future internet

Article

Research on Community Detection of Online Social Network Members Based on the Sparse Subspace Clustering Approach

Zihe Zhou [1] and Bo Tian [2,*]

[1] College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; 981101zzh@nuaa.edu.cn
[2] School of Information Management & Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
* Correspondence: tian.bo@mail.shufe.edu.cn

Received: 23 October 2019; Accepted: 5 December 2019; Published: 9 December 2019

Abstract: The text data of the social network platforms take the form of short texts, and the massive text data have high-dimensional and sparse characteristics, which does not make the traditional clustering algorithm perform well. In this paper, a new community detection method based on the sparse subspace clustering (SSC) algorithm is proposed to deal with the problem of sparsity and the high-dimensional characteristic of short texts in online social networks. The main ideal is as follows. First, the structured data including users' attributions and user behavior and unstructured data such as user reviews are used to construct the vector space for the network. And the similarity of the feature words is calculated by the location relation of the feature words in the synonym word forest. Then, the dimensions of data are deduced based on the principal component analysis in order to improve the clustering accuracy. Further, a new community detection method of social network members based on the SSC is proposed. Finally, experiments on several data sets are performed and compared with the K-means clustering algorithm. Experimental results show that proper dimension reduction for high dimensional data can improve the clustering accuracy and efficiency of the SSC approach. The proposed method can achieve suitable community partition effect on online social network data sets.

Keywords: sparse subspace clustering; community detection; microblog text analysis; online social network

1. Introduction

Smartphones are becoming the main way for people to receive information. Mobile applications are roughly divided into four categories: instant messaging, search engines, online news, and social applications [1]. Among them, social applications have the advantages of convenient and simple operation and have become the main way for people to communicate remotely. The interaction between users forms a virtual social structure based on the Internet platform, which is composed of the directed relationship between individuals [2]. Based on the interpersonal network relationship, social media such as WeChat and Microblog strengthen the user viscosity and play an important guiding role in the process of consumers' access to shopping information, which provides a good solution for e-commerce promotion and reduces the cost of the business. Based on the increased relationship flow and information flow of the platform, the Microblog network has the advantages of a large user base, high user activity, and user interest orientation. Users can independently pay attention to the home page of interest and express their views. The instantaneity and openness of the information platform promotes the gathering of users with common values and interests, and user communities will form gradually [3]. The current community detection methods usually divide users based on their basic attributes (age, gender, educational level, place of birth, etc.), which do not reveal the characteristics of user groups based on product information [4].

The first step to grasp the topics of users' interest and to achieve accurate marketing is the hierarchical management of social network users. This paper takes Microblog as example to research the social network clustering structure of massive and high-dimensional user data. Text mining is the discovery of textual semantics and potentially valuable knowledge from the large-scale text set to assist people in making decisions. The large-scale short text on the mobile social platform has loud data noise and sparsity compared with the conventional text on the traditional Internet platform web page, which increases the difficulty of text analysis [3,4]. Traditional text clustering methods cannot accurately reflect the similarity of texts by calculating the distance of high-dimensional sparse feature sets because of the curse of dimensionality [5,6]. An effective method to solve high-dimensional problems in documents is to transform or select features and reduce the dimension, then cluster the refined data. This kind of method is based on the correlation hypothesis between features, which will lose the original information of the data, affect the clustering results to some extent, and the interpretability of the clustering results is difficult [7]. As a research focus of cluster analysis in recent years, sparse subspace clustering (SSC) is suitable for sparse text analysis [8,9]. Based on the semantic features of the text, this paper proposes a new online social e-commerce consumer grouping detection method based on SSC, which may solve the curse of dimensionality and sparsity in short text clustering to some extent effectively.

2. Literature Review

This paper researches the user community discovery method for online social networks based on SSC. Subsequently, three relative aspects are summarized from text representation, user interest community research, and subspace clustering theory.

The primary problem of text clustering is to transform the unstructured data into information that can be processed by computational method. At present, the vector space model is usually used in the text representation. Document sets are represented as vector sets in space, and the similarity of document semantics is transformed into the similarity of space vectors [7]. The semantics of the words in the document constitutes the meaning of the whole document. Each word is a feature item, and the importance of each word is reflected by the weight of the featured item. Boolean function, frequency function, and term frequency-inverse document frequency (TF-IDF) function are commonly used to calculate the weight of feature words. The TF-IDF function is commonly used because of its accuracy and simplicity. Sahami and Heilman proposed a short text similarity measure based on Web semantic kernel function, which calculated the similarity between short texts by adding semantic information of text features into the kernel function [10]. Yih extended Sahami's Web semantic kernel function to obtain Web-related similarity measures [11]. Li Tiancai calculated the weight of features in each text and extracted keywords to represent the text, but ignored the feature words with lower frequency in the concise text [12]. Yang Bin improved the weight of traditional TF-IDF function and added a synonym to solve the sparse problem of short text [13]. Yan Chao and other scholars proposed a user-based LDA （Latent Dirichlet Allocation）topic model to classify Microblog short texts [14]. With the development of mobile networks, researchers focus more on clustering analysis of short texts. Peng Min transformed the short text clustering problem into frequent item set clustering and retrieved short text based on subject phrases to achieve massive short text clustering more accurately [5]. The accuracy of the existing clustering algorithm is usually low in the high-dimensional sparse short text. To deal with this problem, Li Xiaohong constructed the unweighted graph of the short text set and the document similarity matrix was obtained [15]. It can bring about storage and computation difficulty if the spectral clustering algorithm is used directly in large data sets [6,16,17]. Therefore, a new method for online social network user community detection is proposed based on compressed sensing theory and SSC in this paper.

Users with high degrees of relevance in online social networks usually aggregate to constitute a group [18]. Microblog is a real-time medium, and user interests mainly stem from users interested in friends and acquaintances. The connection between any two people does not mean the interaction between them, and the link structure of the social network cannot reveal the actual interaction between people. So it is necessary to find the hidden social network in the Microblog network. Lei Bing built a Microblog user interest model of user attributes, keywords, and user behaviors [3]. Xu Zhiming personalized user recommendations and achieved good results from the users' personal basic information, released content, tags, and the number of followers as features reflecting user interests [19]. At present, there are three main research directions in the online community discovery: (1) Based on user attributes and interaction relations, the hidden links between users are discovered, and the clustering is transformed into graph partitioning problem; (2) When mining the text of a user's Microblog content or comments, user clustering based on characteristics of interest is realized; (3) Considering a Microblog user's contact and text content, content-based user interest community discovery and interaction-based user contact community discovery are realized separately, and then the two communities are integrated to form a user community with interest and network structure. For example, Yang Kai described the Microblog user relationship in the form of a network which depicts user behavior and influence in the network according to changes in network structural attributes [20]. Wang Yongcheng proposed a user association mining algorithm based on spectral clustering [21]. Sun Yifan proposed the similarity of Microblog users based on common follows and common followers [22]. This paper integrates user structured data and unstructured data to construct eigenvectors in vector space to achieve user clustering structure in the network.

