Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

February 2012, 19(1): 101–111
www.sciencedirect.com/science/journal/10058885

The Journal of China Universities of Posts and Telecommunications

http://jcupt.xsw.bupt.cn

Efficient representation of text with multiple perspectives

PING Yuan[1,2,3] (✉), ZHOU Ya-jian[1], XUE Chao[1], YANG Yi-xian[1]

1. Information Security Center, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Department of Computer Science and Technology, Xuchang University, Xuchang 461000, China
3. Science and Technology on Electronic Control Laboratory, Chengdu 610036, China

Abstract

An effective text representation scheme dominates the performance of text categorization system. However, based on the assumption of independent terms, the traditional schemes which tediously use term frequency (TF) and document frequency (DF) are insufficient for capturing enough information of a document and result in poor performance. To overcome this limitation, we investigate exploring the relationships between different terms of the same class tendency and the way of measuring the importance of a repetitive term in a document. In this paper, a group of novel term weighting factors are proposed to enhance the category contribution for each term. Then, based on a novel strategy of generating passages from document, we present two schemes, the weighted co-contributions of different terms corresponding to the class tendency and the weighted co-contributions for each term in different passages, to achieve improvements on text representation. The prior scheme works in a dimensionality reduction mode while the second one runs in the conventional way. By employing the support vector machine (SVM) classifier, experiments on four benchmark corpora show that the proposed schemes could achieve a consistent better performance than the conventional methods in both efficiency and accuracy. Further analysis also confirms some promising directions for the future works.

Keywords text representation, support vector machine (SVM), class tendency, category contribution, passages

1 Introduction

Currently, text categorization, the task of automatically categorizing unlabeled documents into predefined class labels, is a prospect techniques which is being employed in many cases of content-based document management tasks, such as information fusion, information filtering, information retrieval, user-interactive question answering and even sensitive information analysis.

In order to achieve an expected performance, many researchers pay attention to improvements on the classifiers, such as Naïve Bayes [1], *k*-nearest neighbor (*K*NN) [2] and SVM [3]. However, Leopold et al. [4] pointed out that it is the text representation scheme which dominates the performance of text categorization. Moreover, Xue et al. [5] found that the structural information of data in real-world problem carried by vector is the vital factor for achieving good performance. That is, at least, the way of vectorizing a document with sufficient information, whether employing the prior knowledge implicitly or not, is essential to reaching a further improvement on accuracy. In spite of many forms of terms surveyed in Ref. [6–7], the standard ‘bag-of-word (BOW)’ indexing is widely used for its ability of generalization and performance. Furthermore, to describe the topic of a document, different terms might have distinguished contributions. Therefore, varied term weighting schemes, such as χ^2, information gain (IG), gain ratio (GR), odds ratio (OR), etc., had been developed to measure these differences. Among these popular term weighting schemes, the relevance frequency (RF) which consistently achieves the best categorization performance was introduced by Lan et al. [8–9] and confirmed by analytical evaluations in Ref. [10]. Since these measures are almost based on the TF and DF, with the assumption of

Received date: 15-05-2011
Corresponding author: PING Yuan, E-mail: pingyuan@bupt.edu.cn
DOI: 10.1016/S1005-8885(11)60234-3

independent terms, can they carry sufficient information for representing a document? In order to answer this question, in this section, we would like to consider the following three scenarios.

Actually, for the first, many terms in a document are semantically (conceptually) related. If not, practically, they would also contribute to the same category. For instance, in 20Newsgroup, ‘secret’ is a discriminator (feature) for sci.crypt while another term ‘bank’, whose positive category is sci.med with frequency of 86 also repeats 28 times in sci.crypt. Thus, most of the terms contribute to one or more categories. However, the conventional methods prefer to the supervised term weighting methods with ‘global policy’ [11] where a document will have a global single representation that ignores the multiple categories contributions from each term. Although Dino et al. [12–13] tried to combine these multiple contributions, the solution still suffers from the solely reliance on TF. Therefore, intuitively, an effective scheme which is able to measure and combine these contributions of different terms in a document is expected to be developed.

Secondly, the traditional term weighting schemes subject to some limitations of describing the exact contribution from each term to its positive category. Generally, the term weighting scheme comprises of an importance part of a term in a document and a term weighting factor which is used as a discriminator between categories [14–15]. Considering DF of a term in its positive category and negative category, in the training procedure, the greater of the ratio, the stronger the ability of the term to distinguish among categories is. However, they leave the exact distribution from a term in documents to its positive category out of consideration. Actually, based on the term frequency assumption, multiple appearances of a term in a document are no less important than single appearances [11,16–17]. Therefore, the exact state what we expect for a term weighting factor is that it should appropriately reflect the difference between terms with different TF but the same DF in their positive category.

