single-pass algorithm and fuzzy C-means (FCM) algorithm [2].

Most of early researches concentrate on the selection of clustering algorithms. In [3], fuzzy k-member clustering is applied, to crowd movement analysis based on face image recognition with privacy consideration. Xiaolin et al. proposed an improved single-pass clustering algorithm for topic detection [4]. Researches on topic detection and tracking are developing rapidly. For example, Zhao et al. presented a Social Sentiment Sensor system on Sina Weibo to detect daily hot topics and analyzed the sentiment distributions toward these topics [5]. In [6], TDT is formulated as an online tracking, detection, and learning problem. By learning from historical data using semisupervised multiclass multifeature method, a topic tracker was obtained, which could also discover novel topics from the new stream data. However, most of the above-mentioned methods ignored some topic-indicative terms. References [7, 8] employed the named entity recognition technology in hot topic detection and proved that named entity can improve the performance of hot topic detection. For the inaccuracy of incremental clustering at the initial time, a TDT method of periodic classification and single-pass clustering is proposed [9]. In single-pass based incremental clustering algorithm, the similarities between the news reports and clustering centers of historical topics greatly affect the performance of the topic detection.

General clustering methods mainly contain partition-based clustering method, layer-based clustering method, density-based clustering method, and neural network-based clustering method. Among them, layer-based method FCM algorithm is common. The FCM algorithm [10] uses fuzzy logic where each data point is specified by a membership grade between 0 and 1. However, there are still some defects: (1) its performance relies on initialization of parameters; (2) it consumes a lot of time when data size is large. In order to overcome these shortcomings, a lot of scholars conduct deep research on it to improve parameters of some aspects and propose many new algorithms. For instance, the early research [11] proposed a brFCM algorithm that simplifies data set through quantification and aggregation to improve execution efficiency of FCM algorithm. Kolen and Hutcheson proposed a fuzzy c-means method that reduces time complexity [12]. But it does not consider the problem of clustering centers overlapping with sample points. In [13], a new, supervised, hierarchical clustering algorithm for fuzzy model identification is presented, which solves the problem of global model accuracy, together with the interpretability of local models as valid linearization of the modeled nonlinear system. For spectral clustering, its high computational complexity prevents its application to large-scale datasets. Cao et al. proposed an approximate spectral clustering method to address this complexity [14]. In [15], a hybrid fuzzy K-harmonic means (HFKHM) clustering algorithm based on improved possibilistic C-means clustering and K-harmonic means was presented, which solves the noise sensitivity problem of K-harmonic means and improves the memberships of the improved possibilistic C-Means clustering. In [16], Ding and Fu proposed a kernel-based fuzzy c-means clustering algorithm to optimize fuzzy c-means clustering, based on the genetic algorithm optimization which is combined with the improved genetic algorithm and the kernel technique.

In this paper, a system for detecting public security events is introduced to cater for the significance of this research, which integrates key technique of data mining and machine learning with field of discovery and tracking of hot topics. Then, the shortcomings of those single-pass-based methods, as well as problem of being prone to fall into local optimization of FCM algorithm, are formulated. To optimize the effectiveness of the system and to minimize deviation, an effective and practical framework of a modified FCM clustering method based on adaptive genetic algorithm and semisupervised learning (AG-SL-FCM) for public security events detection system is proposed. In the proposed algorithm, the adaptive genetic algorithm is employed to optimize the determination of initial clustering centers and applies guidance of prior knowledge in semisupervised learning to accelerate convergence of the algorithm. Specially, our main contributions are described in detail as follows.

(1) In order to realize the direction, accurate clustering for news reports is of significance. Therefore, a scheme of detection system for public security events is put forward, in which Chinese words are extracted approximate content and mapped as vector model in three-dimensional space. After preprocessing, clustering analysis operation is carried out on the system.

(2) With regard to the shortcomings of slow convergence speed and being prone to fall into local optimization, we divide the process of clustering analysis into two subproblems which are the determination of initial clustering centers and the iterative optimization of objective function. In the former one, the ability of global optimization of genetic algorithm is adopted to avoid it. In the latter one, idea of semisupervised learning is used to improve accuracy and speed the iterative process.

(3) The closed expression of detection scheme for the system, in which three keywords are used as three metric parameters, is derived. Then, through experiments, it can be proved that the system functions well with the use of the algorithm and the scheme proposed in this paper.

The rest of the paper is organized as follows. Section 2 presents the system model of the whole detection system. Section 3 describes the improved clustering algorithm in detail. Simulation results are given in Section 4. Finally, Section 5 provides the conclusion.

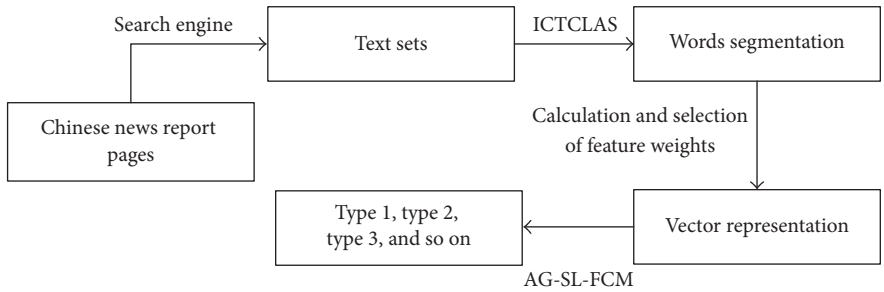

Figure 1: Flow chart of the system.

