International Journal of Computational Intelligence Systems
Vol. **13(1)**, 2020, *pp.* 1218–1226
DOI: https://doi.org/10.2991/ijcis.d.200808.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Tolerance Rough Set-Based Bag-of-Words Model for Document Representation

Dong Qiu[1,*], Haihuan Jiang[1], Ruiteng Yan[2]

[1]*College of Science, Chongqing University of Posts and Telecommunications, Nanan, Chongqing, 400065, P.R. China*
[2]*School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Nanan, Chongqing, 400065, P.R. China*

ARTICLE INFO

Article History
Received 12 May 2020
Accepted 04 Aug 2020

Keywords
Document representation
Tolerance rough set
Bag-of-Words

ABSTRACT

Document representation is one of the foundations of natural language processing. The bag-of-words (BoW) model, as the representative of document representation models, is a method with the properties of simplicity and validity. However, the traditional BoW model has the drawbacks of sparsity and lacking of latent semantic relations. In this paper, to solve these mentioned problems, we propose two tolerance rough set-based BOW models, called as TRBoW1 and TRBoW2 according to different weight calculation methods. Different from the popular representation methods of supervision, they are unsupervised and no prior knowledge required. Extending each document to its upper approximation with TRBoW1 or TRBoW2, the semantic relations among documents are mined and document vectors become denser. Comparative experiments on various document representation methods for text classification on different datasets have verified optimal performance of our methods.

1. INTRODUCTION

With the explosive growth of the Internet, countless text data are accumulated constantly. In addition, unlike numerical data belonging to the structured data type, document or text data are unstructured data. Unstructured data are not appropriate to be directly applied in machine learning or deep learning algorithms. As the basis of natural language processing (NLP) and text mining tasks, efficient text or document representation is particularly important. The main challenges on document representation are the ways of transforming unstructured text data into structured data. For a good document representation, on the one hand, it should be able to truly reflect the content of the document, on the other hand, it should have the ability to distinguish different documents. Additional, it has optimal performance in some NLP applications such as text classification, information retrieval and text clustering.

The bag-of-words (BoW) model [1] is a representative document representation method, which has been widely used in information retrieval, text classification [2] and sentiment analysis [3]. In the BoW model, the word order, grammar and syntax of the document are ignored and every document is only regarded as a set of words, or a combination of words. The emergence of each word in the text is independent and does not depend on the appearances of other words. The BoW model takes all the non-repetitive words of all the documents in the dataset as basis terms. Each document is denoted by a fixed dimension vector, the length of which is equal to the number of the basis terms. Each component of the vector corresponds to the frequency of the basis word that appears in the document. The BoW model has the virtue of briefness and efficiency. However, it suffers from the defects of sparsity and lack of latent semantic relevance. These issues have been addressed many times in earlier research, but we are committed to solving this problem without any prior knowledge and without supervision. Hence, we apply tolerance rough set theory [4] to improve the traditional BoW model, and thus propose two new tolerance rough set-based BOW models, TRBoW1 and TRBoW2, which can expand the semantic space of documents. And they are unsupervised and not need any prior knowledge.

There are many researches have been done on document representation. Among them, the BoW model is one of the most classical method. Owing to that the BoW model is sparse, high-dimensional and lack of latent semantics, some other models emerge as the times require from distinctive perspectives, such as feature selection algorithms [5–8], weight calculation algorithms [9–12] and dimensionality reduction algorithms [13–18]. Text feature selection algorithms contain document frequency [7], Chi square (χ^2) [6], information gain (IG) [5] and mutual information (MI) [8]. To measure the importance of a feature item in document representation better, some works on feature weight calculation have presented, including term frequency inverse document frequency (TF-IDF) [12], entropy [10] and term frequency inverse word frequency (TF-IWF) [9]. TF-IDF is a commonly used one [11]. In order to mine the correlations among words and alleviate the problem of high dimensionality, many scholars have developed a series of latent topic models, the earliest of which is the latent semantic analysis (LSA) model [14]. LSA projects documents into low-dimensional latent semantic

*Corresponding author. Email: dongqiumath@163.com

space by singular value decomposition (SVD) of word frequency-document matrix [19]. Hofmann *et al.* established a probabilistic latent semantic analysis (PLSA) model to solve the polysemy problem of LSA [15,16]. On the basis of PLSA, Blei *et al.* introduced the Dirichlet prior distribution and proposed the Latent Dirichlet Allocation (LDA) [13]. Some other scholars generalized the LDA model and put forward Gaussian LDA model [20], Latent Feature Topic Modeling (LFTM) model [21] and Topical Word Embedding (TWE) model [22]. Although the algorithms above have been successfully used for document representation and played the role of dimensionality reduction, they are still unable to capture the real semantics among words.

Methods based on neural network and deep learning arose at the historic moment. However, majority of them are supervised or need prior knowledge. A sentence encoder-decoder model was proposed in [23], called Skip-Thought model, which learned the representation through predicting the following sentence. But it need a lot of training. Wu *et al.* [24] represented documents based on phrase embeddings by parsing, in which three Phrase2Vec methods are constructed on the basis of Word2Vec. To solve the problem of sparsity, Yao *et al.* [25] utilized word sematic similarity by constructing a neural probabilistic language model. Gao *et al.* [26] proposed the CCTSenEmb model by obtaining the association between adjacent sentences in the process of sentence prediction. The Hybrid-WikiBoC approach was proposed to improve the performance of BoW in [27], taking Wikipedia as the background knowledge. To overcome the limitation of sparsity and inability to dig out semantic significance behind words of the BoW model, Zhao *et al.* combined the fuzzy system with BoW, and proposed a novel fuzzy bag-of-words (FBoW) model [28]. FBoW model converted the original hard mapping into fuzzy mapping, representing the component of document representation vector as the similarity between words and basis terms. Furthermore, the fuzzy bag-of-word clusters (FBoWC) model was developed to solve the problem of high dimension by clustering the basis terms in [28]. However, it is based on the prior knowledge, which needs the word embedding to capture the semantic relevance.