The final scenario is related to the importance part of a term weighting scheme. Considering the way of retrieving information by manual, usually, people try to obtain the crucial information from the special part of a document, i.e., title, keywords, subject, abstract, conclusion. If there is a term t_A appears one time in each of those five parts, and a term t_B repeats 5 times in a passage besides those fives in a document; then they would get the same TF which is 5 for the traditional term weighting schemes. Does the two terms contribute the same to their positive category? Actually, many researchers [6,14,18–20] believe that the difference inherently exists in terms of different positions in a document. However, they either measure the similarity of sentences and title for ranking the importance [18] that is hard to be applied in documents without title or assign categories to each passages and merge the passage categories to the document categories [19] that is insufficient for capturing enough information. Qi et al. [7] suggested employing a function of the first appearance position of a term to measure the importance while Xue et al. [6] carried out three implementations of calculating the compactness of the appearances for a term. Unfortunately, the prior method ignores the contributions of a terms with respect to its appearances except for the first one, and the latter generates a sparse distribution of a term frequently that causes the system unstable.

By integrating the above considerations, this study aims at developing an exhaustive empirical study of enriching information carried by the document vector, and pays special attention to the different ways of combining the contributions from different terms and from each term with different locations. The main novelties of this paper with respect to the referred objectives consist in the following points:

1) Passage is a kind of independent semantic units. we present a strategy of partitioning any document into passages. To capture the structural information, a self-adaptive mechanism is introduced into the strategy to adjust the imbalanced passages.

2) After analyzing the importance of a term with respect to distribution in its positive category, a group of term weighting factors are proposed for accurately capturing the contribution of each term.

3) Since each term could contribute to all categories, based on the inference of passages’ importance [19], a weighted combination of contributions for terms which works in reduced dimension mode is proposed for efficiency while carry structural information in the document vectors.

4) To capture the distinct importance of distributed terms, based on the passage strategy, we present a novel scheme to combine the contributions for each appearance of a term with respect to location. It can be employed as a

substitute for the importance part of any term weighting scheme. Especially, when combined with the proposed term weighing factor, the novel term weighting scheme outperforms the state-of-the-art methods significantly.

The remainder of this paper is arranged as follows: In Sect. 2, we present the strategy of partitioning document into passages, the base form of term weighting factor with its variants, the class tendency-based combination of contributions from different term and the scheme of combining contributions for each term separately, as well as a short discussion of these schemes. We show experiments results in Sect. 3, and finally, we draw conclusions and discuss future works in Sect. 4.

2 Representation of text with multiple perspectives

2.1 Strategy of splitting document into passages

Callan [21] discussed that there are three types of passages: discourse passage, semantic passage and windows passage. On the one hand, although a document may contain a huge number of data with tags, there is no guarantee that these tags are all beneficial to extracting the discourse passages. And even some of them indeed interfere the analysis of the document's structure. Since the removal of these tags might result in varied lengths of discourse passages which can reduce the efficiency of text categorization, it is essential to smooth the sequence of the passages' length. On the other hand, it is difficult and time-consuming for partitioning a document into sematic passages on account of lacking an unique topic in any document. Thus the proposed strategy prefers a compromising solution from the frameworks of discourse passage and window passage without overlapping.

Algorithm 1 Document partition

```
Require: Document $D$ with length $|D|$
Ensure: Passages: $P = \{P_1, P_2, \ldots, P_M\}$
  Initialize the maximal size of passage $L_P$ and the threshold $\lambda$.
  $(P, L)$ = PassageSplit($D$) // $P$ is the set of discourse passages in which the passage is denoted by $P_i (i \in [1, N_P])$, the length $L_i$ for each passage is collected into the set $L$
  $F_P$ = FixPassages($P$) // return the set of fixed passages $F_P$ which have the obvious tags of crucial information
  $P_R = P / F_P$ // the set of passages excluding from $F_P$
  $\bar{L} = \frac{1}{|P_R|} \sum_{i=1, P_i \notin F_P}^{N_P} L_i$, $\delta = \left( \frac{1}{|P_R|-1} \sum_{i=1, P_i \notin F_P}^{N_P} |L_i - \bar{L}|^2 \right)^{\frac{1}{2}}$
  $i \leftarrow 1$
  while $i \leqslant N$ and $P_i \neq$ NULL
     if $P_i \in F_P$ then $i \leftarrow i+1$ continue; end if
     if $P_i \notin F_P$ and $L_i < \bar{L} - \delta$ then
        if $P_{i+1} \notin F_P$ then Merge($P_i, P_{i+1}$), $i \leftarrow i+2$, continue; end if
        if $P_{i+1} \in F_P$ then $i \leftarrow i+1$, continue; end if
     end if
     if $P_i \notin F_P$ and $L_i > \bar{L} + \delta$ then
        SplitPassage($P_i$) // split $P_i$ into two passages in equal length
        $i \leftarrow i+1$
     end if
  end
  Update($P$) // the number of passages is adjusted from $N_P$ into $M$
```