2. System Design and Model

In this section, we present detailed design of the detection system for public security events, as shown in Figure 1. Concrete steps are summarized as follows:

(1) The search engine based on semantics is used to collect news report from various news sites, tweets from microblog, and public opinions from forums.

(2) All of the above document streams are firstly partitioned into segments by a piece of software for words segmentation named ICTCLAS developed by Chinese Academy of Sciences which is one of the most popular tools used for words segmentation.

(3) The TF-IDF method [17] is adopted to calculate the weight of each keyword. And we use classical vector space model (VSM) [18] to extract main points of each sample. Then, the samples are mapped into three-dimensional space after dimensionality reduction.

(4) The algorithms of data mining adopted to cluster the samples act as the core part of the system and the main content of our research.

In the detecting system, the main model for research of each section is introduced as follows.

2.1. Search Engine Model. Web crawlers, as the main composition of the search engine, refers to an application aiming to crawl as fast as possible and collect as many webpages associated with predefined themes as possible. They can operate block collection on the whole Internet and integrate sampling results of different sectors together to improve acquisition coverage rate and page utilization rate of the whole Internet.

To overcome the disadvantages of traditional general web crawler, the capture scheme employed in this paper is based on dynamic topic base [19]. It can update the topic base intelligently during the crawling process, reduce the miscarriage justice of topic relevance of the anchor text, and filter URLs better. Meanwhile, in the process of search, the topic relativity judgement can be carried out on webpages and only needs to be based on the concept set not the whole Internet.

2.2. Text Representation Model. As is known, texts, as character strings formed by symbols with specific meaning that are connected in order, possess two basic characteristics, one of which is the frequency of all the words that constitute texts, and the other is connection order among these words, meaning that texts can be expressed by frequency and interrelation of feature items. In order to represent sequence information of feature items in texts, it is inevitable to utilize directed pointer structure. Thus the whole text would become a complicated graph or net. And one vector is enough to represent frequency information of characteristic items in text.

VSM, firstly put forward by Dr. Salton in 1968, uses vector to represent texts, which has always been the most classic calculation for text representation [17, 18]. Its idea originates from the fact that all the texts and queries contain some independent property that is expressed by some characteristic items to reveal their contents and that can be regarded as one dimension of vector space; thus the texts and queries can be expressed as the collection of these attributes ignoring complicated relation among paragraphs, sentences, and words. Text is expressed as a vector as follows:

$$v(d_t) = (w_1(d_t), w_2(d_t), \ldots, w_n(d_t)), \quad (1)$$

where n is the number of characteristic items when extracting characteristics and $w_i(d_j)$ is the weight of the ith characteristic items in document d_j. The frequency of characteristic words in a document is denoted as tf and the number of documents in which characteristic words occur is denoted as df. Thus, the formula of computing weight of characteristic words is deduced as follows (TF-IDF model):

$$w_i(d_j) = \frac{tf_{ij} \times \log_2(N/N_t + 0.01)}{\sqrt{\sum_{k=1}^{n} \left\{ \left(tf_{ij}\right)^2 \times \left[\log_2(N/N_t + 0.01)\right]^2 \right\}}}, \quad (2)$$

where f_{ij} is the frequency of the ith document occurring in d_j. N is total number of documents. N_i is the number of documents in which the ith characteristic item is occurring. Texts and queries can be expressed using a vector, respectively. And the similarity between them can be measured with distance among vectors.

2.3. Fuzzy C-Means Algorithm Model. Fuzzy C-means (FCM) [16], with a strict theoretical basis of fuzzy mathematics, is an unsupervised clustering algorithm based on a single objective function. Suppose the total number of samples waiting for

clustering analysis is N, C is the number of clusters, and u_{ik} denotes fuzzy membership of the ith sample point belonging to the kth cluster center. The membership degree matrix is given by the following formula:

$$U = [u_{ik}]_{C\times N}. \tag{3}$$

Each row element of matrix U represents membership of pixels belonging to each class. Adding fuzzy exponent, objective function, whose nature is the distance between each point and the clustering centers, is expressed as follows:

$$\begin{aligned} J(U,V) &= \sum_{K=1}^{N}\sum_{i=1}^{C} u_{ik}^{m} d_{ik}^{2}, \\ d_{ik} &= \left\| x_i - V_k \right\|, \end{aligned} \tag{4}$$

where d_{ik} is the Euclidean distance between the ith sample x_i and the kth clustering center V_k. The iterative cycle process in FCM algorithm is actually a process of seeking minimum value of $J_m(U,V)$. Typically, fuzzy exponent m takes the value of 2.

The fuzzy membership degree set is supposed to be established as follows:

$$u_{ik} = \frac{1}{\sum_{j=1}^{C}\left(d_{ik}/d_{ij}\right)^{2/(m-1)}}, \quad 1 \le i \le C,\ 1 \le k \le N. \tag{5}$$

The clustering center adjustment formula is defined as follows:

$$V_K = \frac{\sum_{i=1}^{n}\left(u_{ik}\right)^{m} x_i}{\sum_{i=1}^{n}\left(u_{ik}\right)^{m}}, \quad 1 \le i \le C. \tag{6}$$

The two steps mentioned above, evaluation of fuzzy membership and recomputation of cluster centers, should be executed several times until there is no change in cluster centers and are the necessary condition of $J(U,V)$ tending to the maximum value which can be proved utilizing the Lagrange multiplier.

