International Journal of Database Theory and Application
Vol.9, No.1 (2016), pp.185-200
http://dx.doi.org/10.14257/ijdta.2016.9.1.16

WordNet-based Hybrid VSM for Document Classification

Luda Wang[1,2], Peng Zhang[1*] and Shouping Gao[1]

[1]*Xiangnan University, Chenzhou, China*
[2]*School of Information Science and Engineering, Central South University, Changsha, China*
wang_luda@163.com, mimazp@126.com, gaoshoup@263.net

Abstract

Many text classifications depend on statistical term measures or synsets to implement document representation. Such document representations ignore the lexical semantic contents or relations of terms, leading to losing the distilled mutual information. This work proposed a synthetic document representation method, WordNet-based hybrid VSM, to solve the problem. This method constructed a data structure of semantic-element information to characterize lexical semantic contents, and support disambiguation of word stems. As a template, lexical semantic vector consisting of lexical semantic contents was built in the lexical semantic space of corpus, and lexical semantic relations are marked on the vector. Then, it connects with special term vector to form the eigenvector in hybrid VSM. Applying algorithm NWKNN, on text corpus Reuter-21578 and its adjusted version, the experiments show that the eigenvector performs F1 measure better than document representations based on TF-IDF.

***Keywords**: Document Representation, Lexical Semantic, Classification, VSM*

1. Introduction

Text corpus analysis is an important task. Meanwhile, clustering and classification are key procedures for text corpus analysis. In addition, text classification is an active research area in information retrieval, machine learning and natural language processing. Most classification algorithms based on eigenvector prevail in this field, such as KNN, SVM, ELM, etc. Eigenvector-based document classification is a widely used technology for text corpus analysis. Moreover, the key issue is eigenvector-based classification algorithms depend on the VSM [1].

TF-IDF (term frequency–inverse document frequency) [2] is a prevalent method for characterizing document, and its essence is statistical term measure. Many methods of document representation based on TF-IDF can construct Vector Space Model (VSM) of text corpus. Similarly, many methods of document representation exploit statistical term measures, such as BoS (Bag-of-Words) [3] and Minwise hashing [4]. For document representation, these methods are perceived as statistical methods of feature extraction.

However, in information retrieval field, statistical term measures neglect lexical semantic content. It causes corpus analysis to perform on the level of term string basically, and connives lexical replacement of document original at deceiving the text corpus analysis easily.

Close relationship between syntax and lexical semantic contents of words have attracted considerable interest in both linguistics and computational linguistics. Semantic approach based on VSM [1] is an effectively used technology for document analysis, such as synset vector or BoS model [3]. It can capture the semantic features of word senses,

[*] Corresponding Author

ISSN: 2005-4270 IJDTA

and based on that, characterizes and classifies the document. But the approach, synset vector, fails to characterize lexical semantic relations.

The design and implementation of WordNet-based hybride VSM take account of special terms, lexical semantic contents and relations collectively. Unlike traditional VSM methods of feature extraction, our work developed a new term measure which can characterize both lexical semantic contents and relations, and provides a practical method of document representation which can handle the impact of lexical replacement. The document representation is normalized as eigenvector, consequently, it can apply to current VSM-dependent classification algorithms. Theoretical analysis and a large number of experiments are carried out to verify the effectiveness of this method.

2. Related Work

In information retrieval field, similarity and correlation analysis of text corpus needs to implement corresponding document representations for diverse algorithms. Many practicable methods of document representation share a basic mechanism, statistical term measure. Typical statistical methods of feature extraction include TF-IDF based on lexical term frequency and shingle hash based on consecutive terms [5]. TF-IDF methods of feature extractions employ a simple assumption that frequent terms are also significant [2]. And, these methods quantify the extent of usefulness of terms in characterizing the document in which they appear [2]. Besides, as for some hashing measures based on fingerprinted shingle, people call a sequence of k consecutive terms in a document a shingle. Then a selection algorithm determines which shingles to store in a hash table. And various estimation techniques are used to determine which shingles are copied and from which document most of the content originated [5]. These methods for document representation are perceived as the mode using statistical term measures. As a sort of ontology methods [6], these document representations ignore recognition of lexical semantic contents. It causes the document representation to lose the mutual information [7] of term meanings which comes from synonyms in different document samples. Moreover, lexical replacement of document original cannot be distinguished literally by radical statistical mechanisms of term measure.

Additionally, as a semantic approach, synset vector [3] can roughly characterize the lexical semantic contents of terms. In WordNet, a sense of word is represented by the set of (one or more) synonyms that have that sense [8]. Synonymy is a symmetric relation between words [8]. It is WordNet’s basic relation, because WordNet uses sets of synonyms (synsets) to represent word senses, namely lexical semantic contents. Using only Synonymy causes the document representation to lose the mutual information [7] of term meanings which comes from semantic relations such as Hyponymy *etc*.

Example 1.

Sample A. *Men prize maple.*

Sample B. *Human treasures trees.*

Above comments on statistical term measures and synset vector can be clarified by analyzing a text corpus, *Example 1*. In *Example 1*, the two simple sentences are viewed as two document samples, and comprise the small corpus. Evidently, the meanings of *Sample A* and *Sample B* are extremely similar. Thus, the correlation and semantic similarity between these two documents are considerable. Meanwhile, *Sample B* can be regarded as a derivative from *Sample A* via lexical replacement.