There are mainly two kinds of graphs—modular structure graphs and random intersection graphs. In the modular graphs such as social network, biological modular, and physical systems, the points in the networks are not distributed randomly. The problem of community detection is proposed against this background to deal with the modular structure of some networks. Subspace clustering is a new type of clustering algorithm for high-dimensional data. The idea is that high-dimensional data belongs to a low-dimensional subspace in essence [23]. According to the different subspace search strategies, it can be divided into top–down search strategies and bottom–up search strategies. Subspace clustering algorithms are mainly divided into five types: matrix decomposition-based methods, iterative methods, algebraic methods, statistical methods, and spectral clustering-based methods [24]. The first four kinds of methods must set the number of subspaces and their dimensions in advance, and they are sensitive to the outline points in the data set. The spectral clustering method transforms the clustering problem into the optimal graph partition problem by using the relationship between data points, without knowing the number and dimension of subspaces [25,26]. SSC is a subspace spectral clustering algorithm based on compressed sensing theory. It uses sparse representation theory to represent data as a linear combination of other data in the same subspace, constructs the similarity matrix through the sparse coefficient matrix, and clusters data with a spectral clustering algorithm [27–29]. The SSC optimization models can be solved by different approaches such as the alternating direction method of multipliers (ADMM), orthogonal matching pursuit (OMP), etc. [24,27]. In recent years, scholars have proposed different optimization strategies and wider applications of SSC algorithms. In this paper, SSC is introduced into an online social network community discovery, which expands the research field of this method.

The research on the clustering of short text users in mobile social networks has received extensive attention, Main problems that need to be solved in short text research are high- dimensional features and sparsity of short text. In order to deal with this problem, this paper proposes a new method of social network member community discovery based on SSC, which collects users' structured data (including user attributes and user behaviors) and unstructured data (user comments) to construct a vector space model and establish the similarity of feature words. The matrix projection method is then applied to the high-dimensional data to reduce the dimension. Then, a new community discovery model for short text based on SSC is proposed. At last, experimental results show that the proposed method is an effective way to deal with online short text.

3. User Similarity Measurement for Short Text

3.1. Text Representation

Text data are usually semi-structured or unstructured text formats, and computers cannot process the original text formats. The first problem of text clustering is to transform the text into a structured form. Currently, transformation methods include the Boolean model, probability model, vector space model, and conceptual model [30]. The weight calculation functions for measuring the relationship between documents and feature words are the Boolean function, word frequency function, term frequency–inverse document frequency (TF-IDF) function, etc. TF-IDF function is used in text clustering because of its accuracy and simplicity. But the short text is limited by the length of documents. At the same time, the overlap rate of feature words between documents is low. When calculating the similarity of the sparse short text, there will be a drift phenomenon, which has great impact on the clustering effect. Therefore, the SSC method is used to solve the problem.

3.2. Multidimensional Feature Analysis of Microblog Users

Microblog user information is mainly divided into three categories: ID label information used to uniquely identify user identities; basic information that users can fill in when they register their accounts, which include descriptions of birthdays, age, gender, labels, etc.; users' data information of activities on the Microblog platform, such as number of followees, number of followers, Microblog number, forwarding amount, etc. Labels are the most direct summaries of personal personality traits, hobbies, and areas of expertise [1–3]. The number of followees, followers, and blog posts can indirectly reflect the impact and active degree of users. Based on the user's information search motivation, information sharing motivation, and other factors, a Microblog user's postings, followings, forwarding, comments, '@' and other active behavior characteristics are formed.

Microblog users relationship mining mainly includes two aspects: the community that discovers the user relationship structure; and the core user identification that measures user influence. This paper investigates the Microblog user community discovery method. At present, the clustering methods of Microblog users are mainly based on the user's basic attributes (age, gender, educational level, place of birth), and cannot reveal the user group characteristics based on text information. User information and text information are used as the two most important entities on the Microblog platform, which stores massive amounts of structured and unstructured data. Therefore, a new user interest model is constructed by incorporating text comment information in this paper.

This paper takes Microblog, released by the Xiaomi Company, as an example. The company collects feedback from users and obtains 1200 fans' effective evaluations of the use of Xiaomi phones. At the same time, the other basic information of these users can be used. The interest model of the Microblog user to the product is constructed from three dimensions such as user attributes, user behavior, and text features. Researchers have found that consumers of different genders have different needs and information processing methods, which lead to different brand awareness and consumption behaviors [16]. Labels are the keywords that most directly show user interest. Therefore, the two attributes of gender and label are important factors in showing user interest. The weak contact between ordinary users is mainly based on common follows. The 'follow' in this article refers to the user's concern for the relevant influencer account in a specific field. We selected 166 brand-related influencer accounts from the list of the top 100 followers of popular comments. There are three main categories: enterprise-related accounts, including Xiaomi derivative brand accounts and Xiaomi company founders; digital bloggers, including digital product evaluation accounts and new products and industry information sharing accounts; commodity accounts, mainly the official Microblog account of other mobile phones of the same type. The user comments under the official Microblog of the enterprise include the use of experience, emotional tendency, etc. And the relevance to the brand products is high. Product comments mainly include several categories: product brand awareness, product buying experience, product using experience, hot brand topics discussion, brand activity participation, etc. Users' comment information is unstructured data, which need to be converted into structured data for computer processing.

A vector space model is then used to represent the text set as a vector set consisting of the feature word dimensions and the corresponding weights in the corresponding space. Table 1 shows the classification of the key feature words of user comments.

Table 1. Key feature word types of user comments.

