Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 5054825, 10 pages
https://doi.org/10.1155/2017/5054825

Research Article

Nonuniform Granularity-Based Classification in Social Interest Detection

Wenjuan Shao,[1,2] Qingguo Shen,[2] Xianli Jin,[3] Liaoruo Huang,[2] and Jingjing Chen[3]

[1]*Zijin College, Nanjing University of Science and Technology, Nanjing, Jiangsu 210023, China*
[2]*College of Communication Engineering, PLA University of Science and Technology, Nanjing, Jiangsu 210007, China*
[3]*Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China*

Correspondence should be addressed to Qingguo Shen; shenqg2@163.com

Received 21 November 2016; Revised 16 April 2017; Accepted 9 May 2017; Published 6 July 2017

Academic Editor: Ibrahim Zeid

Copyright © 2017 Wenjuan Shao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Social interest detection is a new computing paradigm which processes a great variety of large scale resources. Effective classification of these resources is necessary for the social interest detection. In this paper, we describe some concepts and principles about classification and present a novel classification algorithm based on nonuniform granularity. Clustering algorithm is used to generate a clustering pedigree chart. By using suitable classification cutting values to cut the chart, we can get different branches which are used as categories. The size of cutting value is vital to the performance and can be dynamically adapted in the proposed algorithm. Experiments results carried on the blog posts illustrate the effectiveness of the proposed algorithm. Furthermore, the results for comparing with Naive Bayes, *k*-nearest neighbor, and so forth validate the better classification performance of the proposed algorithm for large scale resources.

1. Introduction

Automatic classification of user interest from social media has gained much attention in recent years [1]. Users' social interest dataset consists of massive amount and various types of resource, such as video, audio, image, text, and blog posts. These social resources may reveal users' potential preferences, broaden the vision for user market mining, and improve the performance of online recommendation system. Moreover, the user information of these resources can also provide a deeper understanding of our social network architecture.

Nowadays a number of techniques have been used to mine and analyze these social resources, but the most widely accepted method is the classification for the resources. It is worth noting that existing classification approaches are mainly drawn from machine learning techniques such as Support Vector Machine (SVM), Naïve Bayes (NB), and *k*-nearest neighbor (KNN) classifiers [2]. Although these classification technologies are quite mature, research on granularity-based classification is still in its early stages. Hence, it is an open research question.

Granularity refers to "average measurement for particle size" in physics; however, in this paper it refers to "average measurement for information thickness" [3]. The idea of information granulation has been applied successfully in such fields as rough set theory, divide and conquer, machine learning, and cluster analysis databases, since it has been proposed by Zadeh in 1979 [4]. Granular computing is an essential part of information granularity, the basic ingredients of which are subsets, lasses, and clusters of a universe [4, 5]. There are some fundamental issues in granular computing, such as granulation of the universe, description of granules, expression of relationships between granules, and computation with granules. So granular computing can be studied mainly from two related aspects: the construction of granules and the computation with granules. The former deals with the formation, representation, and interpretation of granules, and the latter deals with the utilization of granules in problem solving, and we will concentrate on the former in this paper.

Classification refers to "the task of assigning sample data to one or more predefined categories." Classification is mainly based on a training sample set, and experts in the field indicate which samples belong to a class and which samples belong to another class. Due to the subjectivity for prior knowledge, it often uncoordinates with characteristic of space

and similarity measure function. To avoid the incompatibility for classification, a novel classification approach based on nonuniform granularity is proposed to classify resources in social network.

This paper is organized as follows: Section 2 describes some basic concepts of information granules and rough set. In Section 3 we introduce granularity principle in classification and clustering, framework, and specific process of the proposed classification algorithm. In Section 4 we propose a novel algorithm for classification. Experiments and results are provided in Section 5. Finally, Section 6 ends with the summary of conclusion and future work.

2. Related Work

Recently clustering and classification have drawn many researchers' attention. Although research on classification is quite mature, resources classification based on granularity in social network is still in early stages; hence, it is an open problem for research.

In recent years, much work has been done for classification especially for text classification [6–9]. Using supervised learning technique Support Vector Machine (SVM), Joachims [7] proposed an approach to the text classification. Ng et al. [9] proposed an automated learning approach to text categorization based on perception learning and new feature selection metric called correlation coefficient and finally conducted a usability case study by comparing the performance of such an automated learning approach with traditional approach of text categorization. McCallum and Nigam [8] clarified the confusion between multivariate Bernoulli model and multinomial model in document classification and called both models "Naïve Bayes." By comparing classification performance on some corpora, including Web pages, UseNet articles, and Reuters newswire articles, it was showed that multivariate Bernoulli performs well with small vocabulary size, whereas the multinomial usually performs much better with large vocabulary size. Friedman et al. [6] introduced Bayesian Network Classifiers and proposed a new method called Tree Augmented Naive Bayes (TAN), which outperforms Naive Bayes.

The works mentioned above use machine learning techniques to classify resources. Although information granularity approach is rarely utilized on resources classification, it has been applied into many other fields, such as rough set theory, divide and conquer, machine learning, cluster analysis, and databases. We extend our previous work [10] by adopting a fast-start strategy proposed in [11] as the benchmark for adjusting the cutting granule dynamically. Furthermore, the experiment results prove that the proposed algorithm using granularity principle can achieve better classification performance for large scale resources.

3. Basic Concepts

3.1. Information Granules

3.1.1. Definition

Definition 1 (granular system). A Granular system can be defined in the following three-tuple forms [9]:

$$G = (U, D, F). \tag{1}$$

G is a set of object granules.

U is a finite nonempty set, which defines the object granules to be discussed.

D is a finite nonempty set, which is the description set of all object granules in U.

F defines the relationship between all the object granules in U.

3.1.2. Relationship between Different Granularity Worlds

Definition 2. If X and Y are two sets, then $X \times Y$ is the product of sets X and Y, and if $R \in X \times Y$, we call R one of the relations of $X \times Y$, for $x \in X$, $y \in Y$, and we have xRy.

Definition 3. if $xR1\,y \Rightarrow xR2\,y$, we call $R1$ more detailed than $R2$, marked as $R1 \leq R2$.

3.2. Rough Set Theory. Rough sets theory was proposed by Pawlak in 1982 [12]. It is a mathematic tool to deal with inaccurate, uncertain, or fuzzy knowledge in many branches of artificial intelligence. For the convenience of illustration, we only introduce some basic concepts of rough set theory [13] here.

