J. Cent. South Univ. (2015) 22: 1833−1840
DOI: 10.1007/s11771-015-2702-8

Springer

WordNet-based lexical semantic classification for text corpus analysis

LONG Jun(龙军)[1], WANG Lu-da(王鲁达)[1], LI Zu-de(李祖德)[1], ZHANG Zu-ping(张祖平)[1], YANG Liu(杨柳)[2]

1. School of Information Science and Engineering, Central South University, Changsha 410075, China;
2. School of Software, Central South University, Changsha 410075, China

© Central South University Press and Springer-Verlag Berlin Heidelberg 2015

Abstract: Many text classifications depend on statistical term measures to implement document representation. Such document representations ignore the lexical semantic contents of terms and the distilled mutual information, leading to text classification errors. This work proposed a document representation method, WordNet-based lexical semantic VSM, to solve the problem. Using WordNet, this method constructed a data structure of semantic-element information to characterize lexical semantic contents, and adjusted EM modeling to disambiguate word stems. Then, in the lexical-semantic space of corpus, lexical-semantic eigenvector of document representation was built by calculating the weight of each synset, and applied to a widely-recognized algorithm NWKNN. On text corpus Reuter-21578 and its adjusted version of lexical replacement, the experimental results show that the lexical-semantic eigenvector performs $F1$ measure and scales of dimension better than term-statistic eigenvector based on TF-IDF. Formation of document representation eigenvectors ensures the method a wide prospect of classification applications in text corpus analysis.

Key words: document representation; lexical semantic content; classification; eigenvector

1 Introduction

Text corpus analysis is an important task. Meanwhile, clustering and classification are the 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. Eigenvector-based document classification is a widely used technology for text corpus analysis. Relevant classification algorithms and experiments are typically based on eigenvector of document representation. Moreover, the key issue is eigenvector-based classification algorithms depending 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 Bag-of-Words [3] and Minwise hashing [4]. For document representation, these methods are perceived as statistical methods of feature extraction.

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

Semantic approach is an effectively used technology for document analysis. It can capture the semantic features of words under analysis, and based on that, characterizes and classifies the document. Close relationship between the syntax and lexical semantic contents of words have attracted considerable interest in both linguistics and computational linguistics.

The design and implementation of WordNet-based lexical semantic classification take account of lexical semantic content particularly. Unlike traditional statistical methods of feature extraction, our work developed a new term measure which can characterize lexical semantic contents, and provide a practical method of document representation to can handle the impact of lexical replacement. The document representation is normalized as the eigenvector; consequently, it shall be applied to current VSM-dependent classification algorithms. Theoretical analysis and a large number of experiments

Foundation item: Project(2012AA011205) supported by National High-Tech Research and Development Program (863 Program) of China; Projects(61272150, 61379109, M1321007, 61301136, 61103034) supported by the National Natural Science Foundation of China; Project(20120162110077) supported by Research Fund for the Doctoral Program of Higher Education of China; Project(11JJ1012) supported by Excellent Youth Foundation of Hunan Scientific Committee, China
Received date: 2014−03−21; **Accepted date:** 2014−10−11

Corresponding author: WANG Lu-da, PhD Candidate, Lecturer; Tel: +86−18613082443; E-mail: wang_luda@csu.edu.cn

are carried out to verify the effectiveness of this method.

2 Analysis of statistical term measure

In the 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]. Many TF-IDF-based 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 of shingle. Then, a selection algorithm determines which shingles to store in a hash table. And various estimation techniques are used to determine which shingle is copied and from which most of the content originated [5].

In the above, these methods for document representation are perceived as the mode using statistical term measures. As a sort of ontology methods [6], document representations based on statistical term measures 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 samples. Moreover, lexical replacement of document original cannot be represented literally by radical statistical mechanisms of term measure. Our comment on statistical term measures and document representation can be clarified by analyzing a small text corpus example 1.

Example 1.

Sample A: Men love holiday.

Sample B: Human enjoys vacation.