Product	Name	Mi, Mi 1, Mi 1s, Mi 3, Mi 5, Mi 5s plus, Mi 6, Mi 7, Mi max2, Mi mix, Mi mix2, Mi note, Mi note3, Mi tablet 2
	Function	Video shooting, Burning screen, Electric quantity, Exercise tolerance, Durable, Consume power, Configuration
	Appearance	Color, Black, Gold, Blue Headphone, Nude Machine
	Activity	Xiaomi Festival, Xiaomi home, Endorsement, Price reduction
User	Character	My mother, My grandmother, My girlfriend, My godmother, My aunt, My mother-in-law, My family, My parents, My classmates
	Feeling	Yearning, Attract, Like, Bad, Pleasantly surprised, Accident

3.3. Hierarchical Model of User-Product Interest

To establish the similarity relationship between ordinary users, the product interest hierarchy model is established as shown in Figure 1. The user's interest model can be expressed as an n-dimensional eigenvector, and the corresponding weights are used to indicate the aspects and degrees of the user's interest. In order to describe the user's interest more accurately, label, following account, and text feature word are used as the different dimensions of vector space. Microblog comments are limited by the number of words. The number of occurrences of the same feature word in a user's comment is 1-2 times, and the probability of it appearing in different users is small. The computation of TF-IDF in a vector space model is based on the assumption that the feature words are independent, without considering the similar words and related words in short text documents. The TF-IDF similarity measurement method relies on the number of overlapping words between texts, which results in a drift phenomenon in calculating short text similarity.

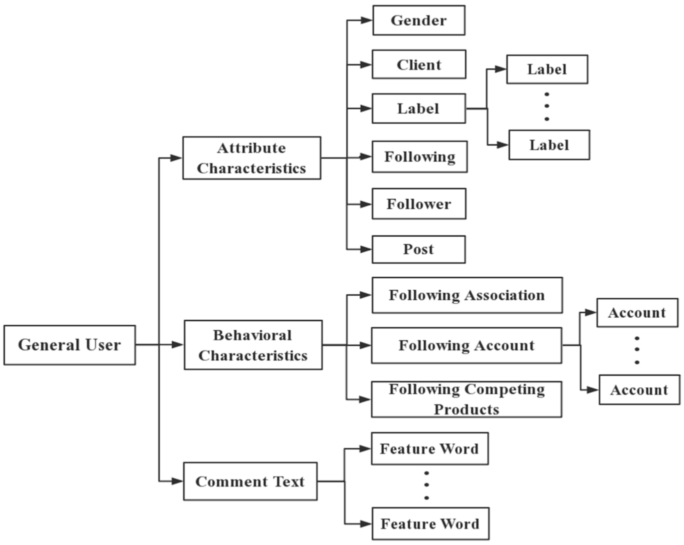


Figure 1. Hierarchical model of product interest for ordinary users.

In this work, the TF–IDF function is used to calculate the weight of document feature words. Then the similarity of feature words in different documents is calculated based on a synonym word forest, which adds feature phrases exceeding the threshold to vectors and calculates new feature weights. This paper uses *The Extended Edition of Synonyms in the Information Retrieval Laboratory of Harbin University of Technology* compiled by the Harbin University of Technology of China, which provides a five-layer coding tree structure with more detailed description as the level of the word meaning increases [29]. Table 2 is the word coding table. Among them, "=“ represents that seven-digit words with the same encoding are synonymous; "#" represents that seven-digit words with the same encoding are related words; "@" represents that there are no synonyms or related words in the dictionary.

Table 2. Word coding table.

Coding Bit	1	2	3	4	5	6	7	8
Examples of symbols	D	a	1	5	B	0	2	=\#\@
Symbolic properties	Big	Middle	Small		Word groups	Atomic word groups		
Level	First	Second	Third		Fourth	Fifth		

There are different meanings of a word in different contexts. Therefore, a word can map multiple meanings. The maximum similarity of meanings is used when calculating the word similarity in this work. Referring to Tian Jiule's method of calculating the similarity of words based on the synonym word forest [31], the similarity between the two meaning items is measured by the number position, which starts from the tree structure of the synonym word forest. The formulas for calculating the similarity of two meanings A and B are as follows. In the following, n represents the number of branches in the branch layer, and k represents the distance between two branches.

1. If A and B are not in the same tree, then the first coding bit is different such as:

$$sim(A, B) = f. \tag{1}$$

2. If A and B are on the same tree and are in the second layer, then the first coding bit is the same, and the second coding bit is different such as:

$$sim(A, B) = 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right) \times \left(\frac{n-k+1}{n}\right). \tag{2}$$

3. If A and B are on the same tree and are in the third layer, then the first two coding bits are the same, and the third, fourth coding bits are different such as:

$$sim(A, B) = 1 \times 1 \times b \times cos\ (n \times \frac{\pi}{180}) \times (\frac{n-k+1}{n}). \tag{3}$$

4. If A and B are on the same tree and are in the fourth layer, then the first four coding bits are the same, and the fifth coding bit is different such as:

$$sim(A, B) = 1 \times 1 \times 1 \times c \times cos\ (n \times \frac{\pi}{180}) \times (\frac{n-k+1}{n}). \tag{4}$$

5. If A and B are on the same tree and are in the fifth layer, then the first five coding bits are the same, and the sixth, seventh coding bits are different such as:

$$sim(A, B) = 1 \times 1 \times 1 \times 1 \times d \times cos\ (n \times \frac{\pi}{180}) \times (\frac{n-k+1}{n}). \tag{5}$$

6. If A and B are on the same tree, the first seven codes are the same. If the last bit is "=", then: $\mathrm{sim(A, B)} = 1$; if the last bit is "#", then:

$$sim(A, B) = e. \tag{6}$$

The parameters in the formula are adjusted in [0,1]. And the parameters a, b, c, d, f have been tested by many trials and manual tests. In our experiment, the similarity of words calculated is consistent with the result based on Zhihu.com when a = 0.65, b = 0.8, c = 0.9, d = 0.96, e = 0.5, and f = 0.1.