Definition 4 (lower and upper approximation). We assume that A consists of object sets, and a pair of approximation operators, $\underline{\text{apr}}$, $\overline{\text{apr}}$, are formally defined as follows:

$$\begin{aligned} \underline{\text{apr}}\,(A) &= \cup \left\{ X \mid X \in \frac{U}{E},\ X \subseteq A \right\}, \\ \overline{\text{apr}}\,(A) &= \cup \left\{ X \mid X \in \frac{U}{E}, A \subseteq X \right\}. \end{aligned} \tag{2}$$

The lower approximation $\underline{\text{apr}}(A) \in U/E$ is the greatest definable set contained in A, and the upper approximation $\overline{\text{apr}}(A) \in U/E$ is the least definable set containing A.

Definition 5 (positive, negative, and boundary regions). $\text{POS}(A) = \underline{\text{apr}}(A)$, and it is the positive region of A. $\text{NEG}(A) = A - \overline{\text{apr}}(A)$, and it denotes the negative region of A. And $\text{BN}(A) = \overline{\text{apr}}(A) - \underline{\text{apr}}(A)$, and it refers to the boundary region of A. The positive region contains the objects that can be definitely described by object set A. The negative region is the set of objects that cannot be defined by object set A. The boundary region consists of objects that may be defined by object set A.

4. Principle and Algorithm for Classification

In this section, we will introduce granularity principles of clustering and classification in granule spaces. Incompatibility between them and corresponding treatment method will also be presented. Ultimately, we describe the framework and specific process of the proposed classification algorithm.

4.1. Granularity Principles for Clustering and Classification

4.1.1. Granularity Principle in Clustering. Clustering is a multivariate and statistical method which can classify sample

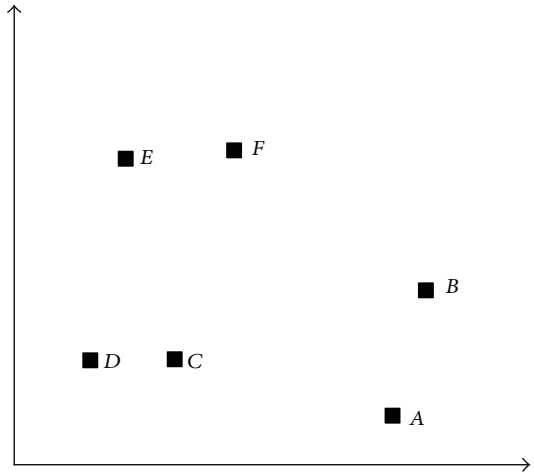


Figure 1: Sample points.

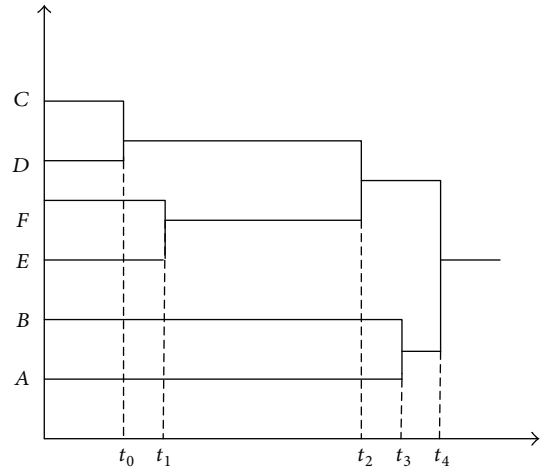


Figure 2: Clustering pedigree chart for sample points.

points. Clustering results usually can be illustrated by clustering pedigree chart drawn according to similarity measure function. In this paper, we use the shortest distance method as similarity measure function to generate clustering pedigree chart. The specific process is summarized as follows, and we take the sample points in Figure 1 as an example:

(1) Regarding every sample point as one class

(2) Calculating the distance between the sample points

(3) Merging the closest pair to be a new class, regarding the distance of the merged classes as the height of the new class, and recalculating the distance between the new class and other classes

(4) According to the name and the distance of the merged sample points, marking them on corresponding location in clustering pedigree chart

(5) Returning to (3), until all sample points can merge into one class.

According to the process above, clustering results of sample points in Figure 1 are shown as Figure 2.

We assume classification threshold is named τ, and the given object set $X = \{A, B, C, D, E, F\}$. The clustering pedigree chart for sample points in Figure 1 is showed in Figure 2. We can see that when the value of τ decreases, different clustering results can be obtained.

(1) If $\tau \geq t_4$, all sample points can be clustered into one class.

(2) If $t_3 < \tau < t_4$, then object set X are clustered into two classes: A and B belong to one class, and the remaining sample points are clustered into another class, $\{C, D, E, F\}$.

(3) If $t_2 < \tau < t_3$, then object set X are clustered into three classes, point A and point B become separate class, $\{A\}, \{B\}$, and the remaining sample points are classified as another class $\{C, D, E, F\}$.

(4) If $t_1 < \tau < t_2$, then object set X are clustered into four classes $\{A\}, \{B\}, \{E, F\}$, and $\{C, D\}$.

(5) If $t_0 < \tau < t_1$, then object set X are clustered into five classes, $\{A\}, \{B\}, \{E\}, \{F\}$, and $\{C, D\}$.

(6) If $\tau < t_0$, every sample point can form a class, respectively, $\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$, and $\{F\}$.

We can find that when τ varies, the corresponding classification results will differ greatly. Specifically, with a lager τ, these sample points will show us a "rough" outline, for some details are ignored, and slightly analogous points will also be clustered into the same class, while, with a smaller τ, some minor differences between sample points are portrayed clearly, and only extremely analogous sample points can be clustered into the same class.

4.1.2. Granularity Principle in Classification. Classification is a data mining or machine learning technique used to predict category for data resources. The goal of the classification algorithm is to explore for the intrinsic quality of sample instances in each class.

Clustering tries to reflect the goal for "holding-together" nature between sample points as faithfully as possible. However, classification is actually a learning process, by which with the given sample points experts in the field classify the points into several category in terms of their prior knowledge. The ideal prior knowledge is the following situation: in feature space, heterogeneous sample points with clear distinction or a small similarity measure are classified into different categories; and sample points with a large similarity measure are gathered into one category.

However, in real life, clustering results are often incompatible with prior knowledge. Experts in the field think some points should be classified into one class, whereas in fact these points often have particularly far distance in feature space, or their similarity measure is very small. On the other hand, those points classified into different categories often have close distance, or their similarity measure is very large. Examples are shown as Figures 3 and 4.