Algorithm 1 gives the detail steps of the strategy to automatically partition a document into passages. Lines 1 initializes the maximal size of passage L_P and the threshold λ. Lines 2–3 extract the discourse passages P and the fixed passages F_P of the document. The fixed passage is a special part of a document which carries the crucial information that is usually focused on by people, for instance, the title, keywords, subject, abstract, conclusion etc. In consideration of the contributions from the fixed passages, lines 4–18 give the steps to smooth the imbalanced passage sequence without affecting the fixed passages. The output of the Algorithm 1 is the set of passages in which the fixed passages and the relatively semantic passages with similar length are stored.

2.2 Term weighting factors with category contribution enhanced

Usually, the way of weighting a term is distinct from the scheme of feature selection. To accurately measure a term, the basic idea of our intuitive consideration is quite simple: the exact capability of a term to distinguish categories originates from both of its interior contribution to positive category and exterior divergence among categories. The former is related to term frequency while the latter corresponds to the document frequency.

Considering a scenario, a term t has equal value sequences $\{a_1,b_1,c_1,d_1\}$ and $\{a_2,b_2,c_2,d_2\}$ in the two-way contingency table of a word *t* and a category C while C_1 and C_2 are the positive categories respectively. Simultaneously, the frequencies of *t* in C_1, C_2 are t_1, $t_2 (t_1 \gg t_2)$. Taking RF as an example, the two weights $W_{\text{RF}}(C_1,t)$ and $W_{\text{RF}}(C_2,t)$ are identical. Is that means any document containing *t* should be sentenced to C_1 and C_2 in the same probability? Obviously, $t_1 \gg t_2$ indirectly reflects a more uniform of distribution of *t* in C_1 than that of in C_2. Similar with the measure of importance for terms occurring in one document by the distributional feature in Ref. [6], the different distributions of a term in corpus by category also show its differences in degree of importance. In fact, the widely used Naïve Bayes classifier has explained the difference by the posterior probability of the inequality $\Pr(C_1 \mid t) \neq \Pr(C_2 \mid t)$. Together with two parts of the intuitive consideration, that is to say, the distribution of inner-class [22–24] related with the TF is as important as the DF in constructing the term weighting scheme.

Although the supervised term weighting factors surveyed in the previous section perform well, they rely on the DF which mainly concentrates in describing the exterior divergence. Even though the TF is combined with one of these factors, the final weight is limited in representing the interior contribution to the positive category since the value of the TF is gathered from one document. Therefore, we propose a group of novel term weighting factors with category contribution enhanced (CCE) to capture the previous basic idea. Its formula is expressed as

$$W_{\text{CCE}}(C_k,t) = \alpha^{\Pr(C_k|t)} W_{\text{RF}}(C_k,t) = \alpha^{\frac{\Pr(t|C_k)\Pr(C_k)}{\Pr(t)}} \text{lb}\left(2 + \frac{a}{\max(1,c)}\right) = \alpha^{t_k \big/ \sum_{k=1}^{N_c} t_k} \text{lb}\left(2 + \frac{a}{\max(1,c)}\right) \tag{1}$$

where t_k is the total appearances of *t* in category $C_k (k \in [1, N_{\text{c}}])$, N_{c} is the total number of categories, *a* and *c* are the document frequencies of positive category and negative category which contain the *t*, respectively. In Eq. (1), a regulatory factor $\alpha (\alpha > 1)$ is suggested to be either 2 or e corresponding to denotations CCE$_2$ and CCE$_{\text{e}}$ respectively. In this paper, we also consider a special case

$$W_{\text{CCE}}(C_k,t)\big|_{\alpha=0} = \Pr(C_k \mid t) W_{\text{RF}}(C_k,t) \tag{2}$$

whose weight is independent with α. Finally, according to the global policy, the value of $W_{\text{CCE}}(t) = \max\limits_{k \in [1,N_{\text{c}}]} W_{\text{CCE}}(C_k,t)$ would be assigned to *t*. For the Eq. (2), we denote it by CCE$_0$ for simplicity.