Table 1. The Statistical Term Measures on Example 1

Term	Men	prize	maple	Human	treasures	trees
Term Frequency Vector *A*	0	0	0	1	1	1
Term Frequency Vector *B*	1	1	1	0	0	0

Obviously, on behalf of statistical term measures, the document representations on *Example 1* shown in Table 1 did not perform well. Comparing two vectors in Table 1, positive weights do not coexist in the same term of two samples. These two orthogonal vectors of term frequency demonstrate the statistical term measures for document representation cannot signify semantic similarity of the corpus *Example 1* effectively. Because they did not recognize and represent the lexical semantic contents of these two documents practically. As a result, these two vectors cannot provide mutual information of lexical semantic contents.

Using WordNet, two samples in Example 1can be represent in BOS (bag of synsets) model. Each document is represented as a synset vector rather than a term frequency vector. Synsets of every term are equivalent to lexical semantic contents, which are structures containing sets of words with synonymous meanings. However, the synset vector based on only Synonymy fails to characterize lexical semantic relations including Antonymy, Hyponymy, Meronymy, Troponomy and Entailment. Taking *Example 1*, the term *maple* and *trees* have no any common synset in WordNet, but *trees* is hypernym of *maple*. Consequently, in the aspect of lexical semantic relations, synset vector cannot provide mutual information. BoS model can not characterize semantic relations in documents effectively.

3. Proposed Program

3.1 The Motivation and Theoretical Analysis

For text corpus analysis, document representations which depend on statistical term measures shall lose mutual information of term meanings. Besides, using BoS model, synset vector can roughly characterize the lexical semantic contents but shall lose mutual information of lexical semantic relations.

To solve these problems, the motivation is connecting a lexical semantic vector with a special term vector to form the hybrid VSM. In hybrid VSM of corpus, the eigenvector consist of lexical semantic spectrum and special term spectrum (shown in Figure 1). On the basis of synset, this method ought to construct a synset VSM for building lexical semantic spectrum, and lexical semantic spectrum shall further characterize semantic relations. Additionally, special term spectrum shall characterize the unusual string in documents, which is neither the recognizable word in WordNet nor function word [9].

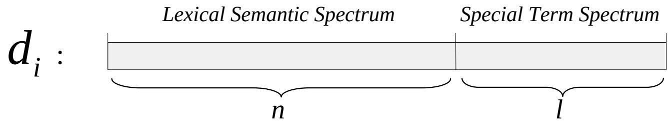

Figure 1. The Eigenvector in Hybrid VSM

In WordNet, because one word or term refers to particular synonym sets, several particular synonym sets can strictly describe the sense of one word for characterizing lexical semantic contents. Then, our method defines these particular synsets as the semantic-elements of word. Thus, lexical semantic spectrum resorts to WordNet [8], a lexical database for English, for extracting lexical semantic contents. Then, the method of document representation shall preliminarily construct a synset VSM of text corpus.

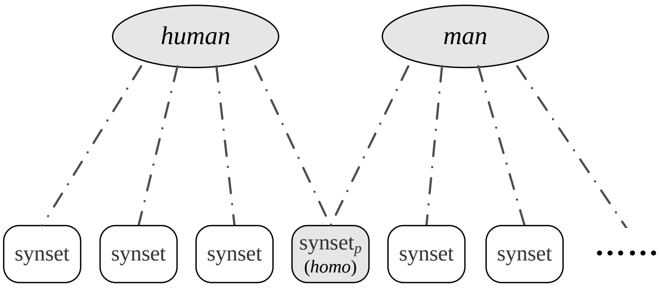

Figure 2. Common Semantic-Element of Words

Based on the above definition, involved semantic-elements can characterize the lexical semantic contents of *Example 1*, which can accomplish feature extraction of lexical semantic contents. For instance, in Figure 2, the words *human* and *man* belong to different documents in *Example 1*, and the common synset *homo* can represent mutual information [7] between lexical semantic contents. Then, lexical semantic spectrum shall capture the mutual information of lexical semantic contents between samples which lies in same semantic-elements of different documents.

Moreover, to characterize the semantic relations of terms, weights of lexical semantic relations shall be marked on the vector in synset VSM, using Antonymy, Hyponymy, Meronymy, Troponomy and Entailment between synsets [8]. Then, lexical semantic spectrum can capture mutual information from lexical semantic relations between documents. In addition, lexical semantic spectrum connects with special term vector to form the eigenvector in hybrid VSM. In the special term vector, the weight of each term is computed using TF-IDF.

According to the statistical theory of communications, our motivation needs further analysis for theoretical proof. The analysis first introduces some of the basic formulae of information theory [2, 7], which are used in our theoretical development of samples mutual information. Now, let x_i and y_j be two distinct terms (events) from finite samples (event spaces) *X* and *Y*. Then, let $\mathcal{X}$ or $\mathcal{Y}$ be random variables representing distinct lexical semantic contents in samples *X* or *Y*, which occur with certain probabilities. In reference to above definitions, mutual information between $\mathcal{X}$ and $\mathcal{Y}$, represents the reduction of uncertainty about either $\mathcal{X}$ or $\mathcal{Y}$ when the other is known. The mutual information between samples, $I\ (\mathcal{X};\mathcal{Y})$, is specially defined to be

$$I\ (\mathcal{X};\mathcal{Y}) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} . \tag{1}$$

In the statistical methods of feature extraction, probability $P(x_i)$ or $P(y_j)$ is estimated by counting the number of observations (frequency) of x_i or y_j in sample *X* or *Y*, and normalizing by *N*, the size of the corpus. Joint probability, $P(x_i, y_j)$, is estimated by counting the number of times (related frequency) that term x_i equals (is related to) y_j in the respective samples of themselves, and normalizing by *N*.