Based on the above similarity calculation method of feature words, a new feature word method is proposed to construct the text vector. In the initial vector space obtained by the text preprocessing, the feature words of the documents i and j composed of any two user comments are $d_i = (a_{i1}, a_{i2}, \ldots, a_{im}), d_j = \left(a_{j1}, a_{j2}, \ldots, a_{jn}\right)$, respectively. The feature word weights calculated by the vector space model and TF-IDF function are represented as vectors $T_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, $T_j = \left(w_{j1}, w_{j2}, \ldots, w_{jn}\right)$. When a feature word a_{is} in the document d_i does not belong to the document d_j, we find the corresponding meanings of the feature word $a_{jt} (j = 1, \ldots, n; j \neq i)$ and a_{is} in the document d_j. $\mathrm{sim}\left(a_{is}, a_{jt}\right)$ is calculated based on the similarity of multiple meanings of the synonym forest, and the maximum corresponding feature word a_{jt} is selected. When $\mathrm{sim}(a_{is}, a_{jt})$ is greater than a certain threshold, a_{is} is added as a new feature word $a_{j,n+1}$ to the feature word set d_j of the document *j*. According to the initial TF-IDF weights w_{is} of a_{is} in document *i*(*i* = 1,..., *n*; *i* ≠ *j*), the new weights are $w_{j,n+1} = w_{is} \times \mathrm{sim}(a_{is}, a_{jt})$ of $\mathrm{sim}(a_{is}, a_{jt})$ in document d_j.

4. The Proposed Community Detection Method

After calculating the similarity matrix, user grouping structure in the Microblog networks can be achieved by using the community detection method. In this work, the community detection method is based on the SSC approach. The traditional clustering method has a good effect in the small sample set, and the distance calculated by the finite feature dimension can accurately reflect the true difference between the sample points. The distribution of short text data in high-dimensional space is sparse, and the traditional clustering algorithm is not accurate enough. SSC is a kind of subspace clustering method based on spectral clustering. The principle is to solve the coefficient matrix of the sample in the high-dimensional space with the sparse representation of the data in the same subspace, and then transform it into the optimal graph partition problem for the data of the similarity matrix.

4.1. Sparse Representation Theory

SSC is based on the assumption that the data in high-dimensional space belong to some low-dimensional subspace essentially. Using the self-expression property of the data set, each data point can be sparsely represented as a linear combination by other data in the same subspace. The similarity matrix is then constructed by the sparse representation coefficient matrix [21]. Let $\mathrm{Y} \in \mathrm{R}^{\mathrm{D} \times \mathrm{N}}$ of D-dimensional data $\{y_i\}_{i=1}^{N}$ be located in the linear subspaces $\{S_l\}_{l=1}^{n}$ of R^{D} space, then the dimensions of corresponding subspaces are $\{d_l\}_{i=1}^{n}$. The data set Y is self-represented in matrix form as

$$\mathrm{Y} = [y_1, \dots, y_N] = [Y_1, \dots, Y_n]\Gamma. \tag{7}$$

In Equation (7), $\mathrm{Y}_l \in R^{D \times N_1}$ is a matrix composed of N_l data in the *l*th subspace S_l, whose rank is the dimension d_l of the subspace S_l. $\Gamma \in R^{N \times N}$ is the coefficient matrix to be solved. For each data point y_i, it can be represented by other data points besides itself, that is, $y_i = Yc_i$. The vector $c_i = [c_{i1}\ c_{i2}\ \dots c_{iN}]^T$ satisfies $c_{ii} = 0$. And when the data points y_i and y_j are not in the same subspace, $c_{ij} = 0$. The number of nonzero elements in vectors is usually represented by ℓ_0-norm, but ℓ_0-norm optimization is an NP-hard(Non-deterministic Polynomial) problem [32,33]. Compressed sensing theory shows that minimization of ℓ_1-norm of the absolute values of vector element is the optimal convex approximation of minimization of the ℓ_0-norm. The ℓ_1-norm sparse optimization model is:

$$min||C||_1 \tag{8}$$

$$s.t. Y = YC,\ diag(C) = 0.$$

In Equation (8), matrix C is $[c_1\ c_2\ \dots c_N] \in R^{N \times N}$, column vector $c_i = [c_{i1}\ c_{i2}\ \dots c_{iN}]^T$ corresponds to a sparse representation of data point y_i, and diag(C) $\in R^N$ is a vector composed of diagonal elements c_{ii} of matrix C. There may be noises and errors in the data set of practical problems. The SSC problem after adding the error term E can be described as:

$$\min_{C,E} \frac{\lambda_z}{2} ||Y - YC - E||_F^2 + ||C||_1 + \lambda_e ||E||_1 \tag{9}$$

$$s.t. diag(C) = 0.$$

In Equation (8), $\lambda_z\ and\ \lambda_e$ are positive penalty parameters that balance the reconstruction error $Y - YC$ and noise term E.

To solve the coefficient matrix C, two auxiliary variables U and Z are introduced as:

$$\underset{C,E,U,Z}{min}\frac{\lambda_z}{2}||Y - YC - E||_F^2 + ||Z||_1 + \lambda_e||U||_1 \tag{10}$$

$$s.t. C = Z, E = U, diag(C) = 0.$$

The corresponding augmented Lagrangian objective function is shown as:

$$\begin{gathered}\underset{C,E,U,Z}{min}\frac{\lambda_Z}{2}\left|\left|Y - YC - E\right|\right|_F^2 + \left|\left|Z\right|\right|_1 + \lambda_e\left|\left|U\right|\right|_1 +< A_C, C - \left(Z - diag(Z)\right) > + \\ \frac{\alpha_C}{2}\left|\left|C - \left(Z - diag(Z)\right)\right|\right|_F^2 +< A_E, E - U > + \frac{\alpha_E}{2}\left|\left|E - U\right|\right|_F^2.\end{gathered} \tag{11}$$

In Equation (11), λ_z , λ_e Z, E and U are the same as in Equation (9). α_C and α_E are positive penalty parameters, A_C and A_E are the multiplier matrixes, constant.