This paper is organized as follows: The tolerance rough set model is reviewed in Section 2. Section 3 presents our proposed tolerance rough set-based BOW models detailedly. Section 4 demonstrates the experimental results. The discussion and analysis of the experimental results is given in section 5. In Section 6, some conclusions are came to.

2. TOLERANCE ROUGH SET

In this section, the tolerance rough set model is reviewed in brief.6pt

Rough set theory, put forward by Polish scholar Pawlak in 1982 [29], is applied to process the problem of uncertainty and fuzziness. It provides both theories and technologies for researchers in a broad variety of fields of artificial intelligence such as data mining, machine learning and NLP. Rough set, which is based on the equivalence relation, uses a pair of concepts, upper approximation and lower approximation, to measure the relationships of one certain object and a set X. The relationships can be represented as belonging to, possibly belonging to and not belonging to. For the shortcoming of traditional rough set model, on different circumstances and needs, researchers have developed plenty of extended rough set models, like probabilistic rough set model [30], decision rough set model [31] and tolerance rough set model [4]. For the reason that equivalence relation contains three properties of reflexivity, symmetry and transitivity, in which the limitation of transitivity leads to the inapplicability in some cases, Skowron *et al.* modified the original equivalence relation to tolerance relation, and the corresponding approximation space to tolerance approximation space [4].

A tolerance space is represented by a quadruple $\mathfrak{R} = (U, I, \nu, P)$ in [32]. Considering that the structural function P has no practical meaning in the definition and has no impact on the whole content, we simplify the quadruple to a triple $\mathfrak{R} = (U, I, \nu)$. In the triple, $U = \{x_1, x_2, \dots, x_n\}$ is the universe of all the objects, $I : U \to 2^U$ is an uncertainty function, and $I(x)$ is a tolerance class on x. If an object shares similar information with x, it belongs to $I(x)$. $\nu : 2^U \times 2^U \to [0, 1]$ is a vague inclusion. Any function having the properties of reflexivity and symmetry can be an uncertainty function $I(x)$, which can be explained that for any $x, y \in U, x \in I(x)$ iff $x \in I(y)$. The vague inclusion ν has the property of monotonicity, i.e., for any $X, Y, Z \subseteq U$ and $Y \subseteq Z$, $\nu(X, Y) \leq \nu(X, Z)$. It measures the degree of inclusion of sets, whether a set X contains the tolerance class $I(x)$ of an object $x \in U$ [32]. The upper approximation $\mathcal{U}(\mathfrak{R}, X)$ and the lower approximation $\mathcal{L}(\mathfrak{R}, X)$ of any $X \subseteq U$ are defined as follows:

$$\mathcal{U}(\mathfrak{R}, X) = \{x \in U \mid P(I(x)) = 1 \& \nu(I(x), X) > 0\}, \tag{1}$$

$$\mathcal{L}(\mathfrak{R}, X) = \{x \in U \mid P(I(x)) = 1 \& \nu(I(x), X) = 1\}. \tag{2}$$

3. TOLERANCE ROUGH SET-BASED BOW MODELS

In this section, we describe the proposed TRBoW models in detail. Now we introduce the definition of the triple of tolerance rough set in document representation. The notations in this section are listed in Table 1, where the vectors and matrixes are written as bold formatting.

Suppose that $D = \{d_1, d_2, \dots, d_v\}$ is the collection of all the documents in the corpus, where v is the document size. $W = \{w_1, w_2, \dots, w_n\}$ denotes all the non-repetitive words in the document corpus, also called basis terms, where n is the vocabulary size. Take the universe as W. For any document $d_i \in D$, let $\boldsymbol{d_i}$ be the representation vector of d_i, which is n-dimensional and defined as $\boldsymbol{d_i} = [wo_{i1}, wo_{i2}, \dots, wo_{ij}, \dots, wo_{in}]$, where

$$wo_{ij} = \begin{cases} 1 & if \quad w_j \in d_i \\ 0 & otherwise. \end{cases} \tag{3}$$

Let c_{ij} denote the times of words w_i and w_j occurring in the same document. For a positive threshold θ, the uncertainty function $\boldsymbol{I}_\theta$ projects each word into a vector as

$$\boldsymbol{I}_{\theta(w_i)} = \left[A(c_{i1}) \; A(c_{i2}) \cdots \; A(c_{in}) \right], \tag{4}$$

where $A(c_{ij})$ $(1 \leq i \leq n, 1 \leq j \leq n)$ is defined as

$$A(c_{ij}) = \begin{cases} 1 & if \quad c_{ij} \geq \theta, or \; i = j \\ 0 & otherwise. \end{cases} \tag{5}$$

Table 1 | Table of notations.

Notation	Description	Notation	Description
D	The document corpus	W	The universe
v	Document size	n	Basis terms size
$\mathbf{d_i}$	The vector representation of document d_i	c_{ij}	The co-occurrence times of words w_i and w_j
θ	Co-occurrence threshold	$\boldsymbol{I}$	n-dimensional unit vector
$\mathfrak{R}$	The tolerance space	$\mathbf{X}$	The document representation matrix by BoW
$\mathbf{I}_{\theta(W)}$	The uncertainty matrix of the basis terms	$\boldsymbol{\mu}(\mathbf{W},\mathbf{D})$	The fuzzy membership matrix of the words and documents
$\mathcal{U}(\mathfrak{R}, d_i)$	The upper approximation set of the document d_i	$\mathcal{L}(\mathfrak{R}, d_i)$	The upper approximation set of the document d_i
$\mathbf{U}(\mathfrak{R}, D)$	The upper approximation matrix of the documents	$\mathbf{L}(\mathfrak{R}, D)$	The lower approximation matrix of the documents
$\mathbf{M_1}(\mathbf{D})$	The document representation matrix by TRBoW1	$\mathbf{M_2}(\mathbf{D})$	The document representation matrix by TRBoW2