Figure 3 shows that sample points of analogous curve shape are clustering into one class according to prior knowledge of experts. However, based on similarity measure

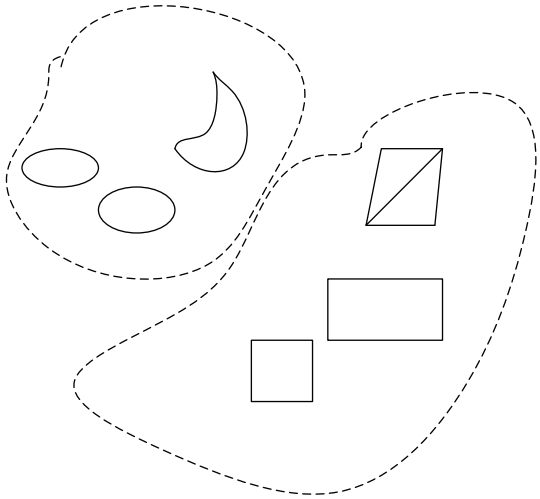


Figure 3: Clustering based on prior knowledge.

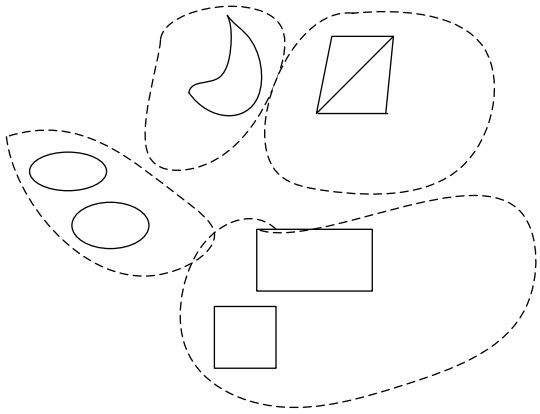


Figure 4: Clustering based on similarity measure.

function, the clustered results are totally different, which are showed in Figure 4, indicating uncoordinated relationship between prior knowledge and similarity measure function.

There are two reasons contributing to this incompatibility. On the one hand, clustering is a process, attempting to objectively reflect the holding-together character of the sample points; once the feature space and similarity measure function are decided, clustering results are determined accordingly. On the other hand, prior knowledge of classification is subjectively extracted by experts; for the same samples, different experts may have different classification results. The inevitable difference between subjective and objective results is called incompatibility. The treatment method for the incompatibility is introduced in the following chapter.

4.1.3. Treatment for the Incompatibility between Prior Knowledge and Similarity Measure. From the content above, classification threshold τ can have equal granularity in granular system to some extent. Considering the characteristics of rough set, we can use it to deal with the incompatibility between prior knowledge and similarity measure function.

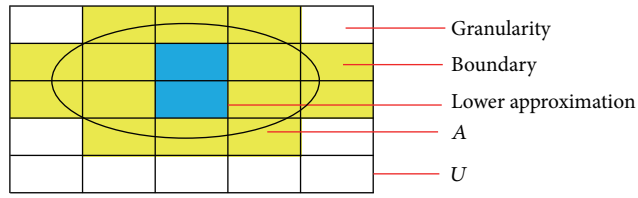


Figure 5: Partition of A based on uniform granularity.

If the existing knowledge of rough set theory can precisely convey object set A which is defined by prior knowledge, it means clustering results and prior knowledge are coordinated; that is, boundary is equal to zero. Otherwise, the incompatibility between them occurs.

The extent of this incompatibility can be quantitatively expressed by the boundary size of A, expressed as boundary (A, R).

Figure 5 shows the classification results for object set A under uniform granularity.

We find that the usage of uniform partition granularity will lead to a big size of *boundary*, shown as the yellow parts in Figure 5, indicating the incompatibility between clustering results and prior knowledge.

There are mainly three strategies to eliminate this incompatibility. One way is to transform the feature space, such as Support Vector Machine (SVM) theory proposed by Yao and Yaohua [14]. Another way is spherical projection algorithm proposed by Professor Zhang [15]. The third is a classification strategy based on granularity. This paper focuses on the third and does not concern the other two strategies.

The size of granularity reflects the number of equivalence classes. With coarse granularity, we can only get coarse equivalence classes, and the number of equivalence classes of A will be smaller compared to finer granularity. But, in extreme case, when granularity size is particularly fine, each object will represent an equivalence class; in this situation, roughness of A is equal to zero. However, the smallest granularity size is not what we really want, and for this kind granularity leads to invalid classification for A. In fact, it is only a simple enumeration for A, since we cannot get any useful information about A.

So our goal is to find a new kind of knowledge. On the one hand, it can precisely express object set A, and on the other hand it can convey the regularity of the elements in A. Based on the thought above, we give up the idea of uniform granularity and use the thought of nonuniform granularity.

For the given instance set A and the equivalent relation k, the process and principle for the partition of nonuniform granularity are summarized below:

(1) For A, we first select a relatively coarse granularity $k1$ (referring to classification threshold when applying clustering algorithm) and calculate the upper approximation, lower approximation, and boundary about A. Because the coarse granularity can make our process simple, we are more inclined to use it as our first selection.

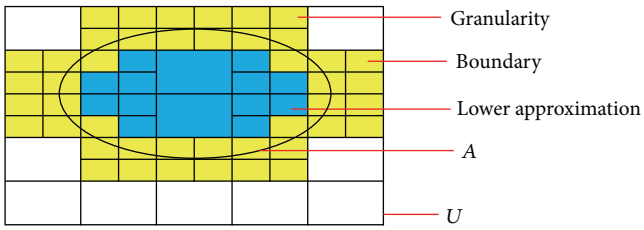


Figure 6: Partition of A based on nonuniform granularity.

(2) We remove the lower approximation accurately expressed by $k1$, for the boundary, apply a much finer granularity than the former, and repeatedly calculate upper approximation, lower approximation, and boundary for the uncertain boundary, until the boundary equals zero. The zero boundary means the object set A can be precisely expressed.