In example 1, the two simple sentences are viewed as two document samples, and these two documents comprise the small corpus. Evidently, the meanings of sample A and sample B are extremely equivalent. Thus, the correlation and semantic similarity between these two documents are considerable. Meanwhile, sample B can be regarded as a derivative of sample A via lexical replacement of document original. The text segmentation shall divide each document into meaningful terms, such as words, sentences, or topics. As to example 1, all words of documents are divided as terms. Obviously, on behalf of statistical term measures, the document representations on Example 1 did not perform well, which are listed in Table 1 and Table 2.

Table 1 Statistical term measures on Sample A

Term	Men	Love	Holiday	Human	Enjoys	Vacation
Weight (frequency)	1	1	1	0	0	0

Table 2 Statistical term measures on Sample B

Term	Men	Love	Holiday	Human	Enjoys	Vacation
Weight (frequency)	0	0	0	1	1	1

Comparing Tables 1 and 2, positive weights do not coexist in the same term of two samples. These two orthogonal vectors of term weight demonstrate that the statistical term measures for document representation cannot effectively signify semantic similarity of the corpus example 1. And 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 term meanings.

3 Proposed program

3.1 Motivation and theoretical analysis

For text corpus analysis, document representations which depend on statistical term measures shall lose mutual information of term meanings. Besides, in different documents, term meanings are relevant to specific synonyms which are involved by lexical semantic contents. Thus, this new work resorts to WordNet [8], a lexical database for English, for extracting lexical semantic contents. Then, the method of document representation will construct a lexical semantic VSM of text corpus to define eigenvector for text classification.

In WordNet, a form is represented by a string of ASCII characters, and a sense is represented by the set of (one or more) synonyms that have that sense [8]. Synonymy is WordNet's basic relation, because WordNet uses sets of synonyms (synsets) to represent word senses. Videlicet, shown as Fig. 1, one word, refers to several synsets.

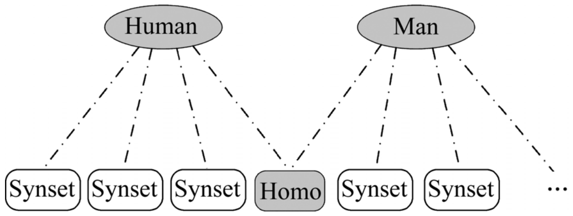

Fig. 1 Common semantic-element of words

In WordNet, because one word or term refers to particular synsets, our motivation is that several particular synsets can strictly describe the meaning of one word for characterizing lexical semantic contents.

Then, our method defines these particular synsets as the semantic-elements of word.

Based on the above definition, involved semantic-elements can character the lexical semantic contents of Example 1, which shall accomplish feature extraction of lexical semantic contents. For instance, in Fig. 1, the words human and man belong to different document samples in Example 1, and the common semantic-element homo that simultaneously describes the meanings of human and man can gain mutual information [7] between term meanings. Moreover, our document representation is able to capture the lexical semantic mutual information between samples which lies in the same synonyms of different documents.

According to the statistical theory of communications, our work 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 variable representing distinct lexical semantic contents in sample X or Y, which occurs 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, $\mathcal{I}(\mathcal{X};\mathcal{Y})$, is specially defined as

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

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 $P(x_i, y_j)=0$ so the samples mutual information $\mathcal{I}(\mathcal{X};\mathcal{Y})=0$. Thus, the analysis verifies that the statistical methods of feature extraction lose mutual information of term meanings.

On the other hand, 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 described by same semantic-elements. Then, lexical semantic mutual information between samples, $\mathcal{I}(\mathcal{X};\mathcal{Y})$, is re-defined to be

$$\mathcal{I}(\mathcal{X};\mathcal{Y})=$$
$$\sum_{x_i\in X}\sum_{y_j\in Y}F(e_{x_i,y_j})\bmod N\lg\frac{F(e_{x_i,y_j})\bmod N}{F(e_{x_i})\bmod N\times F(e_{y_j})\bmod N} \tag{2}$$

To denote probability $P(x_i)$ or $P(y_j)$, function $F(e_{x_i})$ or $F(e_{y_j})$ 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 of semantic-elements in corpus. Meanwhile, e_{x_i,y_j} is the common semantic-elements that simultaneously describe the meaning of x_i and y_j, to denote joint probability $P(x_i, y_j)$, and function $F(e_{x_i,y_j})$ is estimated by calculating the frequency e_{x_i,y_j}, and modulo N.