Then the uncertainty matrix $\boldsymbol{I}_{\theta(W)}$ is defined as follows:

$$\boldsymbol{I}_{\theta(W)} = \begin{bmatrix} 1 & A(c_{12}) & \cdots & A(c_{1n}) \\ A(c_{21}) & 1 & \cdots & A(c_{2n}) \\ \vdots & \vdots & \vdots & \vdots \\ A(c_{n1}) & A(c_{n2}) & \cdots & 1 \end{bmatrix}. \tag{6}$$

For any two documents d_i and d_j, the vague inclusion function is defined as

$$v(\boldsymbol{d_i}, \boldsymbol{d_j}) = \frac{\boldsymbol{d_i} \cdot \boldsymbol{d_j}}{\boldsymbol{d_i} \cdot \boldsymbol{I}}, \tag{7}$$

where $\boldsymbol{I}$ is an n-dimensional unit vector. Then we define the fuzzy membership function μ for $w_j \in W, d_i \in D$ as

$$\mu_{ij} = v(\boldsymbol{I}_{\theta(w_j)}, \boldsymbol{d_i}) = \frac{\boldsymbol{I}_{\theta(w_j)} \cdot \boldsymbol{d_i}}{\boldsymbol{I}_{\theta(w_j)} \cdot \boldsymbol{I}}. \tag{8}$$

And the fuzzy membership matrix of the whole corpus is represented by

$$\boldsymbol{\mu}(\boldsymbol{W}, \boldsymbol{D}) = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{v1} & \mu_{v2} & \cdots & \mu_{vn} \end{bmatrix}. \tag{9}$$

Then the upper approximation $\mathcal{U}(\mathfrak{R}, d_i)$ and the lower approximation $\mathcal{L}(\mathfrak{R}, d_i)$ of any $d_i \in D$ are expressed as

$$\mathcal{U}(\mathfrak{R}, d_i) = \{w_j \in W \mid \mu_{ij} > 0\}, \tag{10}$$

$$\mathcal{L}(\mathfrak{R}, d_i) = \{w_j \in W \mid \mu_{ij} = 1\}. \tag{11}$$

Hence the upper approximation matrix $\boldsymbol{U}(\mathfrak{R}, D)$ and lower approximation matrix $\boldsymbol{L}(\mathfrak{R}, D)$ of the set of documents D are respectively as follows:

$$\boldsymbol{U}(\mathfrak{R}, D) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{v1} & u_{v2} & \cdots & u_{vn} \end{bmatrix}, \tag{12}$$

$$\boldsymbol{L}(\mathfrak{R}, D) = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ l_{21} & l_{22} & \cdots & l_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ l_{v1} & l_{v2} & \cdots & l_{vn} \end{bmatrix}, \tag{13}$$

where

$$u_{ij} = \begin{cases} 1 & if \quad \mu_{ij} > 0 \\ 0 & otherwise, \end{cases} \tag{14}$$

$$l_{ij} = \begin{cases} 1 & if \quad \mu_{ij} = 1 \\ 0 & otherwise. \end{cases} \tag{15}$$

As for the original BoW model, the representation matrix is defined as

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{v1} & x_{v2} & \cdots & x_{vn} \end{bmatrix}, \tag{16}$$

where x_{ij} denotes the number of occurrence times of the word w_j in the document d_i.

Now, we propose two different tolerance rough set-based BOW models, called as TRBoW1 and TRBoW2 according to different weight calculation methods. In the TRBoW1 model we strengthen the weights of words exactly contained in the document. Thus the representation matrix is represented by

$$\boldsymbol{M}_{\mathbf{1}(D)} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ m_{v1} & m_{v2} & \cdots & m_{vn} \end{bmatrix}, \tag{17}$$

where

$$m_{ij} = \begin{cases} x_{ij} \times \mu_{ij} & if \quad w_j \in d_i \\ \mu_{ij} & if \quad w_j \in \mathcal{U}(\mathfrak{R}, d_i) \setminus d_i \\ 0 & if \quad w_j \notin \mathcal{U}(\mathfrak{R}, d_i). \end{cases} \tag{18}$$

In the TRBoW2 model we take the membership matrix as the document representation matrix directly

$$\boldsymbol{M}_{\mathbf{2}(D)} = \boldsymbol{\mu}(\boldsymbol{W}, \boldsymbol{D}) = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{v1} & \mu_{v2} & \cdots & \mu_{vn} \end{bmatrix}. \tag{19}$$

Algorithm 1 is the detailed procedure of the proposed methods.

Example 1. Assume that the corpus is consisted of the following four documents.

1. *It was the best of times.*
2. *It was the worst of times.*

Algorithm 1: Tolerance Rough Set-Based Bag-of-Words Models

Input: A set of text data including m **classes** $K = \{k_1, k_2, \cdots, k_m\}$ **and in total** v **documents** $D = \{d_1, d_2, \cdots, d_v\}$; **the class label** k_i of each document d_i.
Parameters: The co-occurrence threshold: θ; **the dimensionality of vectors for text representation:** h.
Output: Document representation matrix: $\boldsymbol{M}$.
1: Count the term frequencies of all the words occurring in the corpus, and choose the most frequent l words as the basis terms;
2: Calculate the uncertainty function $\boldsymbol{I}_{\theta(w_i)}$ of each basis term according to (4) and get the uncertainty matrix according to (6);
3: The fuzzy membership degree μ_{ij} of each basis word in document d_i can be obtained according to (8), $1 \le i \le v$. And the fuzzy membership matrix $\boldsymbol{\mu}(\boldsymbol{W}, \boldsymbol{D})$ is constructed according to μ_{ij};
4: Acquire the upper approximation $\mathcal{U}(\mathfrak{R}, d_i)$ of each document $d_i \in D$ according to (10). And obtain the upper approximation vector $\boldsymbol{U}(\mathfrak{R}, d_i)$ of each document in the text corpus;
5: if the BoW model is performed then
6: Let $\boldsymbol{M} = \boldsymbol{X}$, which is the document representation using the BoW model;
7: else if TRBoW1 model is performed **then**
8: Apply formula (17) and (18) to compute the representation matrix $\boldsymbol{M}_{\mathbf{1}(D)}$, let $\boldsymbol{M} = \boldsymbol{M}_{\mathbf{1}(D)}$;
9: else if TRBoW2 model is performed **then**
10: Let $\boldsymbol{M} = \boldsymbol{M}_{\mathbf{2}(D)} = \boldsymbol{\mu}(\boldsymbol{W}, \boldsymbol{D})$, using the membership degree as the weight directly, as shown in formula (19);
11: end
12: return $\boldsymbol{M}$