The purpose for the idea of nonuniform granularity is to divide object set A into a number of subclasses, expressed as $A = A1 \cup A2 \cup \cdots \cup An$. Each of the subclasses is the maximum subset with an accurate expression under its corresponding granularity, and the number of connection symbols "$\cup$" in the above formulation can quantitatively determine the degree of incompatibility between similarity measure function (or feature space) and prior knowledge. Specifically, if class A has many connection symbols "$\cup$," it shows the weaker compatibility and indicates that similarity measure function (or feature space) does not support prior knowledge; otherwise it indicates a good coordination between them.

The partition of object set A with nonuniform granularity is shown as Figure 6.

4.1.4. The Impact of Granularity on Classification. Appropriate granularity is important to the classification for resources. When granularity is gradually increasing, the corresponding category varieties will decrease gradually. And, with the decreasing of granularity, classification will become fine, and the category varieties will increase. But when granularity is smaller than a certain value, it will lead to insignificance of the classification result, sample instances which have large similarity measure will be wrongly classified into different categories. Similarly, if granularity is too large, it will result in coarse classification, affecting the classification results, and some details are probably ignored, and various categories in sample instances will be classified into one class.

Examples are shown in Figures 7 and 8.

In Figure 7, there are 3 classes: class A, class B, and class C, respectively. Then the new sample instance is added to the sample set to be tested, represented by N. If we use coarse granularity, then B and C are classified into one class, as shown in Figure 7. Using the barycenter of B and C to represent new class D, when we choose the distance between classes as similarity measure function, by calculating, we find that N is closer to the barycenter of B and C, and N is misclassified into class D. Nevertheless, when we use a fine granularity, as shown in Figure 8, classes B and C become independent class, and, using the same similarity measure function, we find that N is closer to A than the other two classes, so we put it correctly in class A. Hence, we can get the result that appropriate granularity is vital to the classification for resources.

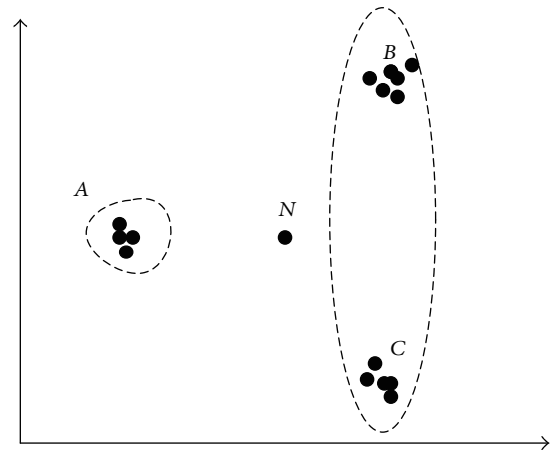


Figure 7: Classification based on uniform granularity.

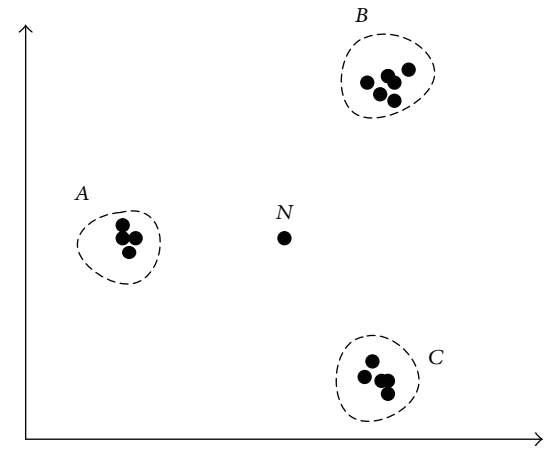


Figure 8: Classification based on nonuniform granularity.

4.2. Classification Based on Nonuniform Granularity in Social Interest. In this paper, social interest resources mainly refer to video, audio, image, text, blog posts, and so forth, which are all widely used on the net. These resources can reflect user social interest, and their scale is large. Algorithm 1 aims to classify these resources.

4.2.1. Framework of the Algorithm. Firstly, we execute clustering algorithm on the social interest resources, and the clustering pedigree chart can be obtained. By using certain classification cutting value to cut it, we can get different branches. Then, we repeat cutting it with finer cutting value (granularity) until stopping conditions are satisfied. Finally, we can get classification results. The framework of classification algorithm based on nonuniform granularity is shown as Figure 9.

Inputs:
U: Social interest resource pool
K: Required number of clusters
τ: Cutting threshold
Outputs: Classification result set: Ψ, cluster number: n.
(1) % **initialize**: choose appropriate classification threshold τ_0 $(\tau_0 > \tau)$ in the light of the pedigree chart.
(2) $V = \varnothing,\ S = \varnothing,\ Z = \varnothing,\ d_{\text{cut}} = \tau_0$;
(3) Perform a hierarchy clustering algorithm to social interest set: U;
(4) A clustering pedigree chart and a hierarchy clustered set: V is obtained;
(5) While $n < K$ and $d_{\text{cut}} > \tau$ do
(6) Do Cutting operation to V
(7) Computer cutting ratio: x, expected cluster ratio: $f(x)$, current cluster ratio: r
(8) if (indivisible subset: Z in V exists)
(9) then {
(10) $V = V \setminus Z,\ S = S \cup Z$;
(11) Update ΔC;
(12) }
(13) Calculate cutting step value: λ;
(14) if $r > f(x)$
(15) then $d_{\text{cut}} = d_{\text{cut}} - \lambda$;
(16) else $d_{\text{cut}} = d_{\text{cut}} + \lambda$;
(17) End while
(18) $\Psi = V \cup S$;
(19) $n = N(V) + N(S)$;
(20) Return Ψ, n

Algorithm 1: Algorithm NGSID: a classification based on nonuniform granularity in social interest detection.

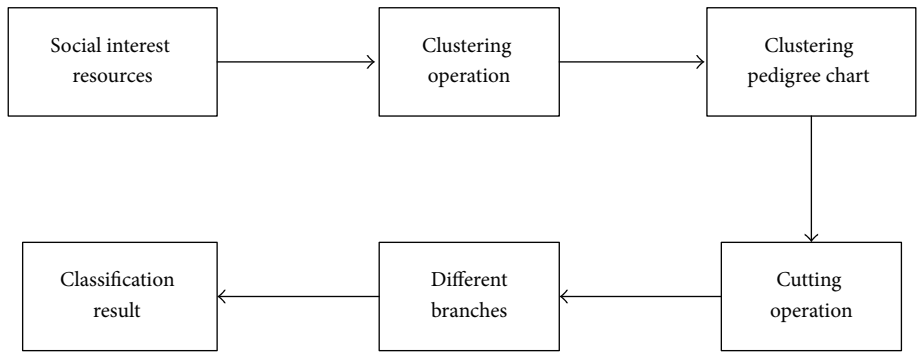


Figure 9: Framework of nonuniform granularity classification.