The alternating directions iterations of multipliers are described as the following updating steps to solve the SSC model Equation (11):

Step 1: Update *C*. When other parameters are given in the *k*th iteration, the objective function is minimized with respect to *C*. The linear equation of the C_{k+1} can be solved as:

$$(\lambda_n Y^T Y + a_c I)C_{k+1} = \lambda_n Y^T(Y - E) + a_c\left(Z_k + diag(Z_k)\right) - A_{c,k}. \tag{12}$$

Step 2: Update *E*. Similar to Equation (12), when other parameters are fixed, and the objective function is minimized with respect to *E*, it can be solved as:

$$E_{k+1} = (1 + \alpha_E)^{-1}\left(Y - YC_{k+1} + \alpha_E U_k - A_{E,k}\right). \tag{13}$$

Step 3: Update auxiliary variable. Update *Z* as:

$$Z_{k+1} = J - diag(J). \tag{14}$$

In Equation (14),

$$J \triangleq \frac{2}{\alpha_C}(C_{k+1} + \frac{2A_{C,k}}{\alpha_C}) \tag{15}$$

$$S_\eta(v) = (|v| - \eta)_+ + sgn(v) \tag{16}$$

$$(.)_+ = \begin{cases}(|v| - \eta), & |v| - \eta \geq 0 \\ 0, & otherwise\end{cases}. \tag{17}$$

Step 4: Update auxiliary variable *U* as:

$$U_{k+1} = S_{\frac{\lambda_e}{\alpha_E}}\left(E_{k+1} + \alpha_E^{-1}A_{E,k}\right). \tag{18}$$

Step 5: Update the multipliers matrix A_E and A_C as:

$$A_{C,k+1} = A_{C,k} + \alpha_C(C_{k+1} - Z_{k+1}) \tag{19}$$

$$A_{E,k+1} = A_{E,k} + \alpha_E(E_{k+1} - U_{k+1}). \tag{20}$$

4.2. Sparse Subspace Clustering

The SSC algorithm flow is as follows [27]: (1) The sparse coefficient matrix *C* is obtained by solving the sparse representation of the data set and normalized to construct the adjacency matrix $W = |C| + |C|^T$. The matrix $|C|$ is the absolute value of the corresponding element of the matrix *C*. (2) The degree matrix D of adjacency matrix W is then calculated, and normalized Laplacian matrix $D^{-1/2}LD^{-1/2} = D^{-1/2}(D - W) D^{-1/2}$ is computed according to the degree matrix D and the adjacency matrix W. (3) The *k* minimum eigenvalues of the Laplacian matrix are solved and the corresponding eigenvectors are normalized by row. The k column vectors obtained above constitute an n×k-dimensional matrix. Each k-dimensional row vector of the matrix corresponds to each original data point. (4) The n×k-dimensional matrix is clustered by spectral clustering algorithm, and the clustering results are obtained.

The spectral clustering algorithm does not need to know the original information of the data, but clusters directly in the adjacency matrix of the sample, which reduces the complexity of the algorithm. To ensure the sparseness of the data point relationship in the weight graph, only the points with high similarity are connected when constructing the undirected weighted graph. Sample points whose similarity is less than the threshold are not adjacent, and the sample points are only related to the nearest points. Through the above analysis, combined with the feature extraction results in the third part, the SSC-based Microblog network user grouping process is shown in Figure 2.

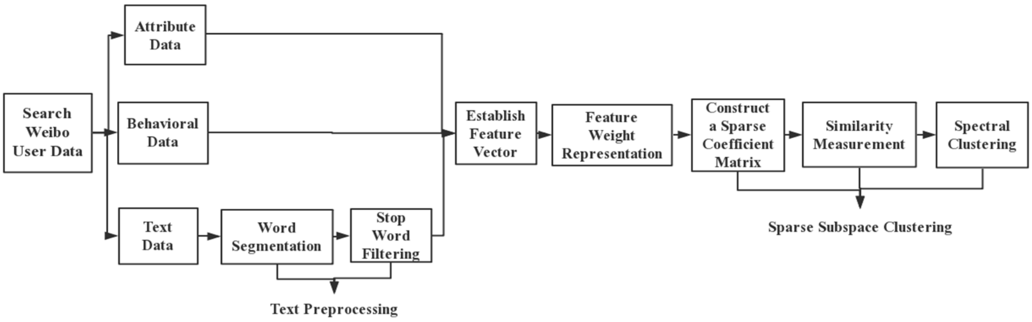


Figure 2. Flow chart of user grouping in a Microblog network based on SSC.

5. Experimental Results and Discussion

The experiment consisted of two parts. First, the standard text data sets marked from UCI were clustered to illustrate the validity of the SSC algorithm in high-dimensional sparse text vector space classification. Then, based on the SSC algorithm, cluster experiments were performed on the interests of actual Microblog users.

5.1. UCI Text Set Document Clustering

The TTC-3600 data set is a collection of 3600 news documents from six well-known news portals and agencies in Turkey from May to July 2015, which cover six different categories of economics, culture, arts, health, politics, sports, and technology [34]. First, the data was preprocessed, which included deleting irrelevant data such as JavaScript, HTML tags, punctuation symbols and operators, and removing stop words. Then, the word stem was extracted based on the Zemberek open-source NLP toolkit. The data consisted of 3600 samples, and a total of 5693 feature words were extracted. At last, the vector space model (VSM) and TF-IDF function to represent the document were combined as:

$$w_{ki} = \frac{tf_{ik}\times log(N/n_k)}{\sqrt{\sum_{k=1}^{n}(tf_{ik}\times log(N/n_k))^2}}. \tag{21}$$

In Equation (21), t_k is the *k*th feature word appearing in the document d_i, and tf_{ik} is the frequency in t_k, d_i. n_k is the number of documents containing t_k in the document set, and *N* is the total number of documents. The text sparseness of vector representation is high. And the high dimension will reduce the efficiency of the algorithm. Feature selection before the SSC algorithm is implemented. Usual feature selection methods include feature selection based on association rules and principal component analysis (PCA). In this work, PCA is used to transform the matrix dimension from D×n to r×n, D = 5693, n = 3600. r is a new feature dimension, which is 30, 40, 50, 60, 100 respectively. The clustering results are evaluated by clustering indicators ACC, ARI, Precision, Recall and NMI [35]. The evaluation results are shown in Table 3.

Experimental results showed that the SSC clustering effect was greatly affected by the data dimension, and the low-dimension representation led to the loss of effective information. With the increase of dimension, the noise and outlier data points would also increase correspondingly. The SSC clustering effect was the best when r was 50 in the data set. The algorithm performed with good accuracy for the data set with appropriate dimension reduction. Compared with the recent literature, the clustering effect of the proposed classification method was better than those of SVM and KNN algorithms [34]. As we know, dimensions of the samples usually have a negative effect in the pattern recognition problem of clustering. In this work, the principal component analysis (PCA) was used to reduce dimensionality. In fact, the suitable dimensionality representation was an open problem. In [34], the step of dimension reduction was not included in the SVM and KNN methods, so the proposed classification method was better than those of SVM and KNN algorithms.

Table 3. Evaluation of the SCS algorithm with TTC-3600 data set.