3. *It was the age of wisdom.*
4. *It was the age of foolishness.*

The universe is the set of all the basis terms, W ={it, was, the, best, of, times, age, worst, wisdom, foolishness} and the vocabulary size is 10. If the BoW model is performed, the document representation matrix is

$$\boldsymbol{X} = \begin{bmatrix} 1&1&1&1&1&1&0&0&0&0 \\ 1&1&1&0&1&1&0&1&0&0 \\ 1&1&1&0&1&0&1&0&1&0 \\ 1&1&1&0&1&0&1&0&0&1 \end{bmatrix}.$$

If the TF-IDF model is performed, the document representation matrix is

$$\boldsymbol{T} = \begin{bmatrix} 0.32&0.32&0.32&0.61&0.32&0.48&0&0&0&0 \\ 0.32&0.32&0.32&0&0.32&0.48&0&0.61&0&0 \\ 0.32&0.32&0.32&0&0.32&0&0.48&0&0.61&0 \\ 0.32&0.32&0.32&0&0.32&0&0.48&0.&0&0.61 \end{bmatrix}.$$

If the TRBoW1 model is performed, setting $\theta = 2$, we have

$$\boldsymbol{I}_{\mathbf{2}(W)} = \begin{bmatrix} 1&1&1&0&1&1&1&0&0&0 \\ 1&1&1&0&1&1&1&0&0&0 \\ 1&1&1&0&1&1&1&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0 \\ 1&1&1&0&1&1&1&0&0&0 \\ 1&1&1&0&1&1&1&0&0&0 \\ 1&1&1&0&1&1&1&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \end{bmatrix}.$$

Then the membership matrix is

$$\boldsymbol{\mu}(\boldsymbol{W}, \boldsymbol{D}) = \begin{bmatrix} \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&1&\frac{5}{6}&1&\frac{4}{5}&0&0&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&1&\frac{4}{5}&1&0&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&\frac{4}{5}&1&0&1&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&\frac{4}{5}&1&0&0&1 \end{bmatrix}.$$

And the upper approximation matrix is

$$\boldsymbol{U}(\mathfrak{R}, D) = \begin{bmatrix} 1&1&1&1&1&1&1&0&0&0 \\ 1&1&1&0&1&1&1&1&0&0 \\ 1&1&1&0&1&1&1&0&1&0 \\ 1&1&1&0&1&1&1&0&0&1 \end{bmatrix}.$$

So the document representation by the TRBoW1 model is

$$\boldsymbol{M}_{\mathbf{1}(D)} = \begin{bmatrix} \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&1&\frac{5}{6}&1&\frac{4}{5}&0&0&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&1&\frac{4}{5}&1&0&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&\frac{4}{5}&1&0&1&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&\frac{4}{5}&1&0&0&1 \end{bmatrix}.$$

If the TRBoW2 model is performed, setting $\theta = 2$, the document representation matrix is

$$\boldsymbol{M}_{\mathbf{2}(D)} = \boldsymbol{\mu}(\boldsymbol{W}, \boldsymbol{D}) = \begin{bmatrix} \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&1&\frac{5}{6}&1&\frac{4}{5}&0&0&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&1&\frac{4}{5}&1&0&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&\frac{4}{5}&1&0&1&0 \\ \frac{5}{6}&\frac{5}{6}&\frac{5}{6}&0&\frac{5}{6}&\frac{4}{5}&1&0&0&1 \end{bmatrix}.$$

Because each basis term appears no more than once in each document, the representation matrixes are the same by the TRBoW1 and TRBoW2 model in Example 1. For the practical corpus is large enough, a word often occurs many times in the same document, the representation matrixes by the two models will be different.

As can be indicated from Example 1, the TF-IDF model emphasizes the weight of each word more than the BoW model. Besides, by applying the rough set theory, our methods capture some latent semantic information. For example, the first two sentences are represented by the first two row vectors of the document representation matrixes, respectively. We can see that $\boldsymbol{X}_{17} = 0$ and $\boldsymbol{T}_{17} = 0$ but $\boldsymbol{M}_{17} = \frac{4}{5}$. In other words, the membership degree of the word "age" belonging to the first two documents is 0 in the BoW and TF-IDF model, however the membership degree of the word "age" belonging to the first two documents is $\frac{4}{5}$ in the TRBoW1 and TRBoW2 model. Hence our models mine the meaning of "age" that may exist in the first two documents potentially.

In the BoW model, the document representation of the first document is expressed as (1 1 1 1 1 1 0 0 0 0); in the TRBoW1 and

TRBoW2 model, the document representation of the first document is expressed as $(\frac{5}{6}\ \frac{5}{6}\ \frac{5}{6}\ 1\ \frac{5}{6}\ 1\ \frac{4}{5}\ 0\ 0\ 0)$. It can be seen that the sparsity of BoW model and TF-IDF model is slightly alleviated and the document matrixes become denser. Furthermore, if we set θ as 1, the document representation matrix of our models is

$$\begin{bmatrix} \frac{3}{5} & \frac{3}{5} & \frac{3}{5} & 1 & \frac{3}{5} & 1 & \frac{4}{7} & \frac{5}{6} & \frac{2}{3} & \frac{2}{3} \\ \frac{3}{5} & \frac{3}{5} & \frac{3}{5} & \frac{5}{6} & \frac{3}{5} & 1 & \frac{4}{7} & 1 & \frac{2}{3} & \frac{2}{3} \\ \frac{3}{5} & \frac{3}{5} & \frac{3}{5} & \frac{2}{3} & \frac{3}{5} & \frac{4}{7} & 1 & \frac{2}{3} & 1 & \frac{5}{6} \\ \frac{3}{5} & \frac{3}{5} & \frac{3}{5} & \frac{2}{3} & \frac{3}{5} & \frac{4}{7} & 1 & \frac{2}{3} & \frac{5}{6} & 1 \end{bmatrix}.$$

It is evident from the matrix above that the sparsity of BoW model is improved considerably. In fact, the sparsities of document representation matrixes in the TRBoW1 and TRBoW2 model are depended on the value of θ, that is, the larger the threshold is, the sparser the representation matrixes are.