3. Proposed Framework

Our research combines idea of adaptive genetic algorithm and semisupervised learning with FCM algorithm to make an improved FCM algorithm: AG-SL-FCM algorithm. It can make full use of high efficiency of FCM algorithm, global search ability of genetic algorithm, and guidance of prior knowledge, which ensures efficiency and precision of clustering.

3.1. The Determination of Initial Clustering Centers Based on Adaptive Genetic Algorithm. The basic idea of adaptive genetic algorithm (AGA) [20] can be summarized as follows: Firstly, it is supposed to produce the initial solution group; then, it is expected to select excellent individuals from solution group according to some index. Then, genetic operators are operated on them to generate a new generation of candidate group. Finally, this process is repeated until the satisfaction of some convergence index. Compared with many other optimization algorithms, genetic algorithm, possessing the obvious global search performance and the robustness of problem solving, has demonstrated its unique advantages and is thus widely used in problems with regard to combinatorial optimization, pattern recognition, machine learning, and image processing.

The formulated problem of being prone to fall into local optimum can be effectively resolved by AGA. As a directed random search technique, AGA adopts natural evolution strategy and works out solution by continually evolving a population of candidate solutions [21]. And the evolution process mainly includes selection, crossover, and mutation. Therefore, a method of determination of the initial clustering centers based on AGA is proposed to work as an optimal global searching method of the system to resolve the problem. The key components of the method are depicted as follows.

FCM algorithm is, respectively, used on the m subsets to split each subset into c categories and to compute their clustering centers V. V^r is the clustering center of the rth subset. $(V^1, V^2, \ldots, V^m)$ is the initial solution group of genetic algorithm; namely, the population size is m and individuals are clustering centers of each subset. The idea of genetic algorithm is used to optimize the population group for obtaining the clustering centers close to global optimum.

(1) Initial Population. Assume that the training set has n samples, number of clustering centers is $P(t)$, and each sample is a k-dimensional vector. These n samples are evenly split into m subsets in which the distribution of different categories is consistent with the original samples. The initial population is a set of potential solutions of which size of individuals is denoted as m, and then the initial population E_0 is represented as follows:

$$E_0 = \left\{V^1, V^2, \ldots, V^m\right\}, \tag{7}$$

where V^r, a matrix with dimension of $c \times k$, is the clustering center of the rth subset (individual).

(2) Encoding. Population individuals constitute a matrix with dimension of $c \times k$ arranged around clustering centers. In this paper, to avoid complexity and improve efficiency, real number coding strategy and concatenated code form are adopted to link c sets of parameters representing clustering centers, which contributes to shortening the length of chromosome and to improving ability of global optimization as well as convergence speed of the algorithm. The result of encoding clustering centers is represented as follows:

$$\begin{aligned} E_0 &= \left\{V^1, V^2, \ldots, V^m\right\} \\ &= \left\{v_{11}, v_{12}, \ldots, v_{1k}, \ldots, v_{m1}, \ldots, v_{mk}\right\}, \end{aligned} \tag{8}$$

where v_{ij} is jth component of V^i.

(3) Fitness Calculation. Fitness is employed to evaluate the performance of each individual in the population, meaning that a better individual returns a higher fitness. As a performance measurement criterion, the fitness function

establishes a mapping of value of the objective function for fitness, which plays an important role in the evolutionary process. In terms of FCM algorithm, optimal clustering results correspond to minimal value of objective function. Motivated by simulated annealing thinking [22], the fitness function carrying out the scale transformation is defined by

$$f = \frac{1}{1 + \sum_{k=1}^{N} \sum_{i=1}^{C} W(g)\, u_{ik}^{m} d_{ik}^{2}}, \tag{9}$$

$$W(g) = w_0 \cdot a_w^g, \tag{10}$$

where $W(g)$ is the annealing temperature function of the evolutional generation and g is the current generation number. In (10), w_0 is the initial annealing temperature and equals $2 \cdot g_{\max}$; $g_{\max}$ is the maximum number of generations; a_w denotes the annealing temperature coefficient whose value is slightly less than 1. With the number of evolutional generations increasing, the differences of the fitness between individuals in the same generation become more evident, thus providing more opportunities to extract better individuals.

(4) Survivor Selection. Selection strategy, playing an important role in algorithm performance, adopts the rule of survival of the fittest in the evolution theory. It says that high potential individuals (parents) will produce better ones (offspring). Individuals in the population are selected to undergo crossover and mutation operations for reproduction using the roulette wheel selection [23]. This selection method is based on the distribution and obtained by (9). According to the theory of roulette wheel selection, a better individual should have a higher chance to be selected. Consequently, after survivor selection, the better individuals survived and the worse ones are eliminated.

In this approach, elite individuals are retained so that they would not take part in crossover or variation operation, but directly enter next generation. For poor individuals, they also do not take part in crossover operation but take part in variation operation, with the probability of variation being higher than normal individuals. Then, computing probability distribution corresponding to fitness function and extracting other individuals of current group to attach crossover and variation operation on them are expected, for improving the average fitness of the group. The selection probability function is defined as follows:

$$\text{Rate}\left(V^i\right) = \frac{f(i)}{\sum_{j=1}^{m} f(j)}, \tag{11}$$

where $f(i)$ is fitness value of individual V^i.