2.3 Co-contributions of terms by class tendency

Definition 1 Class tendency and class bias. In text categorization, once a term existing in at least one document of any category, it should have a positive or negative impact on the decision of classification. The impact of representing a tendency to a category is called class tendency and its degree could be approximately measured by the weight of a term to any positive category. Among these measurements, the label corresponding to the maximal value $\arg\max W_{\text{CCE}}(C_k,t)$ is defined as class bias that represents the exact category might be assigned to the document with the maximum probability in which *t* appears.

Practically, each term of a document would contribute to all the categories more or less. That is why classifiers can hardly obtain hundred percent in accuracy. Those diverse contributions are approximately reflected by the weights of a term to all the positive categories. Therefore, any two documents with the same label do not need to be consisted of by the same terms, because the different terms could have the same class bias. It is worthy of trying to combine these contributions to each category from different terms of a document respectively, and to represent the document by the collection of values on class tendencies. Out of this reason, we also introduce the distributional coefficient [17] which is derived and revised from the Refs. [6,19] and is defined by the weighting function $W_{\text{DC}}(m,t)$ to describe the importance of *t* in the *m*th passage of a document or being the *m*th term of a single passage document. The five weighting functions for distributional coefficient employed in this paper are listed in Table 1 where the parts of a document is logically described by Head, Body and Tail in the same way of Refs. [6,19].

Table 1 Functions of distributional coefficient

$W_{\text{DC}}(m,t)$	Weighting tendency	Denoted by
1	Head=Body=Tail	DC$_{\text{SUM}}$
$1/m$	Head ≫ Body ≫ Tail	DC$_{\text{GI}}$
$1/\text{lb}(m+1)$	Head>Body>Tail	DC$_{\text{GLI}}$
$1 - m/M$	Head>Body>Tail	DC$_{\text{LL}}$
$\{\lvert m-[(M-1)/2]\rvert+1\}/M$	Head=Tail>Body	DC$_{\text{LVL}}$

Note: *M* is the total number of passages of document.

Considering both the class tendency and the distributional coefficient, the Algorithm 2 outlines the strategy of vectorizing a document in reduced dimension way (SVDRD). Since the line 2 travels the whole passage set, lines 3–8 merge all the terms with respect to the class tendency in each passage by calculating $W_{\mathrm{DC}}(m,t_{mj})\times W_{\mathrm{CCE}}(C_k,t_{mj})$. Finally, in lines 10–13, the average of the weighted co-contributions for all the terms on class tendency are employed to form the document vector. Apparently, the method of vectorizing document presented by Refs. [12,25] is a special case of the Algorithm 2 while the $\mathrm{DC}_{\mathrm{SUM}}$ is chosen as the distributional coefficient.

Algorithm 2 The SVDRD algorithm

```
Require:  In corpora with $N_c$ categories, a document
          $D$ consists of $M$ passages with each one
          $P_m(m \in [1,M])$ contains $N_m$ terms
Ensure:   Vector: $\boldsymbol{V}_D = [V_{D_1}\ V_{D_2}\ \dots\ V_{D_{N_c}}]$
  $N \leftarrow 0$   // the total number of terms in $D$
  for $m=1$ to $M$ do
    for $j=1$ to $N_m$ do
      for $k=1$ to $N_c$ do
        $V_{D_k} += W_{\mathrm{DC}}(m,t_{mj})\times W_{\mathrm{CCE}}(C_k,t_{mj})$
      end
      $N \leftarrow N+1$
    end
  end
  for $k=1$ to $N_c$ do
    $V_{D_k} \leftarrow \dfrac{V_{D_k}}{N}$
  end
  return $\boldsymbol{V}_D = [V_{D_1}\ V_{D_2}\ \cdots\ V_{D_{N_c}}]$
```

2.4 Co-contributions of each term with respect to passages

Apart from the SVDRD which combines the contributions from different terms, in this section, we also propose a novel scheme to integrate the contributions of each term with respect to different passages. Xue et al. [6] had proposed a term weighting scheme with the distributional features consisting of two parts: the term weighting factor (IDF was employed) and the strategy of the first appearance (FA) for a term. Although a group of weighted term frequency (WET) features which is formulated by Eq. (3) are carried out by using a weighting function (in sentences) to weight each appearance of a term, they had proved that the WET features perform worse than the FA features.