Taking the *Example 1*, between any term x_i in *Sample A* and any term y_j in *Sample B*, there is not any counting of times that x_i equals y_j. As a result, on corpus *Example 1*, the statistical term measures indicate $P(x_i, y_j) = 0$ and the samples mutual information $I(X;Y) = 0$. Thus, the analysis verifies that the statistical methods of feature extraction lose mutual information of term meanings.

As to semantic approach, for feature extraction of lexical semantic contents, our method uses several particular semantic-elements to describe the meaning of one word or term. In different samples, words can be related to other words by common semantic-elements or lexical semantic relations. Then, lexical semantic mutual information between samples, $I(X;Y)$, is re-defined to be

$$I(X;Y) = \sum_{x_i \in X} \sum_{y_j \in Y} F(\mathrm{e}_{x_i,y_j}) \bmod N \log \frac{F(\mathrm{e}_{x_i,y_j}) \bmod N}{F(\mathrm{e}_{x_i}) \bmod N \times F(\mathrm{e}_{y_j}) \bmod N} . \tag{2}$$

To denote probability $P(x_i)$ or $P(y_j)$, function $F(\mathrm{e}_{x_i})$ or $F(\mathrm{e}_{y_j})$ is estimated by calculating the frequency of semantic-elements that describe the meaning of x_i or y_j in sample *X* or *Y*, and modulo *N*, the total number of semantic-elements in corpus. Meanwhile, to denote joint probability $P(x_i, y_j)$, function $F(\mathrm{e}_{x_i,y_j})$ is estimated by calculating the frequency of common semantic-elements that relate to lexical semantic contents or relations of x_i and y_j, and modulo *N*.

For lexical semantic feature, in *Example 1*, the frequency of semantic-elements are calculated by marking the lexical semantic contents and relations, joint probability $P(x_i, y_j)$ is estimated by counting the frequency of the common semantic-elements, and modulo *N*. For instance, the words *human* and *man* are described by the common semantic-element *homo* (shown in Figure 2). In reality, $P(human, man)$ = $F(homo) \bmod N$ >0 and $P(tree, maple)$ = $F(\text{Hyponymy}(tree, maple)) \bmod N$>0. Note that, the $F(homo)$ denotes the frequency of the common semantic-element, synset *homo*, which is caused by the lexical semantic contents, and the $F(\text{Hyponymy}(tree, maple))$ denotes the frequency of the common semantic-elements, which are caused by one of the lexical semantic relations, Hyponymy. As a result, lexical semantic mutual information between *Sample A* and *Sample B*, $I(X;Y)$, is positive. Thus, the analysis proves that the semantic-elements and extraction of lexical semantic feature can provide the probability-weighted amount of information (PWI) [2] between term meanings of documents on the lexical semantic level.

3.3 Hybrid VSM of Text Corpus

In our method, documents are represented using the vector space model (VSM). Each document is considered to be an eigenvector in the hybrid feature space. For forming the hybrid VSM, the procedures are as follows. In the first place, **(1)** for feature extraction of lexical semantic contents, our work makes a data structure of semantic-element information. Secondly, **(2)** the data structure uses EM modeling to disambiguate word stems. Moreover, **(3)** using semantic-elements, it constructs synset VSM of corpus, and builds lexical-semantic-content vector as template vector of the lexical semantic spectrum in hybrid VSM. Last, **(4)** to characterize lexical semantic relations, it marks each template vector with weights of 5 semantic relations between synsets [8]. Thus, **(5)** eigenvector in hybrid VSM is constructed via connect lexical semantic spectrum to special term spectrum, in which the latter is TFIDF vector of special terms.

(1) The data structure of semantic-element information comprises relevant information of each semantic-element in a document sample. As a data element, The data structure is shown in Table 3. It can record all important information of semantic-elements in a document, such as synset ID, weight, document sample ID and relevant information of words.

Table 3. Data Structure of Semantic-Element Information

Item	Explanation
Synset ID	Identification of synonym set
Set of Synonym	**S**ynonymy is WordNet's basic relation. WordNet uses sets of synonyms (synsets) to represent word senses.[8]
Weight (Frequency)	Frequency of semantic-element in a document sample (sum of Semantic Members Frequency)
Sample ID	Identification of document sample
Semantic Member	A linked list (shown in Figure 3) which carries all Original Words of Terms referring to the semantic-element and their Word Stem(s)
Semantic Members Frequency	A linked list (shown in Figure 4) which carries frequency of each Original Words of Terms (that refer to the semantic-element) one by one

Note that, in a record of the data structure, each original word in inflected form [10] referring to the semantic-element and its word stem(s) in base form [10-11] are recorded by linked list of *Semantic Member* (shown in Figure 3(a)). According to WordNet framework [8], when original word refers to more than 1 word stems, the linked-list of *Semantic Member* will expend the very node of the original word to register all word stems.