4.2.2. Algorithm Description. According to the features of social interest resources, we can get a granular system $G = (U, D, F)$, where G denotes resources pool, U denotes all resources in the resources pool such as blog posts, news reports, and other Web texts, D is the description set of the resources, and F defines the relationship between the resources.

Definition 6 (coarse cutting and fine cutting). According to the cutting ratio, we divide the whole cutting process into two stages: coarse cutting stage and fine cutting stage. During the former stage, the cutting granule is coarse enough so as to form a basic division of the resources. Correspondingly the latter stage uses fine cutting granule to get a more detailed division.

Let x denote the ratio of progress of the cutting process, which is the remaining proportion of current cutting distance to initial maximum interclass distance.

$$x = 1 - \frac{d_{\text{cut}}}{C_{\text{max}}^{\text{init}}}. \tag{3}$$

Different from linear cutting process in which the cutting interval is fixed, our cutting process adopts nonuniform granularity-based cutting policy:

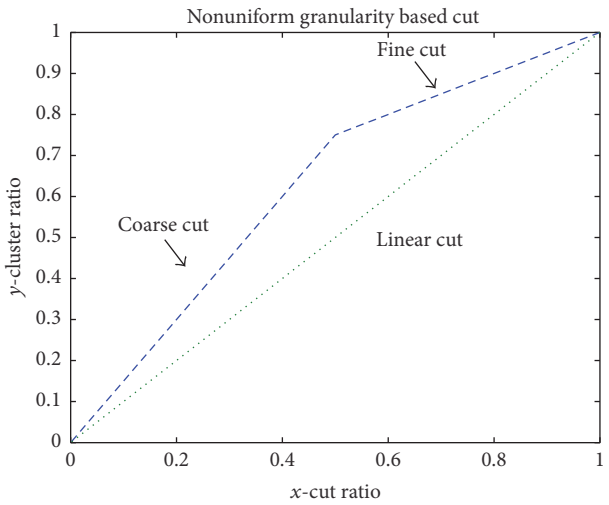


Figure 10: Nonuniform granularity-based cutting.

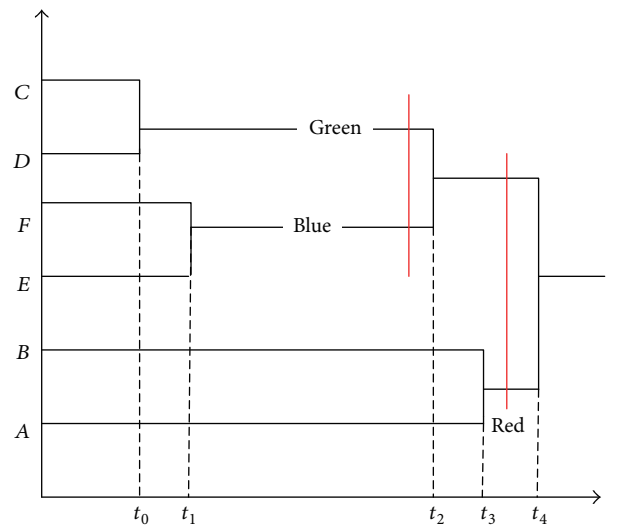


Figure 11: Classification results based on nonuniform granularity.

(i) $0 \le x \le 1/2$: coarse cutting stage, in which the cutting value shrinks quickly.

(ii) $1/2 \le x \le 1$: fine cutting stage, in which the cutting value shrinks relatively slowly.

During the cutting operation, leaf nodes that only belong to one class may be obtained. For these branched, no further cutting is needed. We use a transitional set S to store these clustered results. Ultimately, we have the final classification result set: $\Psi = V \cup S$ (see Notation Summary).

Definition 7. The cutting operation conforms to the object function with fast-start strategy proposed in [11]:

$$y = f(x) = \begin{cases} \frac{3}{2}x & 0 \le x \le \frac{1}{2} \\ \frac{1}{2}x + \frac{1}{2} & \frac{1}{2} \le x \le 1. \end{cases} \quad (4)$$

Here, variable y denotes the expected cluster ratio according to the current cutting progress, which acts as a baseline for analyzing the gap between expected cluster ratio and current cluster ratio.

According to the benchmark showed by the blue dash line (see Figure 10), the cutting value d_{cut} adjusts dynamically: when cluster ratio is below the expectation value, the cutting distance d_{cut} will be added by a current step value λ in the next round; when the cluster ratio is above the expectation value, this cutting operation will do vice versa.

Definition 8. In each stage, the cutting step value λ is in proportion to cutting boundary region ΔC. According to the scenario of Figure 10, the cutting step value is define as

$$\lambda = \frac{\Delta C}{K}. \quad (5)$$

Cutting boundary region is defined as

$$\Delta C = C_{\max} - C_{\min}. \quad (6)$$

In each round cutting boundary region adjusts dynamically in light of both cutting ratio and cutting result set. When new clustered subsets consisting of leaf nodes entirely are formed, these subsets are removed from the main set, then accordingly $C_{\max}, C_{\min}$ may change. Cutting boundary region ΔC will adjust correspondingly.

The size of cutting value d_{cut} represents granularity. Firstly we choose coarse granularity as much as possible when cutting, because if it can reflect the difference among the sample resources, there is no need to use fine granularity. Nevertheless, as for the elements in the boundary, their difference is not easy to be seen in large-grained world, and they are more likely to cause confusion with elements of other classes, so a relatively more fine granularity is needed, which can distinguish them clearly.

To illustrate the cutting process, we use the algorithm NGSID to deal with the classification of samples with clustering pedigree chart shown in Figure 2. Assuming $K = 3$, $t_0 = 0.2$, $t_1 = 0.3$, $t_2 = 0.8$, $t_3 = 0.9$, and $t_4 = 1.0$, let classification threshold $\tau = 0.2$, which rests with initial minimum interclass distance. Without loss of generality, the algorithm randomly chooses initial cutting value, which is slightly smaller than the maximum distance of the initial set and also bigger than τ; for example, $d_{\text{cut}} = 0.95$. In the first round, after cutting the pedigree chart, two clustered sets are formed: $V = \{\{C, D\}\}$, and $S = \{\{E\}, \{F\}, \{A\}, \{B\}\}$. With the boundary region ΔC shrinking from 0.8 to 0.7, the algorithm then adjusts the cutting step $\lambda = 0.2$ and gets a more fine cutting value $d_{\text{cut}} = 0.75$ for the next round. In the second round, using cutting value $d_{\text{cut}} = 0.75$ calculated by last iteration, three clustered sets are formed. Then we get the desired classification result set $\Psi = \{\{A, B\}, \{E, F\}, \{C, D\}\}$, meeting the demand of classification number. The result is shown in Figure 11, assuming that sample instances belonging to the same class are marked with the same color.