4. EXPERIMENTS

In this section, in order to evaluate the performance of our methods, we compare our proposed models with the BoW model, TF-IDF model and average embeddings (AEs) by conducting some popular text classification tasks [1,12]. For better comparison, the dimension of BoW and TF-IDF is also set as 1000. AE is a neural network based method, which represents a document by averaging out the word embedding of each word included in the document. The pretrained word2vec word embeddings are used in the experiments [33], which are 300 dimensional. The other parameter settings are the same.

4.1. Datasets and Preprocessing

In order to verify the validity of our methods and experiments, we will use the commonly used benchmark datasets, Reuters, BBC and BBCsport in text classifications.

Rueters: Reuters is a dataset of news documents generated from the Reuters news website, including 46 categories [34]. Due to the serious imbalance of the document size of per category, we choose the five top categories in our study, a total of 8157 documents. The dataset is beforehand cut into training set with 6533 articles and testing set with 1624 articles. Since we download the data from the toolkit of Keras, and the data is represented in the form of arrays, we firstly converted it into the form of original text.

BBC: BBC dataset is a corpus of 2225 documents containing 5 different categories, which is crawled from the BBC news website [35]. Business, entertainment, politics, sport and tech are respectively the labels of each category. Since the training set and testing set have not been given, we split the corpus into training set with 1225 documents and testing document with 1000 documents randomly.

BBCsport: BBCsport is a collection of 737 documents from the BBCsport website including athletics, cricket, football, rugby and tennis in total of 5 categories [35]. Since the training set and testing set have not been given, we split the corpus into training set with 400 documents and testing set with 337 documents at random.

Table 2 shows the statistical information of the three datasets above. As for the preprocessing of the datasets, the headers and footers of every document have been removed. Then the punctuation and stop words with no practical sense in the document, have also been deleted. The stop words list is obtained from the library *sklearn* [34]. In addition, we convert all the letters into lower case.

4.2. Experiment Setup

We carry out the experiments to compare our proposed TRBoW1 model and TRBoW2 model with BoW model and TF-IDF model and AE method. In order to get the best result, we set the parameters of each classifier in a reasonable range, and the program automatically selects an optimal one. Since the most frequent based keyword extraction method (MF) was verified to be more efficient than other keywords extraction methods, such as term frequency inverse sentence frequency (TF-ISF) and co-occurrence statistical information (CSI) [36], we set the most 1000 frequent words in the corpus as the basis terms in all the experiments. Therefore, the representation vectors in this paper are set as 1000 dimensional.

After learning the document representation, the following four different popular machine learning classifiers are applied for classification tasks:

Support vector machine (SVM): In the SVM theory [37], a linear and a nonlinear kernel can be chose to classify linear data and nonlinear data, respectively. We use the linear one, which is called as linear SVM. We set the parameter C as $\{0.1, 1, 10\}$ and the parameter gamma as $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ additionally. Outcomes of the SVM classifier are chose the optimal one from these different parameters by the system.

K-nearest neighbor (KNN): In the KNN algorithm [38], one certain document belongs to the category that the most of the nearest k documents belong to. We search the best parameter K from 2 to 12.

Random forest (RF): RF algorithms [39] are a set of classification and regression trees derived from the guided samples of training data. The connection among trees and the efficiency of a single tree influence generation error of the classifiers. The depth of RF is searched from 6 to 12.

Ridge regression (RR): RR algorithm [40] is a nonlinear partial estimation method. And it is a biased estimator regression method specially used in the analysis of collinear data.

The other hyper-parameters are set as the default values of the system. To make a comprehensive comparison of the performance of our proposed TRBoW1 and TRBoW2 model for document representation, we use the five-fold cross validation in all the experiments to obtain the best result of every single experiment.

Table 2 Statistical information of the three datasets.

Statistics	Reuters	BBC	BBCsport
Document number	8157	2225	737
Class number	5	5	5
Training/Testing splits	6533/1624	1225/1000	400/337

4.3. Evaluation Metrics

In this section, we exploit precision rate, recall rate and F-measure to measure the performance of text classification.

Precision rate refers to the proportion of the number of texts to the total number of texts correctly classified by the classifier, which is defined as

$$pr = \frac{a}{a+b} \times 100\%, \tag{20}$$

where a denotes the number of texts that correctly classified into a category, and b means the number of texts that incorrectly classified into a category.

Recall rate refers to the proportion of the number of correct texts classified to all documents of the category, which is defined as

$$re = \frac{a}{a+c} \times 100\%, \tag{21}$$

where c represents the quantity of texts that incorrectly excluded from a category.

Precision rate and recall rate are contradictory measures. In general, the improvement of precision rate will lead to the depression of recall rate, and vice versa. So F_measure, harmonic mean of precision and recall, is proposed to make the comprehensive assessment, given as

$$F\_measure = \frac{2 \times pr \times re}{pr + re}. \tag{22}$$

4.4. Experimental Results

The experimental results are illustrated in this subsection. The experimental results of precision, recall rate and F_measure of document categorization on the BBCsport dataset obtained by the original BoW model, TF-IDF model, AE method and the proposed TRBoW1 and TRBoW2 are respectively described in Tables 3–5. Table 6 describes the classification precision of the four methods and four classifiers on the BBC dataset. Table 7 shows the recall and Table 8 shows the F_measure. Tables 9–11 present the study results on the dataset of Reuters. They are individually the precision, recall and F_measure.