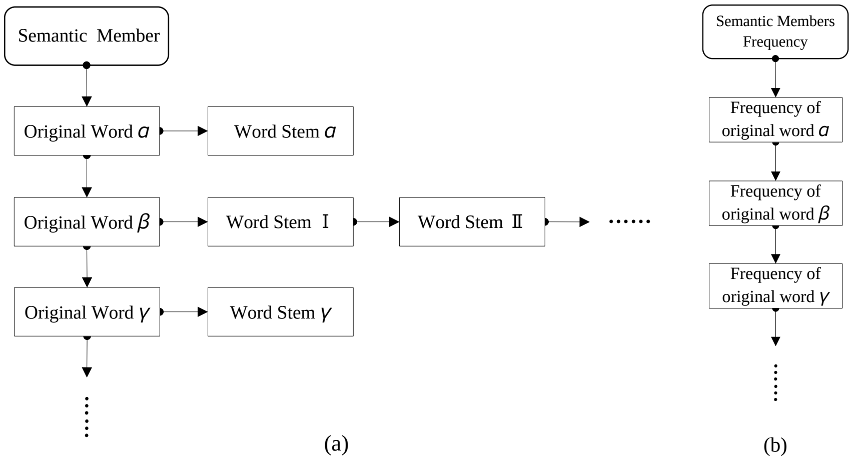

Figure 3. The Linked-lists of Semantic Member (a) and Semantic Members Frequency (b)

Meanwhile, the linked-list of *Semantic Members Frequency* is shown in Figure 3(b). It records the frequency of each original word one by one in node order of *Semantic Member*.

(2) On the basis of above data structure, *Semantic Member* needs to disambiguate word stems of original word. In case of an original word referring to more than 1 word stem in base form, semantic-element must ensure that one original word refers to only 1 word stem. Then, in order to select only 1 word stem for an original word (shown in Figure 4), we employ the Maximum Entropy Model [12].

ME modeling provides a framework for integrating information for classification from many heterogeneous information sources [13]. In our model, we suppose that diversity [14] of *Semantic Member* implies the significance of the semantic-element and the rationality of existing *Semantic Members*.

Assume a set of original words *X* and a set of its word stems *C*. The function $cl(x): X \rightarrow C$ chooses the word stem *c* with the highest conditional probability, which makes sure original word *x* only refers to: $cl(x) = \arg\max_c p(c \mid x)$. Each feature [13] of original word is calculated by a function that is associated to a specific word stem *c*, and it takes the form of equation (3), where S_i is the number of *Semantic Member* of semantic-element *i*, P_j is the proportion of the *Frequency of original word j* to *Weight* in semantic-element *i*, and the $-\sum_{j=1}^{S_i} P_j \cdot \log_2 P_j$ indicates *Semantic Member* diversity of semantic-element *i* in a document, in the form of Shannon-Wiener index [14].

The conditional probability $p(c \mid x)$ is defined by equation (4). The parameter of the semantic-element *i* [13], α_i , is the *Frequency* of original word *x* in semantic-element *i*. *K* is the number of semantic-elements that word stem *c* refers to, and $Z(x)$ is a value to ensure that the sum of all conditional probabilities for this context is equal to 1.

$$f_i(x,c) = \begin{cases} -\sum_{j=1}^{S_i} P_j \cdot \log_2 P_j & \text{if } x \text{ refers to } c \text{ and } c \text{ refers to semantic-element } i, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

$$p(c \mid x) = \frac{1}{Z(x)} \prod_{i=1}^{K} \alpha_i^{f_i(x,c)} . \quad (4)$$

Above equations aim at finding the highest conditional probability $p(c \mid x)$, and using the function $cl(x)$ to ensure that original word *x* refers to only 1 word stem (like Figure 4). After semantic-elements characterizing lexical semantic contents of a document preliminarily, the specified ME modeling is applied to implement disambiguation of word stems. Necessarily, the relevant items in the data structure of semantic-element information shall be modified, such as the *Semantic Member*, the *Frequency of original word*, and the *Weight*. Furthermore, some relevant semantic-elements shall be eliminated.

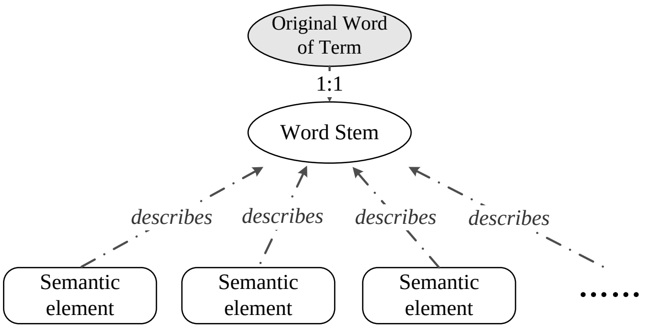

Figure 4. 1:1 Reference of Original Word

(3) As the template of lexical semantic spectrum in hybrid VSM, synset VSM of corpus is constructed using semantic-elements. In text corpus, all referred semantic-elements are fixed by disambiguation of word stems, then, each identical *Synset ID* of all semantic-elements fills one dimension in synset VSM respectively. And, each template vector of lexical semantic spectrum is built in the synset VSM of text corpus and characterizes lexical semantic contents. In synset VSM, each template vector of document is just the lexical-semantic-content vector. Specifically, each document sample identified by *Sample ID* is represented by the template vector of lexical semantic spectrum, for only characterizing lexical semantic contents. The synset VSM represents a document doc_x, using a lexical-semantic-content vector d_x Î R^n, given as

$$d_x = \left(d_{x(1)}, d_{x(2)}, \cdots\cdots, d_{x(n)}\right), \tag{5}$$

where n is the number of identical *Synset ID* of all semantic-elements in corpus, $d_{x(i)}$ is the feature value on the i^{th} synset, given as $d_{x(i)} = FS(s_i, doc_x)$ for all $i=1$ to n. $FS(s_i, doc_x)$ is the *Weight* (frequency) of the i^{th} corresponding semantic-element s_i in document doc_x.