Index Name	r = 30	r = 40	r = 50	r = 60	r = 100
ACC (%)	73.97	75.28	80.86	66.75	66.70
ARI	0.4830	0.5063	0.6001	0.3763	0.4182
Precision	0.5481	0.5761	0.6599	0.4369	0.4829
Recall	0.6001	0.6056	0.6751	0.5672	0.5706
NMI	0.5278	0.5459	0.5871	0.4805	0.4920

Experimental results showed that the SSC clustering effect was greatly affected by the data dimension, and the low dimension representation led to the loss of effective information. With the increase of dimension, the noise and outlier data points would also increase correspondingly. The SSC clustering effect was the best when r was 50 in the data set. The algorithm performed with good accuracy for the data set with appropriate dimension reduction. Compared with the recent literature, the clustering effect of the proposed classification method was better than those of SVM and KNN algorithms [34].

5.2. Experiments on Microblog Network

Experimental data collection was obtained by the Octopus collector [3]. First of all, considering the timeliness of comments and the existence of a large amount of worthless information, we selected six popular Microblog posts published by Xiaomi Company's Microblog homepage, and collected 200 popular comments under each Microblog. A total of 1200 popular followers' popular comments and corresponding user IDs were adopted. Then we randomly selected 100 follower users to get their following lists, and filtered out the 166 influencer accounts (related companies, digital bloggers, competing related products) related to this experiment. Finally, the user's ID link was entered into the Microblog homepage according to the comments obtained. The basic information of the users was collected, including gender, labels, number of follows, followers, and posts. And then we extracted the client terminals which had newly published five posts. The common follow list of the user's homepage refers to the current login account and the account that the user follows collectively. By obtaining a list of common follows of new accounts and ordinary users, we could indirectly get the following situation of ordinary users to 166 influencer accounts in specific areas.

The Jieba word segmentation in Python supports three-word segmentation modes: precise mode, full mode, and search engine mode [4]. This work uses the precise mode of Jieba word segmentation for the textual information of user comments. The text contains a large number of function words such as modal particle, adverb, preposition, connective, etc. The stop word filtering operation is performed after the word segmentation. By integrating the "Hagong University Stop Words" [5] and "Sichuan University Machine Learning Intelligence Lab Stop Words" [6] and other stop words, we got a stop word list with 2079 stop words. Each comment was saved after the word segmenting into a separate document by word segmentation and filtering stop words operation. Then the data representation were transformed into structured data in a vector space model. Each user was represented as a vector in space. The elements of the vector were composed of user attribute features, behavior features, and text features. Different weight calculation methods were set for each feature item. The value of the user's numeric attribute represented the weight of the corresponding feature item, while the weight of the non-numeric attribute was calculated by using the Boolean function. The inconsistency of the dimensions of different features would affect the clustering results. Therefore, the weights of the attribute features (gender, client) were normalized by min-max according to the column. The value range of the feature elements was finally between [0,1]. The CountVectorizer and TfidfVectorizer in the sklearn package of Python were used to calculate the TF-IDF weights for the text comment feature words after the word segmentation. The similarity between two words of all feature words in vector space was calculated, and the feature word-feature word similarity matrix was obtained. The feature words that exceeded the similarity threshold were added to the corresponding documents. The new feature weights in the documents were calculated according to the similarity value and the maximum initial weight of the word in other documents. Table 4 is an example of the calculated similarity of some feature words.

Table 4. Similarity of partial feature words.

Sense Coding	Nephew Ah16A01 =	Families Ah01B01 =	Mother Ah04B01 =	Brother Ah09B01 =	Father-in-Law Ah07A03 =	Mother-in-Law Ah02C01 =
Nephew	1.00	0.14	0.27	0.50	0.41	0.18
Families	0.14	1.00	0.68	0.45	0.54	0.77
Mother	0.27	0.68	1.00	0.59	0.68	0.72
Brother	0.50	0.45	0.59	1.00	0.72	0.50
Father-in-law	0.41	0.54	0.68	0.72	1.00	0.59
Mother-in-law	0.18	0.77	0.72	0.50	0.59	1.00

Through the above steps, the space vector model of user interest was obtained. The SSC was used then to group Microblog users. The main process of SSC clustering was as follows: (1) The similarity matrix was calculated, and the coordinates in each space were normalized. The coordinate range was between [−1,1]. (2) Sparse decomposition was obtained by using the MATLAB convex optimization toolkit to obtain the sparse coefficient matrix. The absolute value was obtained for each element of the sparse coefficient matrix, and then was normalized to construct an adjacency matrix. (3) The Laplacian matrix was obtained according to the adjacency matrix, and the singular values arranged in descending order were obtained by singular value decomposition. The corresponding right singular vector took the smallest n column to form a matrix, and the row vectors of the matrix were regularized. (4) The regularized matrix was clustered to get the user's group label. The user grouping situation and the corresponding key feature words are displayed in a visibility graph.

Experimental results show that the number of clusters had a great influence on the clustering algorithm. The number of clusters by the number index was determined. The number was an evaluation index of intra-cluster density and inter-cluster dispersion. Its value was within the range of [−1,1]. With the larger value of the number index, the higher the similarity of intra-cluster samples was, and the lower the similarity of inter-cluster samples was [33]. The average number of all data points was the total number of clustering results. The number of clusters was generally 2-10. With larger average contour coefficient, the clustering effect was better. The average value of the clustering results of different cluster numbers was calculated. When the number of clusters was 2, the average value was the largest, but the sum of squared errors was not the smallest. When the number of clusters was 3, the average contour coefficient was high while satisfying the error square sum as small as possible. The relationship between the contour coefficient and the number of clusters is shown in Figure 3. In the SSC, the smallest three vectors in the column represent the data points, and the clustering result is shown in Figure 4.

For the related influencer accounts, users in the same cluster recommended each other's following list. At the same time, word clouds were generated from the text information of users in each cluster by word cloud tools. The degree of attention and participation of different user groups on products were obtained. As shown in Figure 5, the three groups of users had a high degree of attention for Xiaomi products and were actively involved in the polling activities initiated by Xiaomi companies. The first group of users paid much attention to the product performance and Xiaomi robot. The second group of users repeatedly mentioned the appearance and the motherboard of the mobile phone, the scratches appearing in the feedback. The third group of users was concerned about the price and quality problems of the products. They were opposed to the "hunger marketing" of the company. Therefore, different marketing strategies can be formulated and implemented according to the user groups with different concerns. Personalized push can be realized. And collaborative filtering can be implemented to recommend the following accounts for grouped users.