(5) Crossover Operation. Crossover operation is employed to reproduce new individuals (offspring) by swapping segments of genetic information of two individuals (parents). In AGA, an adaptive crossover strategy is used to improve the search performance of the algorithm. It is realized by assigning a crossover probability to each crossover operation. The adaptive crossover probability of individuals is defined by

$$p_c = \begin{cases} k_1' - \left(k_1' - k_2'\right) \dfrac{\left(F' - \overline{F}\right)}{F_{\max} - \overline{F}} & F' \geq \overline{F} \\ k_1' & F' < \overline{F}, \end{cases} \tag{12}$$

where $F_{\max}$ and $\overline{F}$ are, respectively, the maximum and the average fitness of the current generation; F' denotes the higher fitness between the two selected individuals; k_1' and k_2' are coefficients fixed at the initialization. When F' is higher than $\overline{F}$, the lower crossover probability is adopted to reduce the chance of destruction of excellent individual. When F' is lower than $\overline{F}$, the higher crossover probability is used to raise the chance of emergence of new and improved individuals. The crossover operation used in the proposed algorithm is arithmetic crossover [24].

Assume that (V^a, V^b) is parent crossover pair and carried out crossover operation to produce the jth bit of offspring $V^{a\prime}$ and $V^{b\prime}$ according to following formula:

$$\begin{aligned} V_i^{a\prime} &= p_c v_i^a + (1 - p_c)\, v_i^b, \\ V_i^{b\prime} &= p_c v_i^b + (1 - p_c)\, v_i^a, \end{aligned} \tag{13}$$

where the number of crossover bits among chromosomes is a random integer in the range of $[1, ck]$.

(6) Mutation Operation. Similar to the gene mutation in genetics, mutation operation is used to change some genetic information of an individual. Each individual in the population has a possibility of mutation. A low probability may hinder the production of new individuals, which is not conductive to the worse individual evolving. However, a high possibility of mutation reduces the searching performance of the random searching algorithm and may go against retaining the better. Thus, an adaptive mutation probability is employed to ensure the searching ability. The adaptive mutation probability of individual V^i is defined by

$$p_m = \begin{cases} k_3' - \left(k_3' - k_4'\right) \dfrac{\left(F_i - \overline{F}\right)}{F_{\max} - \overline{F}} & F_i \geq \overline{F} \\ k_3' & F_i < \overline{F}, \end{cases} \tag{14}$$

where F_i is the fitness of the individual V^i; k_3' and k_4' are predefined parameters. When the fitness of an individual is higher than the average fitness of the current generation, the lower mutation probability is adopted. And when the fitness of the individual is lower than the average fitness, the higher mutation probability is used. In the proposed algorithm, the mutation operation is Gaussian mutation [25]. Each individual in the population should undergo the mutation operation.

Assume that V^a is a mutational individual and it carried out mutation operation according to the following formula to produce the ith bit of offspring $V^{a\prime}$:

$$V_i^{a\prime} = p_m V_i^a + (1 - p_m)\, V_i^b, \tag{15}$$

where b is a random integer except a in the range of $[1, m]$. And bit number of chromosome mutation is a random integer in the range of $[1, ck]$.

After the operation of selection, crossover, and mutation, a new population is generated. The same operations will be repeated until the maximum number of evolutional generations is reached. Finally, the clustering centers close to real optimal clustering centers are obtained. And the obtained clustering centers are initial value for FCM algorithm that is used on the original n samples to obtain global optimal clustering centers for the subsequent process of semisupervised iteration.

In summary, with genetic algorithm integrated into FCM algorithm, it is able to obtain basic steps of the determination of initial clustering centers in AG-SL-FCM algorithm:

(1) Give some genetic algorithm parameters, such as the clustering number c, the population size m, the crossover probability, the variation probability, and maximum generation $T_{\max}$.

(2) Set up the evolution algebra counter $t = 0$. The train set with sample size of n is divided into m subsets of which clustering centers can be computed using fuzzy C-means, resulting in m population individuals $P(t)$.

(3) Set up individual objective function and carry out fitness evaluation.

(4) Heredity operation includes selection, crossover, and mutation.

(5) Compute fitness of offspring and add it to the parent group. Then, remove individuals with low fitness.

(6) If the setting evolutional generation is achieved, it is supposed to output individuals with the highest fitness in current population group and end the algorithm when reaching maximum generation $T_{\max}$. Otherwise, return to step (4) to continue evolving.

3.2. The Process of FCM Iteration Based on Semisupervised Learning. In this paper, influence of prior knowledge is used to improve basic unsupervised FCM algorithm, which belongs to the domain of semisupervised learning. α ($\alpha \in [0, 100]$) is the proportion of known samples accounting for all samples. Then, this part of prior knowledge is as guidance signal to accelerate the convergence speed of cyclic iteration of the algorithm, thus obtaining the part of modified method based on semisupervised learning [26].

The main idea of semisupervised method is the process of utilizing information of known samples for optimization. By means of adding effect of this part of prior knowledge, convergence process of the algorithm is accelerated.