In example 1, 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 Fig. 1). In reality, P(human, man)= F(homo) mod N>0, as a result, lexical semantic mutual information between Sample A and Sample B, $\mathcal{I}(\mathcal{X};\mathcal{Y})$, is positive. Thus, the analysis proves that the semantic-elements and feature extraction of lexical semantic contents can provide the probability-weighted amount of information (PWI) [2] between document samples on the lexical semantic level.

3.2 Lexical-semantic VSM of text corpus

In our work, documents are represented using the vector space model (VSM). The VSM represents each document as a vector of identifiers [1]. Each dimension corresponds to a separate feature value. If a feature occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed.

For organizing the lexical-semantic VSM of text corpus in the lexical-semantic space, the procedures are as follows. In the first place, for feature extraction of lexical semantic contents, our work makes a data structure of semantic-element information. Secondly, the work uses EM modeling to disambiguate word stems. Lastly, it constructs a lexical-semantic space and builds lexical-semantic eigenvectors in the space to characterize document samples.

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

Note that, in a record of the data structure, each

Table 3 Data structure of semantic-element information

Item	Explanation
Synset ID	Identification of synset
Set of synonym	All synonyms in the identical synset 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 Fig. 2(a)) which carries all original words of terms referring to semantic-element and their word stem (s)
Semantic members frequency	A linked list (shown in Fig. 2(b)) which carries frequency of each original words of terms (that refer to semantic-element) one by one

original word in inflected form [9] referring to the semantic-element and its word stem(s) in base form [9−10] are recorded by linked list of semantic member (shown in Fig. 2(a)). And, according to WordNet framework [8], when original word refers to more than 1 word stem, the linked-list of semantic member will expend the very node of the original word to register all word stems.

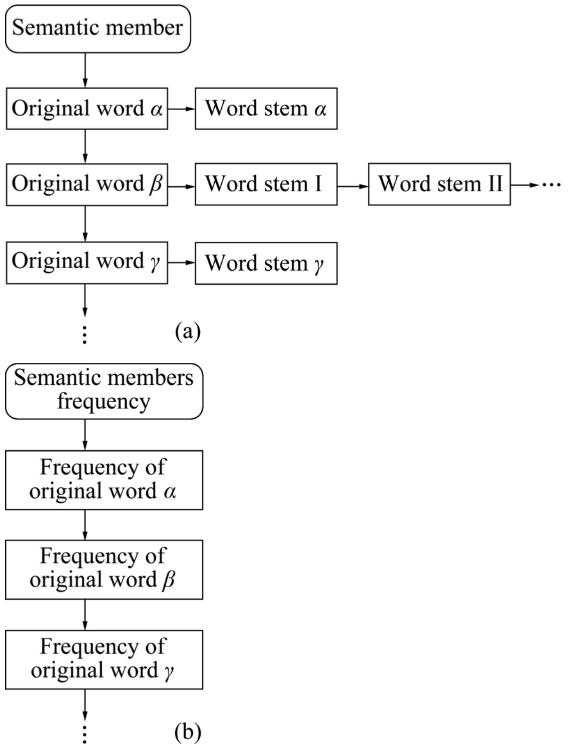

Fig. 2 Linked lists of semantic member (a) and semantic members frequency (b)

Meanwhile, the linked-list of semantic members frequency is shown in Fig. 2(b). It records the frequency of each original word one by one in original word order of semantic member. Both of the two linked lists carry the essential information of original words and word stems in the semantic-element.

2) On the basis of data structure of semantic-element information, semantic member needs to disambiguate word stems of original word. In the 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 Fig. 3), this method employs the maximum entropy model [11]. ME modeling provides a framework for integrating information for classification from many heterogeneous information sources [12].