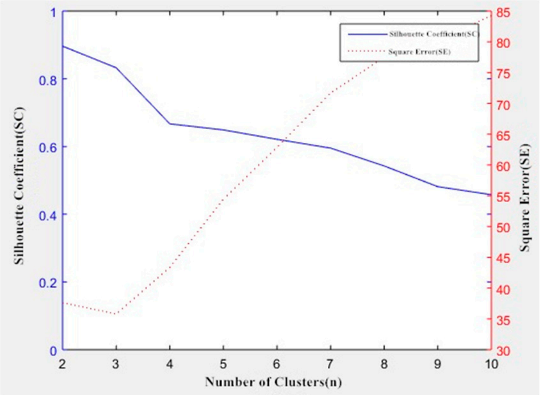


Figure 3. Relation graph of SSC number and cluster number.

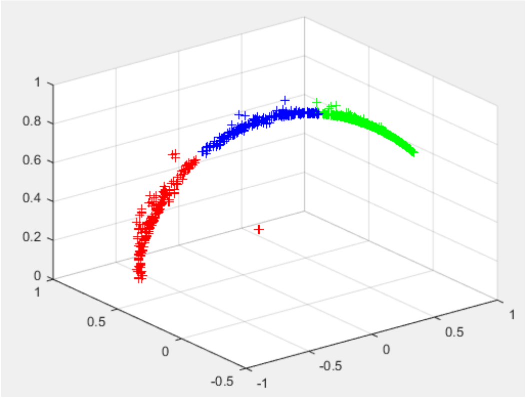
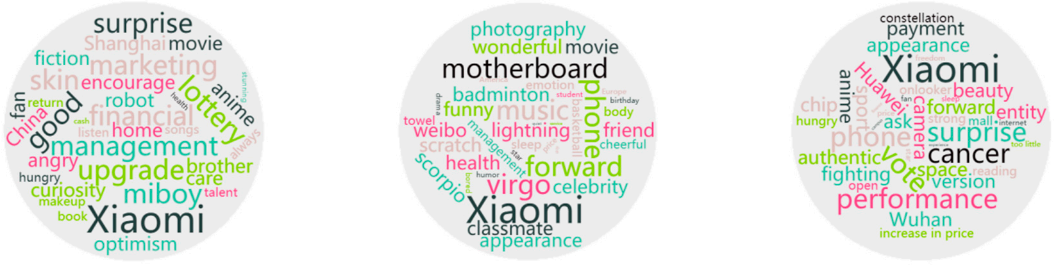


Figure 4. Clustering results of the Microblog network SSC.



(**a**) The group of users paid much attention to the product performance

(**b**) The group of users mentioned the appearance of the mobile phone

(**c**) The group of users was concerned about the price and quality problems

Figure 5. Visualization results of three groups of user comment cloud graphs.

The user clustering of the SSC algorithm and K-means clustering algorithm were compared on the Microblog network data set. The clustering result was evaluated by the modularity index. The modularity index is used to evaluate the results of unlabeled network partitioning [22], which is defined as:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\sigma\left(C_i, C_j\right). \quad (22)$$

In Equation (22), $A_{i,j}$ is an element of the network adjacency matrix which is the edge weights of nodes *i* and *j*; $k_i = \sum_j A_{i,j}$ represent the degree of node *i*; C_i represents the community to which node i belongs; m represents the total number of sides of the community network. When i and j are in a community, $C_i = C_j$, $\sigma(C_i, C_j) = 1$, otherwise $\sigma(C_i, C_j) = 0$. The range of modularity Q is in [−0.5,1]. With a larger value, the better the effect of community division is. Table 4 is the comparing result. From Table 5, it can be seen that the SSC algorithm results are more closely connected within the user group in the case of different cluster numbers compared with the K-means algorithm.

Table 5. Comparison of SSC and K-means clustering results of the Microblog network.

Evaluation Value	Algorithm	*n* = 2	*n* = 3	*n* = 4	*n* = 5	*n* = 6	*n* = 7	*n* = 8	*n* = 9	*n* = 10
Q	SSC	0.18	0.31	0.41	0.41	0.42	0.42	0.42	0.44	0.43
	K-means	−0.05	−0.05	−0.04	−0.04	−0.03	−0.03	−0.02	−0.06	−0.02

Experimental results showed that SSC can effectively cluster short texts. Similar to traditional clustering methods, the reduction of feature dimension will result in the loss of data information and affect the effect of the SSC algorithm. However, as the dimension increases, the introduced noise and abnormal points will increase, and the accuracy of clustering will decrease. We took the post of user product feedback issued by Xiaomi Company as an example. The followers effectively evaluated the use of the Xiaomi mobile phone. This work constructed a value model of Microblog users from three dimensions: attribute characteristics, behavior characteristics, and text features. The user grouping of the SSC algorithm was then implemented. And the word cloud tool was used to extract the keywords from the grouped user comments to obtain the interesting points and concerns of different groups.

6. Conclusions

The text information on the mobile social platform was a short text whose word number is controlled within a certain range. The sparsity of high-dimensional features in short text clustering makes the measurement of sample distance unable to truly reveal the difference between data points. Traditional text mining technology is no longer suitable for sparse short text mining. In this paper, SSC was adopted to massive short text data sets. Text mining was implemented, and we proposed the Microblog user clustering method. The feasibility and effectiveness of the SSC algorithm were verified on UCI's Turkish news text data sets. For the Microblog network, we constructed the user interest model. Product-related comments were collected. And user information of Xiaomi Enterprise's official Microblog was used to construct the Microblog network user grouping method.

In order to deal with the similarity drift phenomenon caused by Chinese short text sparsity, we improved the traditional TF-IDF feature weight calculation method. The similarity of feature words calculated by the synonym word forest was combined with some extended synonyms in the elements of text vector. The corresponding feature word weights made the data point similarity in the sparse vector space consistent with real situation. It was difficult to obtain the prior information of the user group. The number of clusters with the best clustering effect was then introduced. Experimental results show that the proposed SSC method is superior to the K-means clustering on the high-dimensional sparse features of the Microblog network data set.

Author Contributions: All authors have made significant contributions to the paper. The corresponding author B.T. proposed the ideal, methodology, and formal analysis. Z.Z. used software to perform the experiments. The paper is written by Z.Z. and reviewed by B.T.