Table 3 Precision (%) of four methods on the BBCsport dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	78.64	86.65	92.56	**92.58**	**92.58**
RR	96.44	96.71	97.02	97.33	**97.63**
RF	94.96	93.18	89.91	**95.85**	95.25
SVM	97.63	97.33	96.44	**98.22**	97.63

Table 4 Recall (%) of four methods on the BBCsport dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	80.16	86.84	93.65	93.59	**94.04**
RR	96.70	96.66	97.20	97.70	**98.07**
RF	95.28	95.18	90.33	**96.34**	95.60
SVM	97.79	96.37	97.51	**98.37**	98.18

5. DISCUSSIONS

In this section, we discuss the precision rate, recall rate and F-measure of the BoW model, TF-IDF model, AE method and our proposed TRBoW1 model and TRBoW2 for document representations on document categorization tasks by utilizing the KNN, RF, SVM and RR classifier on the BBCsport, BBC and Reuters corpus.

Table 5 F_measure (%) of four methods on the BBCsport dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	79.37	87.03	93.11	93.08	**93.31**
RR	96.57	96.68	97.11	97.52	**97.85**
RF	95.56	95.56	90.12	**96.09**	95.72
SVM	97.71	96.85	96.97	**98.30**	97.90

Table 6 Precision (%) of four methods on the BBC dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	69.30	69.60	**94.70**	91.80	94.10
RR	94.00	88.60	95.60	95.20	**95.80**
RF	91.40	92.10	94.20	93.80	**94.40**
SVM	94.20	94.80	**96.40**	96.10	95.30

Table 7 Recall (%) of four methods on the BBC dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	66.66	66.82	92.64	90.98	**93.66**
RR	93.65	87.85	95.49	94.90	**95.54**
RF	91.43	91.47	93.87	**94.08**	93.96
SVM	93.82	94.40	**96.22**	95.85	95.34

Table 8 F_measure(%) of four methods on the BBC dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	67.96	68.18	93.66	91.39	**93.88**
RR	93.82	88.22	95.55	95.05	**95.67**
RF	91.76	91.78	94.03	93.54	**94.18**
SVM	94.01	94.60	**96.31**	95.47	95.16

Table 9 Precision (%) of four methods on the Reuters dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	74.97	68.84	76.97	**80.36**	79.43
RR	86.51	86.58	86.45	86.02	**87.99**
RF	79.06	79.56	74.88	**87.07**	86.15
SVM	88.18	**90.21**	86.45	89.10	88.55

Table 10 Recall (%) of four methods on the Reuters dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	63.97	54.77	67.28	**70.38**	67.36
RR	78.15	80.06	76.07	77.78	**81.92**
RF	63.76	63.85	48.39	**79.24**	78.24
SVM	81.88	84.48	78.67	**84.66**	83.62

Table 11 F_measure (%) of four methods on the Reuters dataset.

	BoW	TF-IDF	AE	TRBoW1	TRBoW2
KNN	69.02	61.01	71.80	**75.04**	72.90
RR	82.12	83.19	80.93	81.69	**84.85**
RF	70.59	70.84	58.78	**82.97**	82
SVM	84.91	**87.25**	82.38	86.82	86.01

5.1. Performance on the BBCsport Dataset

As can be observed in all the tables of BBCsport dataset, in the comparison of these methods, the optimal results are labeled with bold fonts. And for the TRBoW1 and TRBoW2 method, the tables present the optimal results for the parameter θ from 4 to 20. For the reason that TF-IDF model takes the importance degree of each word in the documents into consideration, it has better performance compared to the BoW model, as shown in the tables. According to these tables, TRBoW1 model and TRBoW2 model both have better performance than BoW, TF-IDF and AE in all the four classifiers. Remarkably, the precision rate has increased by 13.94%, the recall rate has increased by 13.88%, and the precision rate has increased by 13.94% by using KNN. As a whole, the TRBoW1 model results in a greater performance when using RF and SVM as the classifier, and the TRBoW2 model results in a greater performance when using KNN and RR as the classifier.

5.2. Performance on the BBC Dataset

Data emerged in Tables 6–8 are searched the best results, when parameter θ ranges from 8 to 18. Compared with BoW, TF-IDF and AE, our methods achieves 96.10% in precision, 95.85% in recall and 95.67% in F_measure, especially the precision rate of TRBoW2 has increased by 24.91% than the BoW model, having significant improvements. It can be seen that our methods have captured more latent semantics than the BoW method.

5.3. Performance on the Reuters Dataset

Data in Tables 9–11 display that the TRBoW1 model outperforms the other models by the KNN, RF and SVM classifiers, and the TRBoW2 model outperforms the other models by the RR classifier. Because the theory of each classifier is different, it is acceptable that the experimental results of different classifiers are a little various. On account that we do not employ all the data of Reuters, the advantage of semantic mining may be not fully displayed. So compared with other datasets, performances of Reuters dataset do not have a more remarkable increment.

5.4. Co-occurrence Degree Threshold Value

In the tolerance rough set model, the co-occurrence threshold value θ determines the tolerance space. Hence the selection of the parameter θ has a tremendous impact on the performance of document representation. When the value of θ is too high, the upper approximation space may be too small, which will lead to the insufficient exploiting of potential relations behind documents. In the contrary, when the value of θ is too low, the upper approximation space may be too large, which will increase redundancy and noisy information among documents. Generally speaking, the selection of the value of θ relies on the corpus size and the vocabulary size. The larger the sizes, the higher the value is.

Figures 1 and 2 display the impacts of the value of θ on the precision and F_measure, which are the experimental results of our TRBoW1 and TRBoW2 model using the classifiers aforementioned on the BBCsport corpus in our study. The x label value of θ ranges from 4 to 16. Similarly, the influence on the BBC corpus are illustrated in Figures 3 and 4. The x label value of θ ranges from 8 to 18. As indicated in these figures, it is apparent that in majority cases, the trend of the curves generally satisfies the regular inferred above. On account that results predicted by RF have some randomness, it is explainable when some points are not satisfactory.