At first, the distance from samples of prior knowledge to initial clustering centers is denoted as

$$\eta\left(x_i', V_k\right) = \left\|x_i' - V_k\right\|_2. \tag{16}$$

Then, membership of each sample of prior knowledge is demonstrated as

$$u_{ik}' = \begin{cases} 1; & \eta\left(x_i', V_k\right) = 0 \\ \dfrac{1}{\sum_{i=1}^{C}\left(\eta\left(x_i', V_k\right)/\eta\left(x_i', V_m\right)\right)} & \text{otherwise.} \end{cases} \tag{17}$$

Thus the new objective function can be established

$$J = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2} + \alpha \sum_{i=1}^{C}\sum_{k=1}^{N}\left(u_{ik} - u_{ik}' b_i\right)^{m} d_{ik}^{2}, \tag{18}$$

where b_i is represented using Boolean logic as follows:

$$x_i = \begin{cases} 1, & x_i \text{ is prior knowledge} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

In (17), u_{ik}' is the membership relation among this part of prior knowledge computed. u_{ik} is the membership value of $g = 0$ in the whole cycle. The impact factor of prior knowledge α is proportional to the ratio of amount of prior knowledge M to the total number of samples. d_{ik}, which is defined as the Mahalanobis distance [27] from sample x_i to clustering center V_k, which can be expressed as

$$d_{ik} = \left(x_i - V_k\right)^{T} M_i \left(x_i - V_k\right), \tag{20}$$

where M_i is the association correlation matrix among characteristic vector of each sample. We introduce Lagrange operator and then calculate the minimal value of objective function J without constraints; the membership degree matrix can be established as follows:

$$u_{ij} = \frac{1}{1+\alpha}\left\{\frac{1 + \alpha\left(1 - b_j \sum_{i=1}^{C} u_{ij}'\right)}{\sum_{i=1}^{C}\left(d_{kj}/d_{ij}\right)^{2}} + \alpha u_{ik}' b_j\right\}, \tag{21}$$

where α is the impact factor of prior knowledge, which is proportional to the ratio of amount of prior knowledge M to the total number of samples.

Adjustment formula of clustering centers is defined as follows:

$$V_k = \frac{\sum_{k=1}^{C}\left(u_{ik}\right)^{m} x_i}{\sum_{k=1}^{C}\left(u_{ik}\right)^{m}}, \quad 1 \le i \le C, \tag{22}$$

where C is the number of clusters and m is the fuzzy exponent and equals 2.

Therefore, the whole AG-SL-FCM algorithm is described in Algorithm 1.

4. Simulation Results and Analysis

In order to verify the effectiveness of the detection system proposed in this paper, we compare the AG-SL-FCM algorithm with traditional FCM algorithm [16] through experiments. The simulation was carried out on the computer with the Intel Pentium dual CPU (3.20 GHz), and the proposed algorithm is simulated in MATLAB environment.

Input $k_1', k_2', k_3', k_4', a_w$, n samples, number of subsets m, the number of clustering centers c and population size L;
Initialize population E_0, generation number $g = 0$;
Calculate the fitness of each individual in the initial population E_0 according to Eq. (8);
For all $l = 1$ to L do
  Select survivor according to the wheel selection rule [23];
End for
For all $\beta \in (0, 1)$ to $g_{\max}$ do
  For all $l = 1$ to L do
    Select randomly two individuals and calculate the crossover probability p_c according to Eq. (12);
    Generate a random floating-point number β, $\beta \in (0, 1)$;
    If $\beta \leq p_c$ then
      Perform arithmetic crossover operation [24];
    End if
  End for
  For all $l = 1$ to L do
    Calculate the mutation probability p_m according to Eq. (14);
    Generate a random floating-point number β, $\beta \in (0, 1)$;
    If $\beta \leq p_m$ then
      Perform Gaussian mutation operation [25];
    End if
  End for
  Calculate the fitness of each individual in the new population E_{g+1};
  If $g = g_{\max}$ then
    Return individuals with the highest fitness;
  Else if
    Return to selecting survivor;
  End if
End for
Output final population E_L
Initialize obtained the excellent clustering centers, the number of iterations I, the error log *error*, and the cut-off error ε;
Establish new objective function according to Eq. (16), (17) and (18);
For $I = 0$ to *Loop*
  If *error* $> \varepsilon$
    Establish new membership degree matrix according to Eq. (21);
    Obtain new formula of clustering centers according to Eq. (22);
    $U^{(I+1)} = [U_{ik}^{(I+1)}]$, $error = |U^{(I+1)} - U^I|$;
  Else if
    Convergence;
  End if
End for

Algorithm 1: The AG-SL-FCM algorithm.

In this paper, we randomly select 2000 entries as experimental dataset from a dataset released by Hylanda Information Technology Co. Ltd. in Tianjin of China which is comprised of almost all news report containing keywords of public security events. In the dataset, the public security events are divided into three types: violent terrorist attacks, campus attacks, and explosions. The selected experimental dataset has been classified manually. Each category contains at least one training set and one test set. Artificial markers of training sets are supposed to be eliminated before clustering. In our research, with regard to the comparison of clustering result, simulation experiments of both algorithms are, respectively, conducted. Two evaluation indicators which are precision ratio and recall ratio, respectively, are defined as parameters to evaluate results.