In our model, it is supposed that diversity [13] 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 that original word x only refers to: $c|(x)=\arg\max_c p(c|x)$. Each feature [12] of original word is calculated by a function that is associated to a specific word stem c, and it takes the form of Eq. (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 [13−14].

The conditional probability $p(c|x)$ is defined by Eq. (4). The parameter of the semantic-element i [12], 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.

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

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

Above equations aim at finding the highest conditional probability $p(c|x)$, and using the function $c|(x)$ to ensure that original word x refers to only 1 word stem (like Fig. 3). After semantic-elements characterizing lexical semantic contents of a document preliminarily, the specified ME modeling is applied to implementing 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.

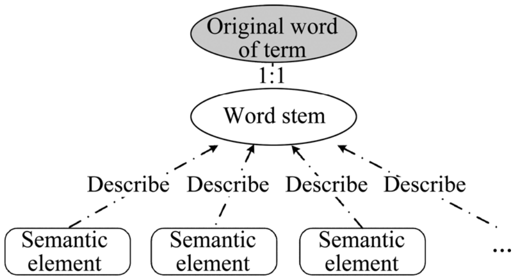

Fig. 3 1:1 reference of original word

3) The document representation uses the vector space model (VSM). 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 lexical-semantic VSM respectively. In lexical-semantic VSM, each document representation is marked in the lexical-semantic space of text corpus. Specifically, each document sample identified by sample ID is represented by the lexical-semantic eigenvector. The lexical-semantic VSM represents a document doc$_x$, using a lexical-semantic eigenvector $d_x \in R^{\mathrm{m}}$, given as

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

where m is the number of identical synset ID of all semantic-elements in corpus; $d_{x(i)}$ is the feature value on the ith synset, given as $d_{x(i)}=FS(s_i,\ \mathrm{doc}_x)\cdot F_{\mathrm{IDF}}(S_i)$ for all i=1 to m. $F_{\mathrm{S}}(s_i,\ \mathrm{doc}_x)$ is the weight (frequency) of the ith corresponding semantic-element s_i in document doc$_x$. And $F_{\mathrm{IDF}}(s_i)=\lg(D/N_{\mathrm{DF}}(s_i))$ is the inverse document frequency of s_i, where D is the sum of the documents in corpus, $N_{\mathrm{DF}}(s_i)$ is the number of documents in which the ith synset appears at least once.

4 Experiment and its results

To test the lexical-semantic VSM and verify the lexical semantic classification, this work uses two sorts of eigenvector to represent document in two datasets, and employs an effective algorithm to classify the documents. After that, contrast between our method and typical statistical method displays the effect of this work.

4.1 Eigenvectors for document representation

In our work, experiments use two sorts of eigenvector to represent document sample: 1) lexical-semantic eigenvector in the lexical-semantic VSM shown in Eq. (5); 2) term-statistic eigenvector in the term-space which takes different numbers of selected features using information gain [15]. Using the typical statistical method of feature extraction, TF-IDF, the term-statistic eigenvector, $d_x \in R^{\mathrm{n}}$, is given as [2]

$$d_x = (d_{x(1)}, d_{x(2)}, \cdots, d_{x(n)}) \tag{6}$$

where n is the number of terms in corpus; $d_{x(j)}$ is the feature value on the jth term, given as $d_{x(j)}=F_{\mathrm{TF}}(w_j,\ \mathrm{doc}_x)\cdot F_{\mathrm{IDF}}(w_j)$ for all j=1 to n, and $F_{\mathrm{TF}}(w_j,\ \mathrm{doc}_x)$ is the frequency of the term w_j in document doc$_x$ and $F_{\mathrm{IDF}}(w_j)$ is the inverse document frequency of w_j.

4.2 Datasets

Our experiments use two corpora: Reuter (http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html) and an adjusted corpus based on Reuter-21578.

1) Reuter. The Reuters-21578 text categorization test collection contains documents collected from the reuters newswire in 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 (listed in Table 4).

2) Adjusted corpus (based on Reuter-21578). After selecting the subset of Reuter-21578, the datasets 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 similar, and the meanings of them are extremely equivalent.