From Figure 11, we find that when classification threshold d_{cut} is larger than a certain value, sample instances of red,

blue, and green color are all classified into one class, and it is clear that this kind of classification results is not appropriate, from which we cannot get any useful information. Whereas, with a slight small threshold d_{cut}, some minor differences between sample instances can be well portrayed. Sample instances of red color are classified into one class, and the rest are instances form another class, yet it is still not quite appropriate, because instances of green and blue color are classified into the same class. A much smaller threshold d_{cut} can classify the sample instances into three classes correctly, namely, green class, blue class, and red class.

5. Experiment

5.1. Experiment Setup. Among the large scale resources of social interest such as video, audio, image, text, and blogs, we choose blogs widely used by people to do our experiments. There are two reasons to choose blogs: (1) the update frequency of blogs includes an indication of the writer's interest varying relatively fast and timely compared to other traditional media and, (2) in the meanwhile, its potential information structure reveals abundant semantic which provides an efficient way to detect bloggers' interest [16, 17]. In addition, recent automatic classification on blogs using NLP (Natural Language Processing) has drawn many researchers' attention and been proved to achieve a satisfied accuracy on social interest detection [18–20].

Due to its size and coverage, Wikipedia, a freely available online encyclopedia, can be utilized similar to an ontology or taxonomy to identify the topics discussed in a document. Based on the fact that Wikipedia is used in a large variety of research areas in Information Retrieval (IR) and Machine Learning, like categorization and clustering, NLP, machine translation, multimedia IR, entity search, and so forth [21], we use category tree structure of Wikipedia, which mainly consists of 12 categories: Physics and Nature, Arts and Culture, Philosophy and Thinking, Geography and Geology, History and Events, Mathematics and Logics, Society and Social Sciences, Economics, Health and Fitness, Technology and Sciences, Military, and Sports.

Experiment setup is as follows.

Step 1. We choose blog posts retrieved from a public dataset (http://www.nlpir.org/), which consists of a large collection of 214544 blog posts, 40050 bloggers, extracted from two of the most fashionable platforms in online social networking domain (http://t.qq.com/, http://weibo.com/). To avoid the classification inaccuracy due to sparse text, we restrict the lower boundary of blog instance to 100 words. We randomly choose five days' blogs (e.g., from November 1, 2011, to November 5, 2011) and then get 2472 blog posts in total. To make analysis simple, we assume that each blog belongs to one category. We finally classify the dataset into 12 categories manually.

Step 2. To reduce the large amount of noise due to blogs' shortness, marks, and irregular words, we do conventional text preprocessing involving the following steps: online text cleaning, words segmentation, white space removal, stop words removal, tokenizing, low frequency words removal, and so forth.

Step 3. We select training set and test set according to [22]. To improve the accuracy of training corpus, we use 12-fold cross-validation: the blog corpus above will be divided into 12 parts, one of which is selected as an open test set, and the remaining 11 parts are defined as training set and closed test set, making sure each of them can be an open test set turn by turn. Classification operation will be performed totally 12 times and the average accuracy for classification will be calculated.

Step 4. We use TFIDF as feature selection and construct VSM (Vector Space Model) to represent the feature of each blog and LSI (Latent Semantic Indexing) as feature extraction. Besides, distances between blog corpuses are measured by the cosine similarity.

5.2. Evaluation

5.2.1. Classification Effectiveness of the Algorithm. The proposed algorithm (NGSID) is applied to the dataset mentioned in Algorithm 1. We adopt three parameters related to classification accuracy: precision, recall, and F-Score to evaluate our algorithm's effectiveness.

Table 1 shows NGSID's classification accuracy for the given taxonomy used in our experiments. As we can see, NGSID works efficiently from the perspective of precision: the total precision can achieve 83.79% at the sample scale of 2472. Due to accurate feature representation in mathematical specific field, Mathematics and Logics outperforms other categories in terms of precision. When it comes to recall ratio, Arts and Culture has the lowest value 73.40%. Actually, according to the clustered result, the blogs in Arts and Culture are mainly classified into three categories: Arts and Culture (181), Economics (22), and Society and Social Sciences (27), where the latter two categories are mismatched.

The reason for NGSID accuracy performance lies in that, with fine granularity, it is beneficial for NGSID to embody the differences between categories. Subtle difference in similarity measure of sample instances will be reflected dedicatedly. Besides, we also attribute part of misclassification to the confusion caused by overlapping of correlative categories (e.g., certain concepts in Arts and Culture are overlapped in Society and Social Sciences), which will later induce additional manual classification work in common.

5.2.2. Comparisons with Other Algorithms. To validate the performance of the proposed algorithm (NGSID), we also apply diverse classification algorithms (NB, KNN, and SVM) to make comparisons.

As shown from Figure 12, the performance of SVM classifiers is better compared to the NB and KNN classifiers in almost all cases. But, for NGSID, at the beginning it performs worse than the other algorithms (NB, KNN, and SVM). That is, mainly relying on the case when granularity is too large, it will result in coarse classification, affecting the classification results, some details are probably ignored, and blog posts of various types will be classified into one class.

Table 1: Classification accuracy of NGSID.