First of all, in order to analyze public opinion trend, it is supposed to use the Chinese words segmentation system ICTCLAS to separate words. Secondly, the weight of each keyword is computed through (2). And then, texts are mapped as a character vector through vector space model (VSM). After the dataset vectorized, the clustering calculation is able to be carried out.

4.1. Qualitative Analysis. It is known from prior knowledge that there are three kinds of topics in the dataset. After clustering, different colors are assigned to three categories to improve visual readability.

Table 1: The clustering center of FCM and AG-SL-FCM.

	FCM			AG-SL-FCM		
	Word 1	Word 2	Word 3	Word 1	Word 2	Word 3
Category 1	0.3021	0.6045	0.3108	0.3162	0.6892	0.3083
Category 2	0.2708	0.3728	0.6498	0.2823	0.3312	0.6543
Category 3	0.5059	0.4896	0.5339	0.5057	0.5002	0.5023

(Category 1: violent terrorist attacks; Category 2: campus attacks; Category 3: explosion).

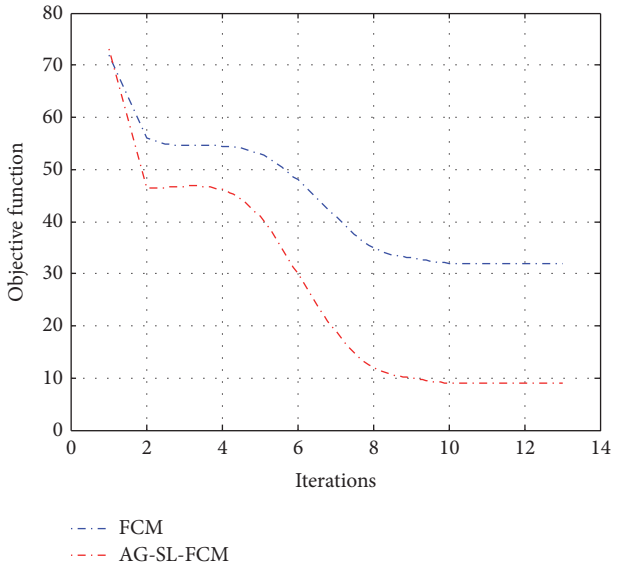

Figure 2: The convergence process of objective function in FCM algorithm and AG-SL-FCM algorithm.

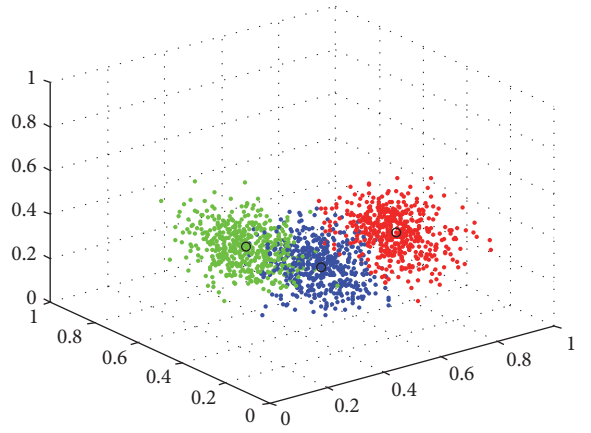

Figure 3: The distribution of sample sets in AG-SL-FCM algorithm under the influence of the three parameters.

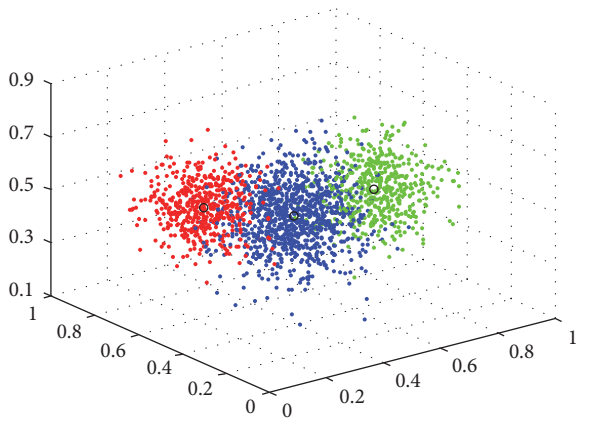

Figure 4: The distribution of sample sets in FCM algorithm under the influence of the three parameters.

At first, FCM algorithm and AG-SL-FCM algorithm are adopted to operate on the experimental dataset, respectively. Three keywords are defined as metric parameters, in which word 1 is violent terrorist attacks, word 2 is campus cutting, and word 3 is explosion. The obtained clustering centers are shown in Table 1. Through prior knowledge we can verify that the sum of variance of known samples in AG-SL-FCM is less. Therefore, after optimization of genetic algorithm, the distribution of clustering centers in AG-SL-FCM algorithm is more ideal.

The convergence and the iterative process for the two algorithms are shown in Figure 2. It indicates that the convergence speed of the AG-SL-FCM algorithm is only little slower than of FCM algorithm. When determining initial clustering centers, iterative process of genetic algorithm increases the time consumption. However, the guidance of prior knowledge in the optimization process of objective function is accelerated. Therefore, the iterative speed does not obviously slow down.