4.3 Classification using NWKNN algorithm

In the text corpus analysis, KNN classification is especially effective on selection of data eigenvectors. To tackle unbalanced text corpus, the experiments select an optimized KNN classification, the NWKNN (Neighbor-Weighted K-Nearest Neighbor) algorithm [14]. For NWKNN classification, each document d is represented

Table 4 Distribution of all categories in subset of Reuter-21578

Category	Cotton	Earn	Cpi	Rubber	Sugar	Money-fx	Bop	Grain	Heat	Money-supply
Sample	27	761	75	40	145	574	47	489	16	113
Category	Silver	Tin	Crude	Hog	Nat- gas	Jobs	Cocoa	Trade	Housing	Nickel
Sample	16	32	483	16	48	50	59	441	16	5

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% last year, up from 15% 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% last year, upward from 15% in 1985.

[15−16] using both lexical-semantic eigenvector and term-statistic eigenvector. Formally, the decision rule [14] in NWKNN classification can be written as

$$\text{score}(d,c_i)=\text{weight}_i\left(\sum_{d_j\in KNN(d)}\text{Sim}(d,d_j)\delta(d_j,c_i)\right)\tag{7}$$

$$\delta(d_j,c_i)=\begin{cases}1, & d_j\in c_i\\ 0, & d_j\notin c_i\end{cases}\tag{8}$$

where $KNN(d)$ indicates the set of K-nearest neighbors of document d; Sim(d, d_i) denotes the similarity between document d and d_i using cosine value between eigenvectors of d and d_i [14]; $\delta(d_j, c_i)$ is the classification for document d_j with respect to class c_i. Besides, according to experience of NWKNN algorithm [14], a parameter of weight$_i$, exponent [14], ranges from 2.0 to 6.0.

4.4 Performance measure

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

$$F1=\frac{2\times R_{\text{Recall}}\times P_{\text{Precision}}}{R_{\text{Recall}}+P_{\text{Precision}}}\tag{9}$$

where R_{Recall} is the recall; $P_{\text{Precision}}$ is the precision.

Using $F1$ measure, it can display the effect of different kinds of data on a text classification system [17]. For ease of comparison, our experiments summarize the $F1$ scores over the different categories using the macro-averages of $F1$ scores; in the same way, the Macro-Recall and Macro-Precision can be obtained [17].

Besides, to express the dimension reduction relatively, the experiments compare the numbers of dimension and increasing document number, which indicates the corpus scale. The comparison between two sorts of eigenvectors on dimensions can display optimization to the dimension problem [18].

4.5 Experimental results

To accomplish three-fold cross validation, the experiments conduct the training-test procedure on datasets Reuter and adjust corpus three times alternately, and use the average of the three performances as final result.

Using the NWKNN classification, Fig. 4 manifests the $F1$ measure curves of lexical-semantic eigenvector and term-statistic eigenvector on Reuter. Note that, the exponent takes 3 empirically [14]. It is obvious that the lexical-semantic eigenvector beats term-statistic eigenvectors under all selected feature numbers of term-space [16] by 4%−7% on Reuter.

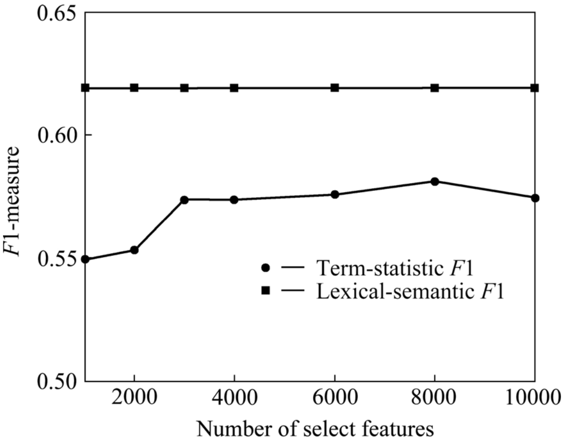

Fig. 4 Classification result of lexical-semantic eigenvector and term-statistic eigenvector with different term-space feature numbers on Reuter

Using different exponents in NWKNN, Fig. 5 illustrates the $F1$ measure comparison between lexical-semantic eigenvector and term-statistic eigenvector on Reuter, respectively. Note that the feature number of term-space takes 10000. With the increase of exponent, the lexical-semantic eigenvector performs better on Reuter, and beats term-statistic eigenvector by approximate 4% averagely.