Category	Category size	Precision	Recall	F-score
Arts and Culture	247	78.40%	73.40%	75.82%
Physics and Nature	100	82.50%	80.40%	81.44%
Philosophy and Thinking	64	87%	83.60%	85.27%
Geography and Geology	120	82.10%	79.60%	80.83%
History and Events	60	83.50%	80.80%	82.13%
Mathematics and Logics	26	90.10%	84.70%	87.32%
Society and Social Sciences	530	89.70%	85.20%	87.39%
Technology and Sciences	528	82.80%	78.02%	80.34%
Health and Fitness	92	83.40%	81.40%	82.39%
Military	25	82.90%	80.04%	81.44%
Sports	150	88.50%	86.15%	87.31%
Economics	530	80.30%	78.10%	79.18%

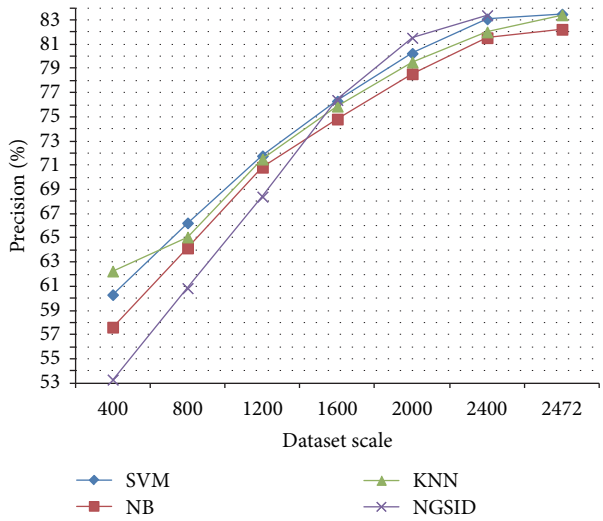


Figure 12: Precision comparisons based on the increasing sample scale.

Nevertheless, with the increasing scale of the dataset, its precision increases gradually with a higher increasing rate. At the scale of 1600, it achieves a better precision than NB and KNN, approximately equivalent precision to SVM. From then on, NGSID can achieve a better performance than SVM while still maintaining precision of 81.53% and above, benefiting from the dataset's gradually increasing scale. The reason is that as more blogs are continuously added to the dataset, more fine granularities are provided to make the cutting operations more precise.

While, at the scale range from 2000 to 2400, the classification precision of NGSID outperforms other algorisms, its increasing rate is steadily dropping with the trend of approaching a convergent value.

From the results of experiments, we can draw the following conclusions: by using NGSID we can get a better classification efficiency at a cost of larger sample scale. For the aspect of classification accuracy, NGSID's performance is sensitive to the size of granularity: with a coarse granularity formed by sparse sample instances, which will affect the classification results, some details are probably ignored, and blogs of various types are classified into one class by NGSID. It leads to the result that the performance of NGSID is worse than those of SVM, NB, and KNN. However, with a fine granularity formed by intensive sample instances, blogs of various types will be classified into their dedicated categories, and the category varieties will increase accordingly. Thus NGSID can achieve an equivalent or even better performance compared to SVM and other classifiers.

6. Conclusion and Future Work

For the classification problem of massive amount and various types of resources in social network, we present an efficient classification algorithm based on nonuniform granularity. Clustering algorithm is used to generate clustering pedigree chart. And most resources can be classified into the correct class by modifying cutting value d_{cut} (granularity) to cut the clustering pedigree chart. The size of cutting value is vital to the performance of the proposed algorithm. Through comparing with existing typical algorithms, we show that our proposed algorithm can improve the performance of classification.

NGSID is flexible and can be extended to medium-sized resource classification widely. Such application scenarios use similarity degree in common to weigh the comparability between resources. However, NGSID will meet its efficiency constraint when the resources' volume or quantity increases extremely fast, which will bring much more complex work in massive similarity calculation or clustering pedigree chart construction in pretreatment stage.

Our work will continue in the following directions. Firstly, the impact of the granularity size and initial cutting threshold on the classification performance will be analyzed. Furthermore, for large scaled resources, the impact of huge dimensions on accuracy also essentially needs to be addressed. Finally, with more social interest datasets continuously added (e.g., video, audio, and image), analyses and simulations will be further carried out to prove the algorithm's adaptability.

Notation Summary

λ: Current step value
x: Cutting ratio
r: Cluster ratio, which denotes the proportion of the number of clustered classes to the total class
d_{cut}: Current cutting value
U: A hierarchy clustered set of social interest
V: Clustered result set after cutting operation on U
Z: An indivisible clustered set, consisting of leaf nodes entirely, with each branch only belonging to one class
S: A transitional result set, storing the result set Z during cutting operation
$N(V)$: Number of clusters in V
$N(S)$: Number of clusters in S
τ: A predefined cutting threshold to avoid overcutting, referring to a minimum cutting distance between branches
$C_{\max}$: The maximum interclass distance
$C_{\min}$: The minimum interclass distance
ΔC: Cutting boundary region
K: Required number of clusters.

Conflicts of Interest

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

Acknowledgments

This work was financially supported by the Natural Science Foundation of China (Project no. 61271254).

References

[1] M. Kumar, "Automatic identification of user interest from social media," *Text Classification*, 2015.

[2] R.-L. Liu, "Interactive high-quality text classification," *Information Processing & Management*, vol. 44, pp. 1062–1075, 2008.

[3] J. Shao, *Information Granularity Computing Based on Rough Sets*, Institute of Automatics, Chinese Academy of Sciences, Beijing, China, 2000.

[4] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, no. 2, pp. 111–127, 1997.

[5] Y. Yao and B. V. Dasarathy, "Information tables with neighborhood semantics," in *AeroSense 2000*, pp. 108–116, 2000.

[6] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[7] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning*, pp. 137–142, 1998.

[8] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *Proceedings of the AAAI-98 workshop on learning for text categorization*, pp. 41–48, 1998.

[9] H. T. Ng, W. B. Goh, and K. L. Low, "Feature selection, perceptron learning, and a usability case study for text categorization," *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, vol. 31, no. 1, pp. 67–73, 1997.

[10] W. Shao, Q. SHen, X. Jin, and L. Huang, "A novel granularity-based classification in cloud environment," in *Proceedings of the 2017 2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)*, pp. 687–690, Beijing, China, March 2017.

[11] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. De Amorim, "Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding," *Pervasive and Mobile Computing*, vol. 8, no. 5, pp. 682–697, 2012.

[12] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, pp. 341–356, 1982.

[13] G.-Y. Wang, *Rough set theory and knowledge acquisition*, vol. 1, Xi'an Jiaotong University Press, Xi'an, China, 2001.

[14] Y. Y. Yao and C. Yaohua, "Rough set approximations in formal concept analysis," in *Proceedings of the IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS '04*, vol. 1, pp. 73–78, 2004.

[15] L. Z. B. Zhang, *Theory of Problem Solving and Its Ap2 Publication*, Tsinghua University Press, Beijing, China, 1990.