From Figures 3–7, the different colors are adopted to represent different topic categories, in which red represents violent terrorist attacks, green represents campus chopping, and blue represents explosion. It shows that three keywords are adopted as characteristic parameters to judge events categories. With comprehensive analysis of these three parameters, 2000 sample points of public security events are divided into three categories so as to be judged visually and correctly. Figures 3 and 4 show the clustering results of FCM algorithm and AG-SL-FCM algorithm, respectively, in three-dimensional space, in which three coordinate axes represent the weight of three indicators computed through equations in Section 2. Experimental results show that obtained clustering centers of AG-SL-FCM algorithm are more ideal than FCM algorithm, which means that AG-SL-FCM algorithm can obtain better clustering results than FCM algorithm.

In Figures 5, 6, and 7, as the effectiveness of AG-SL-FCM has been testified in the previous experiments, further

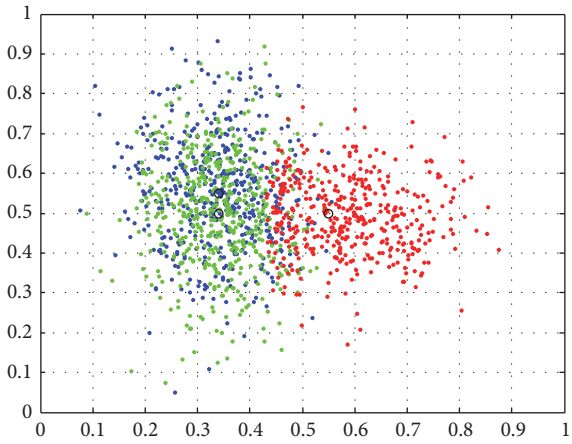

Figure 5: The distribution of sample sets in AG-SL-FCM algorithm under the influence of "violent terrorist attacks" and "campus chopping."

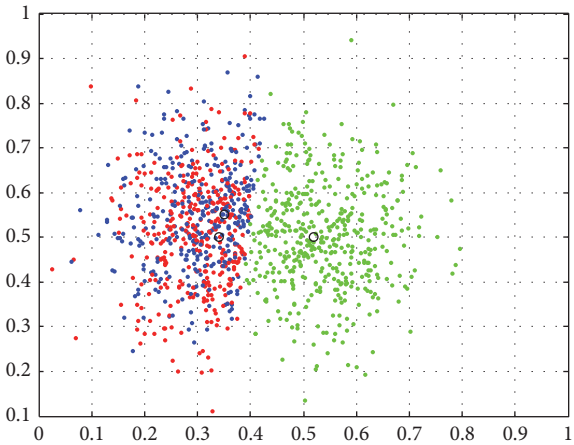

Figure 6: The distribution of sample sets in AG-SL-FCM algorithm under the influence of "explosion" and "campus chopping."

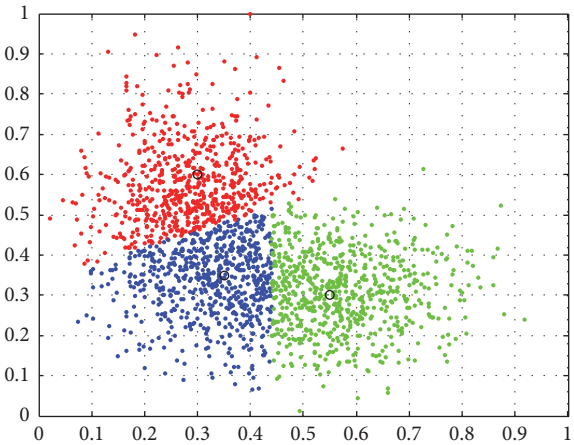

Figure 7: The distribution of sample sets in AG-SL-FCM algorithm under the influence of "violent terrorist attacks" and "explosion."

research is conducted on AG-SL-FCM in two-dimensional space. In Figures 6 and 7, it can be concluded that due to the complexity and semantic relevance of Chinese language (in Chinese language, the scope included by violent terrorist overlaps with the other two parameters partly) the classification of samples differs when using different indicators and that mainly on the basis of parameters "violent terrorist" and "explosion" we can get a relatively accurate determination of events types, while under the influence of indicators "violent terrorist" and "campus cutting" the detection results remain dislocated. In Figure 7, we can also conclude that in the axis of "violent terrorist attacks" when there are a huge number of points gathering in the range of 0.3 to 0.6, it maybe implies that some events have caused no small impact on public opinion, which is latent for inducing some unstable fact especially in such a country, China, that malignant events caused by the nationalism occurred from time to time.

5. Conclusions

In this paper, an improved clustering method based on genetic algorithm and semisupervised learning (AG-SL-FCM) is proposed for a detection system of public security events. In the proposed algorithm, adaptive genetic algorithm is employed to select optimal initial clustering centers. Meanwhile, motivated by semisupervised learning, guiding effect of prior knowledge is used to accelerate iterative process. Result of simulation reveals that compared with FCM algorithm the improved algorithm can not only effectively realize fuzzy clustering of data but also have relatively good convergence speed. With the use of this algorithm, the proposed detection system can analyze discrimination of types from massive news report and blog message, improve efficiency of detecting spread of public opinion associated with public security events, and provide relative departments with a reliable basis for preventing and handling of public security events.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work has been partly supported by Henan Provincial Department of Science and Technology Research Project (172102210307), Henan Province Institution of Higher Learning Youth Backbone Teachers Training Program (2016GGJS-036), and Key Science Research Program of Henan Province (17A480004, 16A413010).