Using different exponent in NWKNN, Fig. 6 describes the macro-precision and macro-recall comparison between lexical-semantic eigenvector and term-statistic eigenvector on Reuter, respectively. Note that the feature number of term-space takes 10000. It is an apparent phenomenon that with the increase of exponent, the curves accord with the experience of NWKNN [14]. Meantime, macro-precision or macro-recall of lexical-semantic eigenvector is superior to the term-statistic eigenvector by 7% or 8% on Reuter averagely.

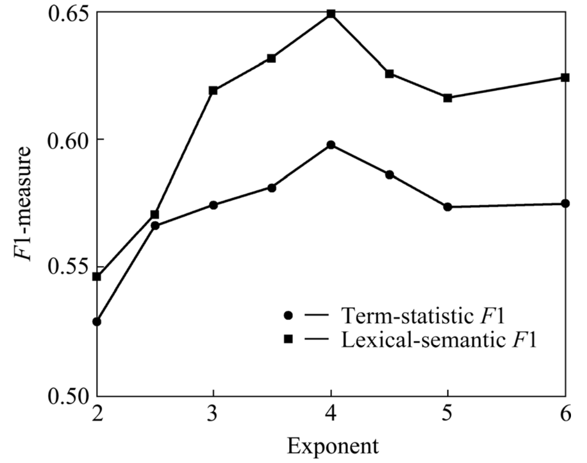

Fig. 5 Classification result of l lexical-semantic eigenvector and term-statistic eigenvector with different exponents on Reuter

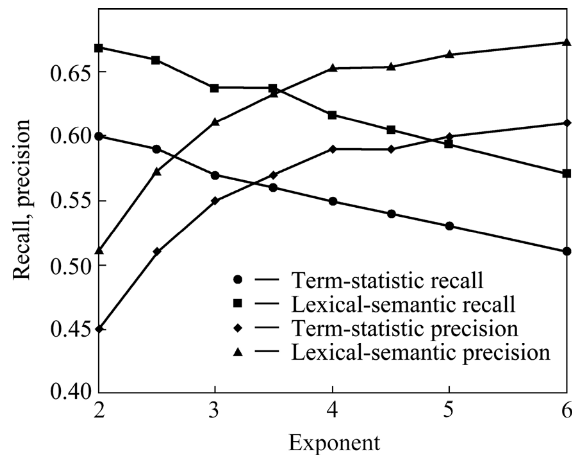

Fig. 6 Classification recall and precision of lexical-semantic eigenvector with different exponents and term-statistic eigenvector on Reuter

Using the NWKNN classification, Fig. 7 manifests the $F1$ measure curves for lexical-semantic eigenvector and term-statistic eigenvector on adjusted corpus. Note that, the exponent takes 3 empirically [14]. It is obvious that the lexical-semantic eigenvector beats term-statistic eigenvectors under all selected feature numbers of term-space [16] by 10%−13% on adjusted corpus.

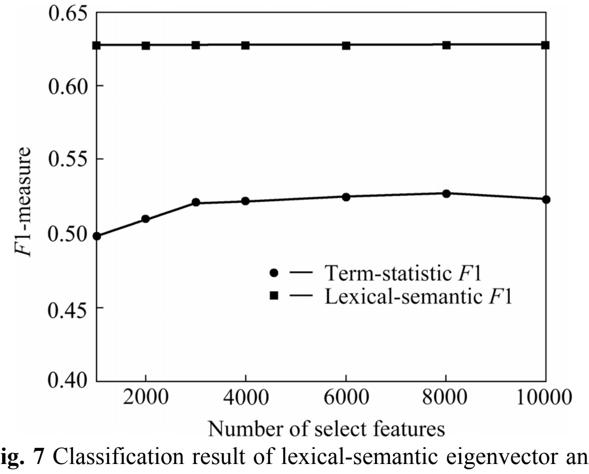

Fig. 7 Classification result of lexical-semantic eigenvector and term-statistic eigenvector with different term-space feature numbers on adjusted corpus

Using different exponents in NWKNN, Fig. 8 illustrates the $F1$ measure comparison between lexical-semantic eigenvector and term-statistic eigenvector on adjusted corpus, respectively. Note that the feature number of term-space takes 10000. With the increase of exponent, the lexical-semantic eigenvector performs better on adjusted corpus, and beats term-statistic eigenvector by 10% averagely.