(4) On the basis of synset VSM, to characterize lexical semantic relations, it marks Antonymy, *Hyponymy*, *Meronymy*, *Troponomy* and *Entailment* on each dimension of the template vector. The processing is formulized as

$$\mathrm{D}d_{x(j)} = \overset{n}{\underset{i=1}{\mathring{a}}} R(i,j) \times d_{x(i)}, \tag{6}$$

$$R(i,j) = \begin{cases} 0.5 & \textit{Antonymy} \\ 0.2 & \textit{Hyponymy, Meronymy,} \\ & \textit{Troponomy} \text{ or } \textit{Entailment}, \\ 0 & \text{Unrelated} \end{cases} \tag{7}$$

where i and $j=1$ to n, n is dimensional number of the template vector, and $d_{x(i)}$ is value of the i^{th} template vector element. The $\mathrm{D}d_{x(j)}$ is semantic relation increment to the j^{th} dimensional value of template vector, and function $R(i, j)$ denotes semantic relation coefficient for $\mathrm{D}d_{x(j)}$. Specifically, when the synset of j^{th} dimension is related to synset of j^{th} dimension via semantic relation such as Antonymy, Hyponymy, Meronymy, Troponomy or Entailment [8], the $R(i, j)$ assignment is shown in equation (7).

(5) Consequently, eigenvector in hybrid VSM is constructed via connect lexical semantic spectrum to special term spectrum, in which the latter is TFIDF vector of special terms.The hybrid eigenvector, $\underset{\cdot}{d}_x$ Î R^{n+l}, is given as

$$\underset{\cdot}{d}_{x(j)} = \begin{cases} d_{x(j)} + \Delta d_{x(j)} & j = 1 \text{ to } n \\ TF(w_{j-n}, doc_x) \times IDF(w_{j-n}) & j = n+1 \text{ to } n+l \end{cases}, \tag{8}$$

where $\underset{\cdot}{d}_{x(j)}$ is the feature value on the j^{th} special term, given as $\underset{\cdot}{d}_{x(j)}=TF(w_j, doc_x)\times IDF(w_j)$ for all $j=n+1$ to l, and $TF(w_{j-n}, doc_x)$ is the frequency of the term w_{j-n} in document doc_x, $IDF(w_{j-n})$ is the inverse document frequency of w_{j-n}.

3.2 Algorithm NWKNN

In the text corpus analysis, KNN classification is a classical algorithm, which is effective on selection of data eigenvectors especially. To tackle unbalanced text corpus, we select an optimized KNN classification, the NWKNN (Neighbor-Weighted K-Nearest Neighbor) algorithm defined to be equation (9) [15]. As for this algorithm, each document *d* is considered to be an eigenvector in the lexical-semantic space or the term-space [15, 17]. And, in the term-space, the weight of each word [2] is computed using TF-IDF.

$$score(d, c_i) = \text{Weight}_i \left(\sum_{d_j \in KNN(d)} Sim(d, d_j)\delta(d_j, c_i) \right). \tag{9}$$

In the process of equation (9), this algorithm uses cosine value between eigenvectors of document *d* and d_i [15] to calculate the $\text{Sim}(d, d_i)$. Besides, according to experience of NWKNN algorithm [15], the parameter of Weight_i, Exponent [15], ranges from 2.0 to 6.0.

4. Experiment and Result

4.1 Experiment Setup

In our work, experiments use 2 sorts of eigenvector to represent document sample 1) the eigenvector based on lexical semantics and TF-IDF in hybrid VSM, given as equation (8). 2) TF-IDF eigenvector in the term-space which takes different numbers of selected features using Information Gain [16]. The former stands for the feature extraction of lexical semantics and special term, and the latter stands for the typical statistical method of feature extraction, TF-IDF.

Our experiments use two corpora: Reuters-21578 and an adjusted corpus based on Reuter-21578.

Reuter. The Reuters-21578 Text Categorization Test Collection contains documents collected from the Reuters newswire in 1987, and was last modified in 16 Feb 1999. It is a standard text categorization benchmark and contains 135 categories. Our experiments used its subset: one consisting of 20 categories, which has approximately 3500 documents (Table 4).

Adjusted corpus (based on Reuter-21578). After selecting the subset of Reuter-21578, we unite lexical-replacement documents deriving from 10% of the subset originals with it. Specifically, each lexical-replacement document is changed from an original document in the subset. For instance, in Table 5, the semantic contents of the *Lexical replacement* and *Original* are the same, and the meanings of them are extremely equivalent.