References

[1] K.-L. Nguyen, B.-J. Shin, and S. J. Yoo, "Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information," in *Proceedings of the International Conference on Big Data and Smart Computing, BigComp 2016*, pp. 223–230, China, 2016.

[2] J. H. Ruan, X. P. Wang, F. T. S. Chan, and Y. Shi, "Optimizing the intermodal transportation of emergency medical supplies using balanced fuzzy clustering," *International Journal of Production Research*, vol. 54, no. 14, pp. 4368–4386, 2016.

[3] K. Honda, M. Omori, S. Ubukata, and A. Notsu, "Fuzzy clustering-based k-anonymization of eigen-face features for crowd movement analysis with privacy consideration," *International Journal of Innovative Computing, Information and Control*, vol. 12, no. 4, pp. 1375–1384, 2016.

[4] Y. Xiaolin, Z. Xiao, K. Nan, and Z. Fengchao, "An improved Single-Pass clustering algorithm internet-oriented network topic detection," in *Proceedings of the 4th International Conference on Intelligent Control and Information Processing (ICICIP '13)*, pp. 560–564, Beijing, China, 2013.

[5] Y. Zhao, B. Qin, T. Liu, and D. Tang, "Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 8843–8860, 2016.

[6] G. Li, S. Jiang, W. Zhang, J. Pang, and Q. Huang, "Online web video topic detection and tracking with semi-supervised learning," *Multimedia Systems*, vol. 22, no. 1, pp. 115–125, 2016.

[7] H. Li, G. Chen, T. Huang, and Z. Dong, "High-performance consensus control in networked systems with limited bandwidth communication and time-varying directed topologies," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2016.

[8] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[9] X.-Y. Dai, Q.-C. Chen, X.-L. Wang, and J. Xu, "Online topic detection and tracking of financial news based on hierarchical clustering," in *Proceedings of the International Conference on Machine Learning and Cybernetics, ICMLC 2010*, pp. 3341–3346, Qingdao, China, 2010.

[10] Y. Wang, Y. Wang, D. Yu, J. Yu, and F. C. Lau, "Information exchange with collision detection on multiple channels," *Journal of Combinatorial Optimization*, vol. 31, no. 1, pp. 118–135, 2016.

[11] H. Li, G. Chen, T. Huang, Z. Dong, W. Zhu, and L. Gao, "Event-triggered distributed average consensus over directed digital networks with limited communication bandwidth," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3098–3110, 2016.

[12] J. F. Kolen and T. Hutcheson, "Reducing the time complexity of the fuzzy *c*-means algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 263–267, 2002.

[13] B. Hartmann, O. Bänfer, O. Nelles, A. Sodja, L. Teslić, and I. Škrjanc, "Supervised hierarchical clustering in fuzzy model identification," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 6, pp. 1163–1176, 2011.

[14] J. Cao, P. Chen, Q. Dai, and W.-K. Ling, "Local information-based fast approximate spectral clustering," *Pattern Recognition Letters*, vol. 38, no. 1, pp. 63–69, 2014.

[15] X. Wu, B. Wu, J. Sun, S. Qiu, and X. Li, "A hybrid fuzzy K-harmonic means clustering algorithm," *Applied Mathematical Modelling*, vol. 39, no. 12, pp. 3398–3409, 2015.

[16] Y. Ding and X. Fu, "Kernel-based fuzzy *c*-means clustering algorithm based on genetic algorithm," *Neurocomputing*, vol. 188, pp. 233–238, 2016.

[17] R. Ñanculef, I. Flaounas, and N. Cristianini, "Efficient classification of multi-labeled text streams by clashing," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5431–5450, 2014.

[18] Y. Du, W. Liu, X. Lv, and G. Peng, "An improved focused crawler based on semantic similarity vector space model," *Applied Soft Computing Journal*, vol. 36, pp. 392–407, 2015.

[19] J. Akbari Torkestani, "An adaptive focused Web crawling algorithm based on learning automata," *Applied Intelligence*, pp. 1–16, 2012.

[20] X. Tao, S. Wang, Y. Huangfu, S. Wang, and Y. Wang, "Geometry and power optimization of coilgun based on adaptive genetic algorithms," *IEEE Transactions on Plasma Science*, vol. 43, no. 5, pp. 1208–1214, 2015.

[21] H. Wang, H. Li, C. Tang et al., "Modeling, metrics, and optimal design for solar energy-powered base station system," *Eurasip Journal on Wireless Communications and Networking*, vol. 2015, no. 1, 2015.

[22] P. Siarry, "Simulated Annealing," in *Metaheuristics*, Springer International Publishing, 2016.

[23] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, New York, NY, USA, 1996.

[24] Z. Michalewicz, C. Z. Janikow, and J. B. Krawczyk, "A modified genetic algorithm for optimal control problems," *Computers and Mathematics with Applications*, vol. 23, no. 12, pp. 83–94, 1992.

[25] S. N. Sivanandam and S. N. Deepa, *Introduction to Genetic Algorithms*, Springer, Berlin, Germany, 2007.

[26] M. Amini and N. Usunier, *Learning with Partially Labeled and Interdependent Data*, Springer International Publishing, 2015.

[27] E. Fetaya and S. Ullman, "Learning local invariant mahalanobis distances," *Computer Science*, pp. 162–168, 2015.