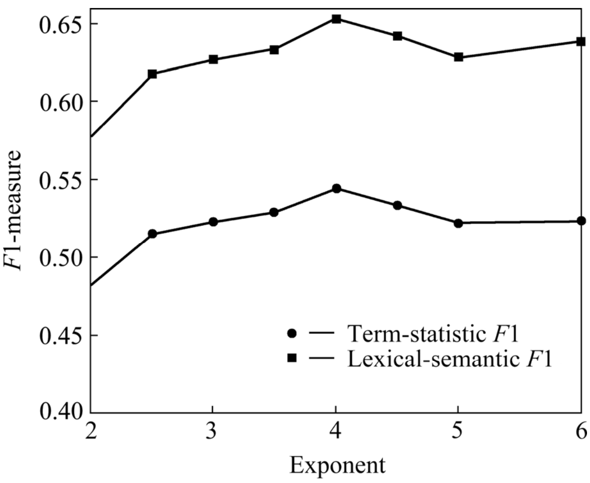

Fig. 8 Classification result of l lexical-semantic eigenvector and term-statistic eigenvector with different exponents on adjusted corpus

Using different exponents in NWKNN, Fig. 9 describes the macro-precision and macro-recall comparison between lexical-semantic eigenvector and term-statistic eigenvector on adjusted corpus, respectively. Note that the feature number of term-space takes 10000. It is an apparent phenomenon that with the increase of exponent, the curves accord with the

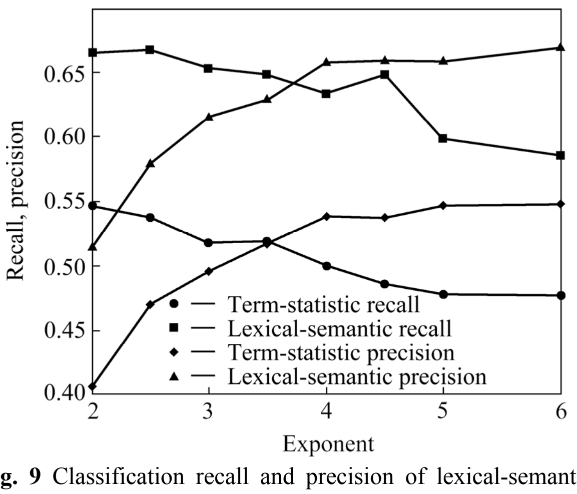

Fig. 9 Classification recall and precision of lexical-semantic eigenvector with different exponents and term-statistic eigenvector on adjusted corpus

experience of NWKNN [14]. Meantime, macro-precision or macro-recall of lexical-semantic eigenvector is superior to the term-statistic eigenvector by 11% or 12% on adjusted corpus averagely.

According to Eqs. (5) and (6), Fig. 10 reports the dimensionalities of lexical-semantic eigenvector and term-statistic eigenvector on Reuter. Note that number of document ranges from 200 to 2800. After the number of document reaching 1650, the dimensionality of lexical-semantic eigenvector is less than that of the term-statistic eigenvector. It indicates the improvement of dimension reduction in our method.