Table 4. The Distribution of All Categories in the Subset of Reuter-21578

Category	Sample	Category	Sample	Category	Sample
Cotton	27	Grain	489	Nat-gas	48

Earn	761	Heat	16	Jobs	50
Cpi	75	Money-supply	113	Cocoa	59
Rubber	40	Silver	16	Trade	441
Sugar	145	Tin	32	Housing	16
Money-fx	574	Crude	483	Nickel	5
Bop	47	Hog	16		

Table 5. Lexical Replacement of <REUTERS …… NEWID="40">

Original	Lexical replacement
"Stable interest rates and a growing economy are expected to provide favorable conditions for further growth in 1987," president Brian O'Malley told shareholders at the annual meeting. Standard Trustco previously reported assets of 1.28 billion dlrs in 1986, up from 1.10 billion dlrs in 1985. Return on common shareholders' equity was 18.6 pct last year, up from 15 pct in 1985.	"Unchanging accrual rates of deposit and an uprising economy are anticipated to render favourable status for further increment in 1987," president Brian O'Malley said to stockholders at the yearly meeting. Standard Trustco antecedently covered assets of 1.28 billion dlrs in 1986, upward from 1.10 billion dlrs in 1985. Return on common stockholders' equity was 18.6 percent last year, upward from 15 percent in 1985.

To evaluate the text classification system, we use the $F1$ measure [17]. This measure combines recall and precision in the following way:

$$F1 = \frac{2' \ \text{Recall}' \ \text{Precision}}{(\text{Recall}+ \text{Precision})}. \tag{10}$$

Using $F1$ measure, we can observe the effect of different kinds of data on a text classification system [17]. For ease of comparison, we summarize the $F1$ scores over the different categories using the macro-averages of $F1$ scores, in the same way, we can obtain the Macro-Recall and Macro-Precision [17].

4.2 The Results

Our experiments split the each dataset into three parts. Then we use two parts for training and the remaining third for test. To accomplish three-fold cross validation, we conduct the training-test procedure three times alternately, and use the average of the three performances as final result.

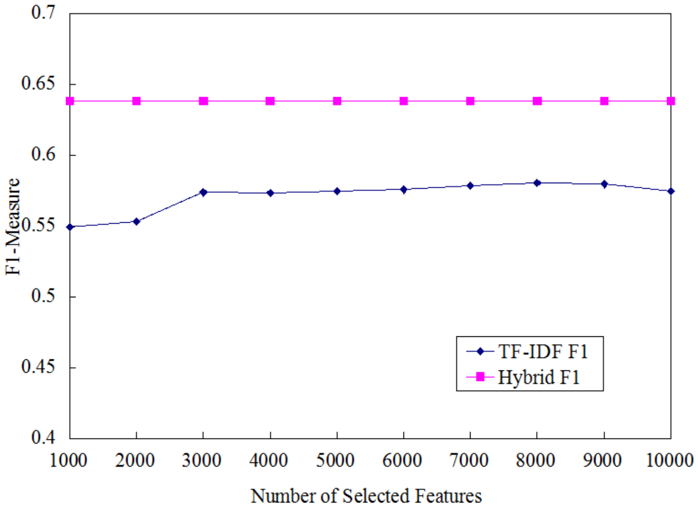

Figure 5. Classification Result of the Hybrid and the TF-IDF Eigenvector with Different Term-Space Feature Numbers on Reuter

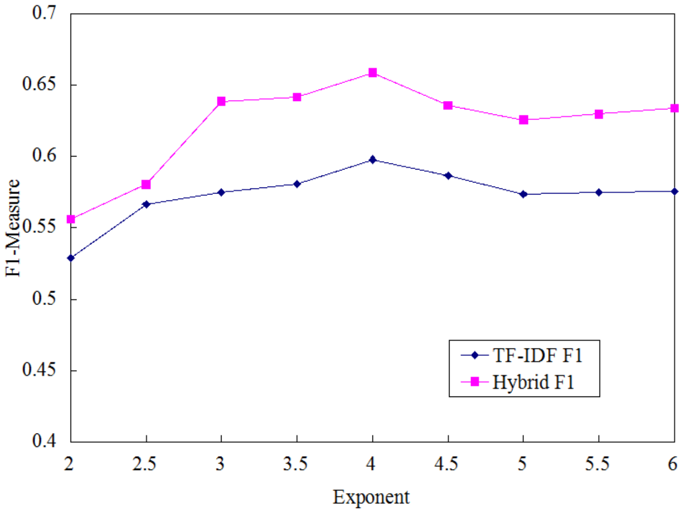

Figure 6. Classification Result of the Hybrid and the TF-IDF Eigenvector with the Exponent on Reuter

Figure 5 manifests the classification performance curves for hybrid eigenvector and TF-IDF eigenvector on *Reuter*, after using the NWKNN algorithm. Note that Exponent takes 3 empirically [15]. From the figure, it is obvious that our hybrid eigenvector beats TF-IDF eigenvectors under all selected feature numbers of term-space [16] by about 5%~7.5% on *Reuter*.

Figure 6 illustrates the classification performance comparison between hybrid eigenvector and TF-IDF eigenvector using different Exponent on *Reuter*, respectively. Note that the feature number of term-space takes 10,000. In the figure, with the increase of Exponent, our hybrid eigenvector performs better on *Reuter*, and beats TF-IDF eigenvector by 5% averagely.

Figure 7 describes the Macro-Precision and Macro-Recall comparison between hybrid eigenvector and TF-IDF eigenvector using different Exponent on *Reuter*, respectively. Note that the feature number takes 10,000. From the two figures, it is an apparent phenomenon that with the increase of Exponent, the curves accord with the experience of

NWKNN [15]. Meantime, Macro-Precision or Macro-Recall of hybrid eigenvector is superior to the TF-IDF eigenvector by 7.5% or 9% on *Reuter* averagely.