6. CONCLUSION

In this paper, we have proposed two novel document representation learning models, TRBoW1 model and TRBoW2 model, which adopt the tolerance rough set model to improve the traditional BoW model. The proposed TRBoW1 model and TRBoW2 model extend each document to its upper approximation. The extended upper approximation can mine the latent semantic relations behind documents, which solves the problems of lacking of latent semantics and

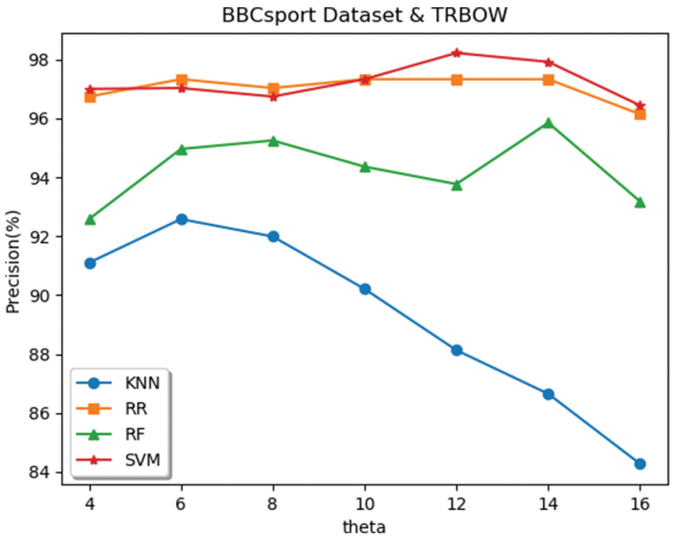

Figure 1 Precision using different classifiers with different threshold value θ.

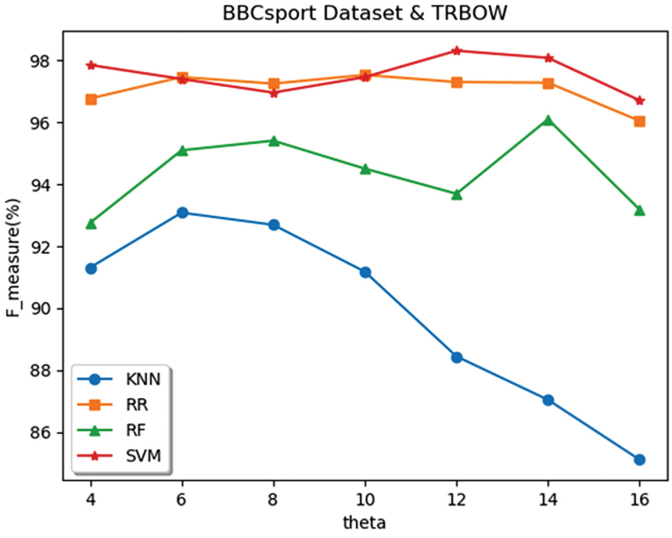

Figure 2 F_measure using different classifiers with different threshold value θ.

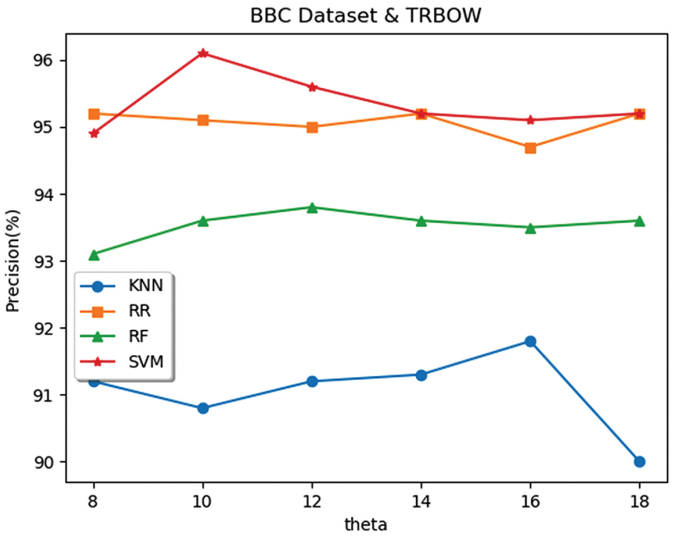

Figure 3 Precision using different classifiers with different threshold value θ.

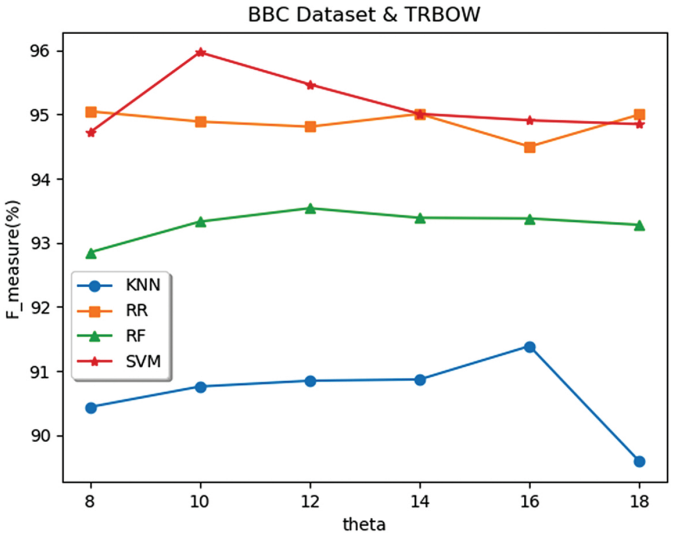

Figure 4 F_measure using different classifiers with different threshold value θ.

the sparsity of the original BoW model. They can learn the document representation without any training or prior knowledge. The experiments have carried out on various document representation methods for text classification on different datasets using classifiers including KNN, RR, RF and SVM. The results of the experiments indicate that the performances have been improved remarkably and allow us to obtain the following conclusions: the highest F_measure of BBCsport is up to 98.30%; the proposed representation methods enrich the representation of BoW, making the improvement in performance up to 27%.