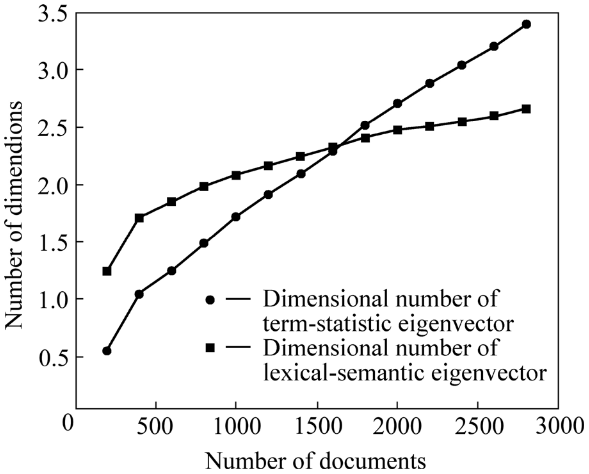

Fig. 10 Dimensionalities of lexical-semantic eigenvector and term-statistic eigenvector on Reuter-21578

5 Conclusion and future work

1) A 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.

2) The lexical-semantic eigenvector using the NWKNN algorithm achieves better performance of classification than term-statistic eigenvector which stands for the typical statistical method of feature extraction, especially, for impact of lexical replacement.

3) Our method of document representation demonstrates the improvement of dimension reduction for text classification.

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

References

[1] JING L P, NG M K, HUANG JOSHUA Z. Knowledge-based vector space model for text clustering [J]. Knowledge and Information Systems, 2010, 25(1): 35−55.

[2] ZHANG Wen, YOSHIDA Taketoshi, TANG Xi-jin. A comparative study of TF*IDF, LSI and multi-words for text classification [J]. Expert Systems with Applications, 2011, 38(3): 2758−2765.

[3] ZHANG Yin, JIN Rong, ZHOU Zhi-hua. Understanding bag-of-words model: a statistical framework [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1/2/3/4): 43−52.

[4] LI P, SHRIVASTAVA A, KONIG A C. b-Bit minwise hashing in practice [C]// Proceedings of the 5th Asia-Pacific Symposium on Internetware. New York: ACM, 2013: 13−22.

[5] HAMID A O, BEHZADI B, CHRISTOPH S, HENZINGER M. Detecting the origin of text segments efficiently [C]// Proceedings of the 18th International Conference on World Wide Web. New York: ACM, 2009: 61−70.

[6] SANCHEZ D, BATET M. A semantic similarity method based on information content exploiting multiple ontologies [J]. Expert Systems with Applications, 2013, 40(4): 1393−1399.

[7] CHURCH K W, HANKS P. Word association norms, mutual information, and lexicography [J]. Computational linguistics, 1990, 16(1): 22−29.

[8] MILLER G A. WordNet: A lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39−41.

[9] LINTEAN M, RUS V. Measuring Semantic similarity in short texts through greedy pairing and word semantics [C]// Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference. Marco Island, USA: AAAI, 2012: 244−249.

[10] MIT. MIT Java Wordnet interface (JWI) [EB/OL]. [2013−12−20]. http://projects.csail.mit.edu/jwi/api/edu/mit/jwi/morph/WordnetStemmer.html/.

[11] ZHAO Ling-yun, LIU Fang-ai, ZHU Zhen-fang. Frontier and future development of information technology in medicine and education: Identification of evaluation collocation based on maximum entropy model [M]. 1st ed. New York: Springer, 2013: 713−721.

[12] HWANG M, CHOI C, KIM P. Automatic enrichment of semantic relation network and its application to word sense disambiguation [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(6): 845−858.

[13] KEYLOCK C J. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy [J]. Oikos, 2005, 109(1): 203−207.

[14] TAN S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus [J]. Expert Systems with Applications, 2005, 28(4): 667−671.

[15] AGGARWAL C C, ZHAI C X. Mining text data: A survey of text classification algorithms [M]. 1st ed. New York: Springer, 2012: 163−222.

[16] TATA S, PATEL J M. Estimating the selectivity of tf-idf based cosine similarity predicates [J]. ACM Sigmod Record, 2007, 36(2): 7−12.

[17] van RIJSBERGEN C. Information retrieval [M]. London: Butterworths Press, 1979.

[18] YAN Jun, LIU Ning, YAN Shui-cheng, YANG Qiang, FAN Wei-guo, WEI Wei, CHEN Zheng. Trace-oriented feature analysis for large-scale text data dimension reduction [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 1103−1117.

(Edited by YANG Hua)