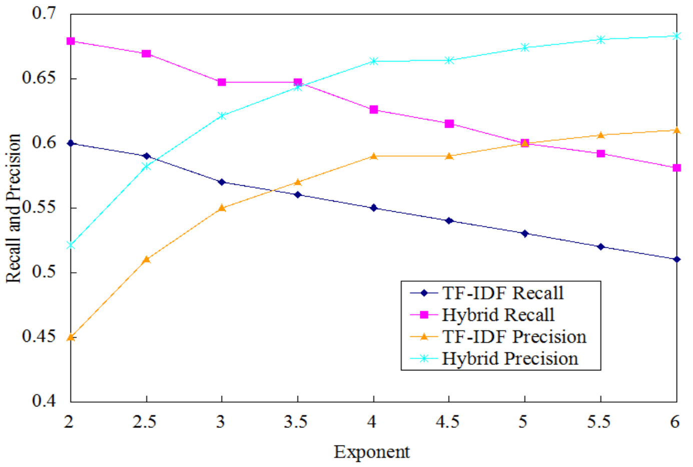

Figure 7. Classification recall and Precision of Hybrid Eigenvector with the Exponent and TF-IDF Eigenvector on Reuter

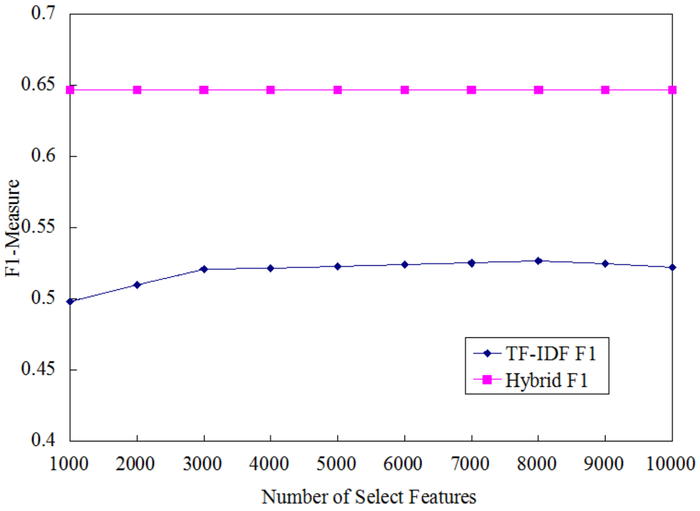

Figure 8. Classification Result of the Hybrid and the TF-IDF with Different Term-Space Feature Numbers on Adjusted Corpus

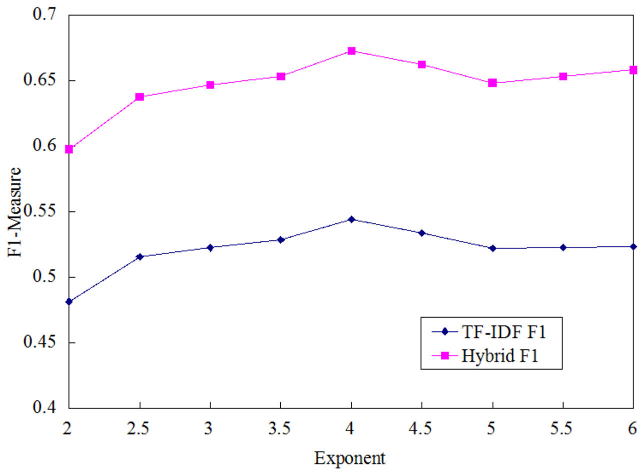

Figure 9. Classification Result of the Hybrid and the TF-IDF with the Exponent on Adjusted Corpus

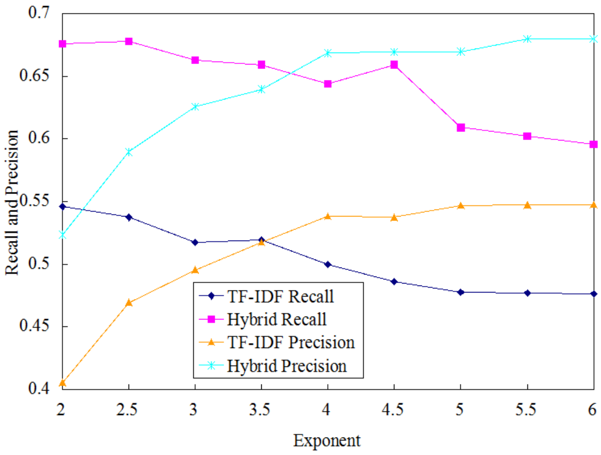

Figure 10. Classification Recall and Precision of Hybrid Eigenvector with the Exponent and TF-IDF Eigenvector on Adjusted Corpus

Figure 8 manifests the classification performance curves for hybrid eigenvector and TF-IDF eigenvector on *Adjusted corpus*, after using the NWKNN algorithm. Note that Exponent takes 3 empirically [15]. From the figure, it is obvious that our hybrid eigenvector beats TF-IDF eigenvectors under all selected feature numbers of term-space [16] by about 11%~15% on *Adjusted corpus*.

Figure 9 illustrates the classification performance comparison between hybrid eigenvector and TF-IDF eigenvector using different Exponent on *Adjusted corpus*, respectively. Note that the feature number of term-space takes 10,000. In the figure, with the increase of Exponent, our hybrid eigenvector performs better on *Adjusted corpus*, and beats TF-IDF eigenvector by 12% averagely.