[16] S. J. Barry, "Using social media to discover public values, interests, and perceptions about cattle grazing on park lands," *Environmental Management*, vol. 53, no. 2, pp. 454–464, 2014.

[17] H.-J. Wu, I.-H. Ting, and K.-Y. Wang, "Combining social network analysis and web mining techniques to discover interest groups in the blogspace," in *Proceedings of the 2009 4th International Conference on Innovative Computing, Information and Control, ICICIC 2009*, pp. 1180–1183, Kaohsiung, Taiwan, December 2009.

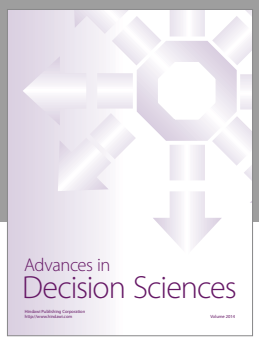
[18] A. Bakliwal, P. Arora, A. Patil et al., "Towards enhanced opinion classification using NLP techniques[C]," in *Proceedings of the 5th international joint conference on natural language processing (IJCNLP)*, pp. 101–107, Chiang Mai, Thailand, 2011.

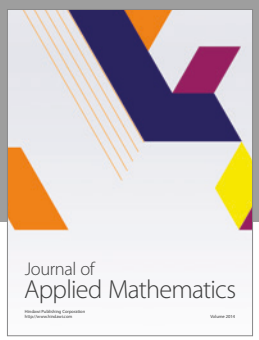
[19] D. Ikeda, H. Takamura, and M. Okumura, "Semi-supervised learning for blog classification," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 1156–1161, July 2008.

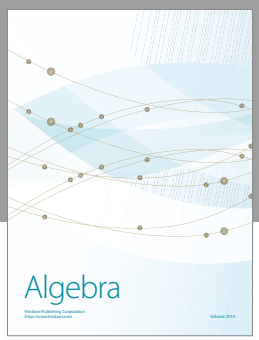
[20] N. Xiaochuan, W. Xiaoyuan, and Y. Yong, "Automatic identification of Chinese weblogger's interests based on text classification," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI'06*, pp. 247–253, China, December 2006.

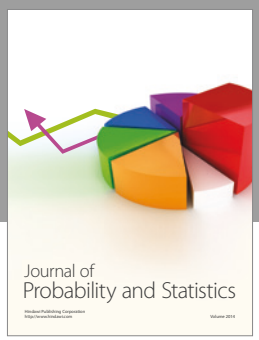
[21] P. Schönhofen, "Identifying document topics using the Wikipedia category network," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI'06*, pp. 456–462, China, December 2006.

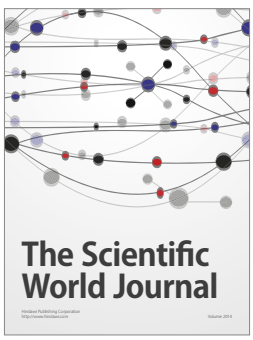
[22] H. X. Jing, *Retrieval, Classification and Summarization of Large Scale Chinese Text*, Fudan University, 1998.

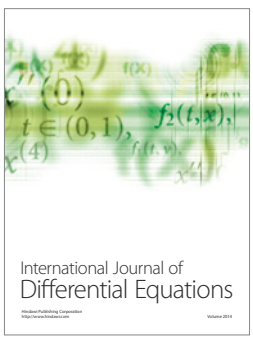


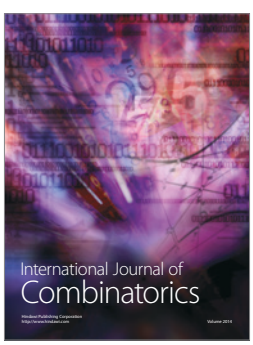


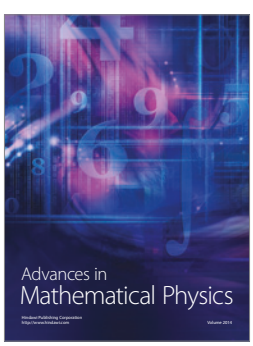


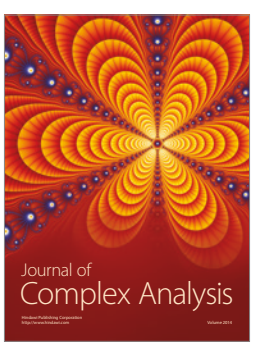


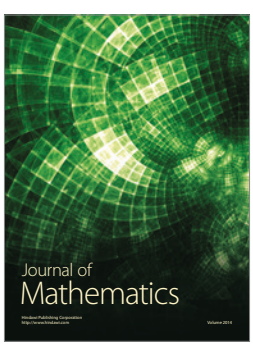


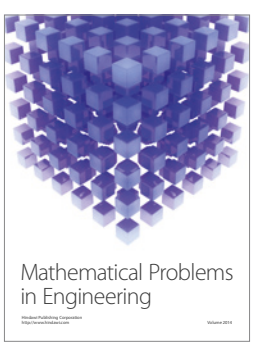


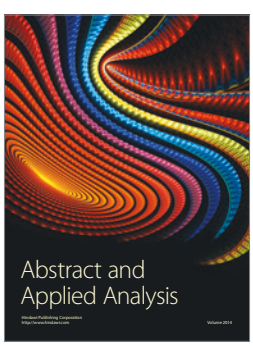


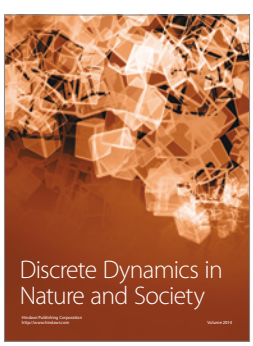


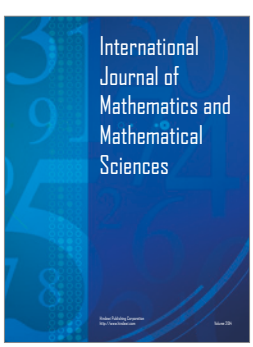


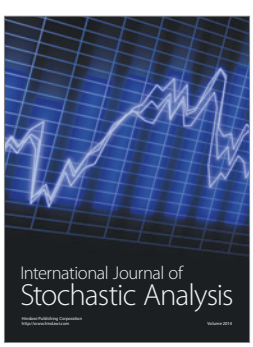












Hindawi

Submit your manuscripts at
https://www.hindawi.com