Except for text categorization tasks, the TRBoW1 and TRBoW2 model can be applied in other domains such as information retrieval and document clustering, which will be further studied in the future. Besides, we apply it in the sentence similarity calculation on the basis of this work.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

These authors contributed equally to this work.

ACKNOWLEDGMENTS

The authors would like to thank the referee for his valuable remarks. This work was supported by The National Natural Science Foundation of China (Grant no. 11671001).

REFERENCES

[1] G. Salton, Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing, Boston, MA, USA, 1989.
[2] M.M. Mironczuk, J. Protasiewicz, A recent overview of the state-of-the-art elements of text classification, Expert Syst. Appl. 106 (2018), 36–54.
[3] S. Vashishtha, S. Susan, Fuzzy rule based unsupervised sentiment analysis from social media posts, Expert Syst. Appl. 138 (2019), 1–15.
[4] A. Skowron, J. Stepaniuk, Tolerance approximation spaces, Fund. Inform. 27 (1996), 245–253.
[5] C. Zhong, Y. Chen, J. Peng, Feature selection based on a novel improved tree growth algorithm, Int. J. Comput. Intell. Syst. 13 (2020), 247–258.
[6] Y. Matsuo, M. Ishizuka, Keyword extraction from a single document using word co-occurrence statistical information, Int. J. Artif. Intell. Tools. 13 (2004), 157–169.
[7] R.G. Rossi, R.M. Marcacini, S.O. Rezende, Analysis of domain independent statistical keyword extraction methods for incremental clustering, Int. J. Comput. Int. Syst. 12 (2014), 17–37.
[8] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in Proceeding ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, (1997), pp. 412–420.
[9] R. Basili, A. Moschitti, M.T. Pazienza, A text classifier based on linguistic processing, In: Proceedings of IJCAI99, Machine Learning for Information Filtering, Citeseer, Stockholm, Sweden, (1999), pp. 1–6.
[10] C. Qiyue, Structure entropy weight method to confirm the weight of evaluating index, Syst. Eng. Theory Pract. 30 (2010), 1225–1228.
[11] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (1988), 513–523.
[12] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972), 11–21.
[13] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003), 993–1022.
[14] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990), 391–407.

[15] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (2001), 177–196.
[16] T. Hofmann, Probabilistic latent semantic indexing, ACM SIGIR Forum. 51 (2017), 211–218.
[17] M. You, J. Liu, G.Z. Li, Y. Chen, Embedded feature selection for multi-label classification of music emotions, Int. J. Comput. Intell. Syst. 5 (2012), 668–678.
[18] F.J. Pulgar, F. Charte, A.J. Rivera, M.J. Del Jesus, AEkNN: an autoencoder kNNCbased classifier with built-in dimensionality reduction, Int. J. Comput. Intell. Syst. 12 (2018), 436–452.
[19] E.R. Henry, J. Hofrichter, Singular value decomposition: application to analysis of experimental data, Methods Enzymol. 210 (1992), 129–192.
[20] R. Das, M. Zaheer, C. Dyer, Gaussian lda for topic models with word embeddings, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 2015, pp. 795–804.
[21] D.Q. Nguyen, R. Billingsley, L. Du, M. Johnson, Improving topic models with latent feature word representations, Trans. Assoc. Comput. Linguistics. 3 (2015), 299–313.
[22] Y. Liu, Z. Liu, T.S. Chua, M. Sun, Topical word embeddings, in Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA, AAAI Press, pp. 2418–2424.
[23] R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in Advances in Neural Information Processing Systems, Montreal, Canada 2015, pp. 3294–3302.
[24] Y. Wu, S. Zhao, W. Li, Phrase2Vec: phrase embedding based on parsing, Inf. Sci. 517 (2020), 100–127.
[25] D. Yao, J. Bi, J. Huang, J. Zhu, A word distributed representation based framework for large-scale short text classification, in 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 2015.
[26] Y. Gao, Y. Xu, H. Huang, Q. Liu, L. Wei, L. Liu, Jointly learning topics in sentence embedding for document summarization, IEEE Trans. Knowl. Data Eng. 32 (2020), 688–699.
[27] M.A. Mouriño-García, R. Pérez-Rodríguez, L. Anido-Rifón, M. Vilares-Ferro, Wikipedia-based hybrid document representation for textual news classification, Soft Comput. 22 (2018), 6047–6065.
[28] R. Zhao, K. Mao, Fuzzy bag-of-words model for document representation, IEEE Trans. Fuzzy Syst. 26 (2018), 794–804.
[29] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982), 341–356.
[30] Y. Yao, Probabilistic rough set approximations, Int. J. Approx. Reason. 49 (2008), 255–271.
[31] Z. Pawlak, R. Sowinski, Rough set approach to multi-attribute decision analysis, Eur. J. Oper. Res. 72 (1994), 443–459.
[32] T.B. Ho, N.B. Nguyen, Nonhierarchical document clustering based on a tolerance rough set model, Int. J. Intell. Syst. 17 (2002), 199–212.
[33] http://code.google.com/archive/p/word2vec/
[34] http://scikit-learn.org/stable
[35] D. Greene, P. Cunningham, Practical solutions to the problem of diagonal dominance in kernel document clustering, in Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, ACM,2006, pp. 377–384.
[36] A. Onan, S. Korukoğlu, H. Bulut, Ensemble of keyword extraction methods and classifiers in text classification, Expert Syst. Appl. 57 (2016), 232–247.
[37] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011), 27.
[38] M.L. Zhang, Z.H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (2007), 2038–2048.
[39] A.S. More, D.P. Rana, I. Agarwal, S. Vallabhbhai, Random forest classifier approach for imbalanced big data classification for smart city application domains, Int. J. Comput. Intell. Syst. 1 (2020), 260–266.
[40] Y. Yang, J. Zhang, B. Kisiel, A scalability analysis of classifiers in text categorization, in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, Toronto, Canada, 2003, pp. 96–103.