Funding: This research is supported by the National Natural Science Foundation of China (71302080), Ministry of Education Research of Social Sciences Youth Foundation Projects (13YJC630149), Special Research Foundation for the Doctoral Program of Higher Education of China (20120078120001).

Acknowledgments: Thanks for the hard work of editor and reviewers' comments to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lianxi, W.; Shengyi, J.; Guansong, P.; Meiling, W. A literature review of user relationship mining on Microblog. *J. Inf.* **2012**, *31*, 91–97.
2. Li, H. Microblog user feature analysis and core user mining. *Inf. Stud. Theory Appl.* **2011**, *34*, 121–125.
3. Bing, L.E.I.; Wei, L.I.U. Research on user Interest model of Microblog following recommendation service. *Inf. Sci.* **2015**, *33*, 126–130.
4. Guangmin, L.; Xinshan X.; Lei Z. Research on product opinion mining in Microblog. *J. Inf.* **2014**, *33*, 135–138.
5. Peng, M.; Huang, J.; Zhu, J.; Huang, J.; Liu, J. Mass of short texts clustering and topic extraction based on frEquationuent item sets. *J. Comput. Res. Dev.* **2015**, *52*, 1941–1953.
6. Xiong, S.-F.; Ji, D.-H. A short text sentiment-topic model for product review analysis. *Acta Autom. Sin.* **2016**, *42*, 1227–1237.
7. Qiang, B.; Jian L.; Yulai B. A new text clustering method based on semantic similarity. *New Technol. Libr. Inf. Serv.* **2016, 12**, 9–16.
8. Vidal, R.; Favaro, P. Low rank subspace clustering (LRSC). *Pattern Recognit. Lett.* **2014**, *43*, 47–61.
9. Ding, Z.; Zhang, X.; Sun, D.; Luo, B. Low-rank subspace learning based network community detection. *Knowl. -Based Syst.* **2018**, *155*, 71–82.
10. Sahami, M.; Heilman, T.D. A web-based kernel function for measuring the similarity of short text snippets. In Proceedings of the 15th international conference on World Wide Web. AcM, Edinburgh, UK, 23–36 May 2006; doi:159593 323 9/06/0005.
11. Yih, W.; Meek, C. Improving similarity measures for short segments of sext proc. *AAAI* **2007**, *7*, 1489–1494.
12. Tiancai, L.I.; Yaoyi X.; Bo W.; Jiamin Z. Improved short text hierarchical clustering algorithm. *J. Inf. Eng.* **2015**, *16*, 743–748+752.
13. Yang; B.; Qingwen H.; Min L.; Yapeng Z. Short text classification algorithm based on improved TF-IDF weight. *J. Chongqing Univ. Technol.* **2016**, *30*, 108–113.
14. Chao, P.; Shibing X.; Min J. Research on Microblog user clustering based on improved LDA theme model. *Inf. Stud. Theory Appl.* **2016**, *39*, 135–139.
15. Xiaohong, L.; Men X.; Huifang M.; Yannian H. A short text clustering algorithm based on spectral cut. *Comput. Eng.* **2016**, *42*, 178–182.
16. Wan, X.; Li, L. Community division method with structure and attribute. *Comput. Technol. Dev.* **2017**, *27*, 97–101.
17. Xiaoyan C.; Guanzhong D.; Libing Y. Community-finding algorithm in complex networks based on spectral clustering. *Comput. Sci.* **2009**, *36*, 49–50.
18. Javed, M.A.; Younis, M.S.; Latif, S.; Qadir, J.; Baig, A. Community detection in networks: A multidisciplinary review. *J. Netw. Comput. Appl.* **2018**, *108*, 87–111.
19. Xu, Z.-M.; Li, D.; Liu, T.; Li, S.; Wang, G.; Yuan, S.L. Measuring Similarity between Microblog Users and Its Application. *J. Comput.* **2014**, *37*, 207–218.
20. Yang, K.; Zhang, N. Structure and Cluster Analysis on Microblog User's Relationship Networks. *Complex Syst. Complex. Sci.* **2013**, *10*, 37–43.
21. Yongcheng W.; Yanjie C. User Association Mining Based on Spectral Clustering. *Telecommun. Eng.* **2016**, *56*, 32–37.
22. Sun, Y.F.; Li, S. Similarity-Based Community Detection in Social Network of Microblog. *J. Comput. Res. Dev.* **2014**, *51*, 2797–2807.
23. Agrawal, R.; Gehrke, J.E.; Gunopulos, D.; Raghavan, P. Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Record* **1998**, *27*, 94–105.
24. Elhamifar, E.; Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications.. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2765–2781.
25. Gu, N.N.; Fan, M.Y.; Wang, D.; Jia, L.H.; Du, L. Semi-supervised classification based on affine subspace sparse representation. *Sci. China* **2015**, *45*, 985–1000.
26. Parsons, L.; Haque, E.; Liu, H. Subspace clustering for high dimensional data: A review. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 90–105.
27. Mahmood, A.; Small, M. Subspace based network community detection using sparse linear coding. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 801–812.
28. Zhu, X.; Suk, H.I.; Lee, S.W.; Shen, D. Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 607–618.

29. Tian, B.; Li, W. Community detection method based on mixed-norm sparse subspace clustering. *Neurocomputing* **2018**, *275*, 2150–2161.
30. Wenwen, M.; Yigui, D. New feature weight calculation method for short text. *J. Comput. Appl.* **2013**, *33*, 2280–2282.
31. Jiu-le, T.; Wei, Z. Words Similarity Algorithm Based on Tong yi ci Cilin in Semantic Web Adaptive Learning System. *J. Jilin Univ. (Inf. Sci. Ed.*) **2010**, *28*, 602–608.
32. Shi, P.; He, K.; Bindel, D.; Hopcroft, J.E. Locally-biased spectral approximation for community detection. *Knowl.-Based Syst.* **2019**, *164*, 459–472.
33. Ma, X.; Di, D.; Wang, Q. Community detection in multi-Layer networks using joint nonnegative matrix factorization. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 273–286.
34. Kilinç, D.; Özçift, A.; Bozyiğit, F. TTC-3600: A new benchmark data set for Turkish text categorization. *J. Inf. Sci.* **2015**, *43*, 174–185.
35. Nan, D.Y.; Yu, W.; Liu, X.; Zhang, Y.P.; Dai, W.D. A framework of community detection based on individual labels in attribute networks. *Phys. A* **2018**, *512*, 523–536.