Figure 10 describes the Macro-Precision and Macro-Recall comparison between hybrid eigenvector and TF-IDF eigenvector using different Exponent on *Adjusted corpus*, respectively. Note that the feature number takes 10,000. From the two figures, it is an apparent phenomenon that with the increase of Exponent, the curves accord with the experience of NWKNN [15]. Meantime, Macro-Precision or Macro-Recall of hybrid eigenvector is superior to the TF-IDF eigenvector by 12% or 14% on *Adjusted corpus* averagely.

5. Conclusion

In this work, data structure of semantic-element information is constructed to record relevant information of each semantic-element in document sample. It can characterize lexical semantic contents and be adapted for disambiguation of word stems. The hybrid eigenvector using the NWKNN algorithm achieve better performance of classification than TF-IDF eigenvector which stands for the typical term-statistical method of feature extraction, especially, for impact of lexical replacement.

The future research includes using more current algorithms based on the hybrid eigenvector for text corpus analysis, and developing a method for representing semi-structured document such as XML on the basis of semantic-element.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (Grant No. 61402165), Nature Science Foundation of Hunan Province of China (No. 2015JJ3058, 12JJ3069), Scientific Research Fund of Hunan Provincial Education Department (No. 15C1288) and Informational-education Application Project of universities in Hunan Province. We would like to thank the anonymous referees for their helpful comments and suggestions. The authors are also grateful for the help of Xin-pan Yuan.

References

[1] J. Li-Ping, M. K. Ng, H. Z. Joshua, “Knowledge-based vector space model for text clustering”, Knowledge and Information Systems, vol. 25, no. 1, **(2010)**, pp. 35-55.

[2] W. Zhang, T. Yoshida and X. Tang, “A comparative study of TF*IDF, LSI and multi-words for text classification”, Expert Systems with Applications, vol. 38, no. 3, **(2011)**, pp. 2758–2765.

[3] Y. Zhang, R. Jin and Z.-H. Zhou, “Understanding bag-of-words model: a statistical framework”, International Journal of Machine Learning and Cybernetics, vol. 1, no. 1, **(2010)**, pp. 43-52.

[4] P. Li, A. Shrivastava and A. C. König, “b-Bit minwise hashing in practice”, Proceedings of the 5th Asia-Pacific Symposium on Internetware, New York, **(2013)**, pp. 13-22.

[5] A. O. Hamid, B. Behzadi, S. Christoph and M. Henzinger, “Detecting the origin of text segments efficiently”, Proceedings of the 18th International Conference on World Wide Web, New York, **(2009)**, pp. 61-70.

[6] D. Sanchez and M. Batet, “A semantic similarity method based on information content exploiting multiple ontologies”, Expert Systems with Applications, vol. 40, no. 4, **(2013)**, pp. 1393–1399.

[7] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography”, Computational linguistics, vol. 16, no. 1, **(1990)**, pp. 22–29.

[8] G. A. Miller, “WordNet: a lexical database for English”, Communications of the ACM, vol. 38, no. 11, **(1995)**, pp. 39–41.

[9] “Function words”, **(2015)**, http://www.sequencepublishing.com/academic.html.

[10] M. Lintean and V. Rus, “Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics”, Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, Marco Island, USA, **(2012)**, pp. 244-249.

[11] “MIT. MIT Java Wordnet Interface”, **(2013)**, http://projects.csail.mit.edu/jwi/api/edu/mit/jwi/.

[12] Z. Ling-Yun, L. Fang-Ai and Z. Zhen-Fang, Editors, “Frontier and future development of information technology in medicine and education”, Identification of evaluation collocation based on maximum entropy model, Springer Publishers, New York, **(2013)**.

[13] M. Hwang, C. Choi and P. Kim, “Automatic enrichment of semantic relation network and its application to word sense disambiguation”, IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, **(2011)**, pp. 845-858.

[14] C. J. Keylock, “Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy”, Oikos, vol. 109, no. 1, **(2005)**, pp. 203-207.

[15] S. Tan, “Neighbor-weighted k-nearest neighbor for unbalanced text corpus”, Expert Systems with Applications, vol. 28, no. 4, **(2005)**, pp. 667-671.

[16] S. Tata and J. M. Patel, “Estimating the selectivity of tf-idf based cosine similarity predicates”, ACM SIGMOD Record, vol. 36, no. 2, **(2007)**, pp. 7-12.

[17] C. C. Aggarwal and Z. Cheng-Xiang, Editor, “Mining Text Data: A survey of text classification algorithms”, Springer Publishers, New York, **(2012)**.

Authors

Luda Wang, He received the M.S. degree of Computer Application Technology from Hunan University in 2009, and he is currently a Ph. D. candidate of Computer Science and Technology in Central South University (CSU). His research interests include AI, Data Analysis and Information Retrieval.

Peng Zhang, She received the M.S. degree of Computer Application Technology from Hunan University in 2010. She has been a faculty member of Computer Science at Xiangnan University, China, since 2004, where she is currently a lecturer. Her research interests include Information Retrieval and Information Fusion.

Shouping Gao, He received the Ph. D. degree of Applied Mathematics from Tongji University in 2003, He has been a faculty member of Computer Science at Xiangnan University, China, since 1991, where he is currently a professor. His research interests include Data Analysis and Symbolic Computation.