$$W_{\mathrm{WET}}(t,D)=\frac{\sum_{q=0}^{N_s-1}s_q W_{\mathrm{DC}}(q,t)}{s(D)} \tag{3}$$

In the Eq. (3) , the distributional sequence of the word *t* is $\{s_0,s_1,\dots,s_{N_s-1}\}$ corresponding to N_s sentences in the document D with $s(D)$ words. However, since both of the WET features and the FA features are counted based on sentence, a great amount of additional computational cost and storage are required. Therefore, in this paper, we develop a group of novel weighting term frequency (NWET) features and a group of normalized NWET (N2WET) which are calculated by Eq. (4) and Eq. (5) respectively based on the previously proposed passages.

$$W_{\mathrm{NWET}}(t,D)=f_t\left(\sum_{m=1}^{M}W_{\mathrm{DC}}(m,t)\right) \tag{4}$$

$$W_{\mathrm{N2WET}}(t,D)=\frac{\sum_{m=1}^{M}W_{\mathrm{DC}}(m,t)}{f_t} \tag{5}$$

In the above equations, *M* is the number of passages in document *D* partitioned by the Algorithm 1, f_t is the term frequency of *t*. Corresponding to the distributional coefficient functions in Table 1, four NWET features and four N2WET features can be generated, i.e., $\mathrm{NWET}_{\mathrm{GI}}$, $\mathrm{NWET}_{\mathrm{GLI}}$, $\mathrm{NWET}_{\mathrm{LL}}$, $\mathrm{NWET}_{\mathrm{LVL}}$, $\mathrm{N2WET}_{\mathrm{GI}}$, $\mathrm{N2WET}_{\mathrm{GLI}}$, $\mathrm{N2WET}_{\mathrm{LL}}$, and $\mathrm{N2WET}_{\mathrm{LVL}}$. Obviously, in comparison with Eq. (3), the proposed two group of features which can only be generated synchronously with the scanning process require less computational cost and storages.

3 Experiments and analysis

3.1 Data corpora

The first corpus is the gathered from the top 10 largest categories of the “ModApte” split of the Reuters-21578 [http://kdd.ics.uci.edu/databases/reuters21578/]. With reduplicate versions existing in their multi-topics separately, 9 990 news stories have been partitioned into a training set of 7 199 documents and a test set of 2 791 documents. In this preprocessing, the Porter’s stemming [26] is done to reduce words to their base forms. We also conduct experiments by omitting the words that occur less than 3 times or is shorter than 2 in length.

The 20Newsgroups collection [27] which has approximate 20 000 newsgroup documents nearly evenly divided among 20 discussion groups is used as the second corpus in our experiments. We partition it into ten subsets in equal size and conduct a number of experiments by randomly selecting three subsets for training and the remaining seven subsets for testing. The Subject and Keywords parts of each document are preserved as fixed passages. The other operations are the same with those done for Reuters-21578.

The third corpus is WebKB [28]. Following the work of Ref. [6], four categories, course, faculty, project and student, which contain 4 199 documents are experimented. Among these categories, 300 samples of course, 400 samples of faculty, 200 samples of project and 500 samples of students are selected randomly for training while the remaining 2 799 documents for testing. Except for some tags useful to the Algorithm 1, such as '<H>' and '</H>' for Title, '<p>' and '</p>' for paragraph etc., the others are removed. After stop words removal, stemming and omitting the short words or words of frequency lower than 3, the top 4 000 features in the resulting vocabulary is tried for experiments.

The last Ohsumed collection used here is a subset of clinically oriented MEDLINE from year 1987 to year 1991, consisting of 348 566 references form 270 medical journals over five year period. 50 216 documents of Ohsumed in year 1991 have abstracts. Following the works of Refs. [8–9,15], we use the first 10 000 documents for training and the second 10 000 documents for testing. This corpus including medical abstracts from the MeSH categories are related to 23 cardiovascular diseases, and each disease corresponds to one category label. After the removal of the duplicate issues, there are 6 286 documents and 7 643 documents retained for training and testing respectively. According the preprocesses done for the prior three corpora, the top 16 000 features are selected for the experiments.

3.2 Experiments settings

To explore the effectiveness of the proposed methods, i.e., CCE, SVDRD, NWET and N2WET, we conduct three series of experiments under various experimental circumstances.

The purpose of the first series of experiments is to check whether the term weighting factor with category contribution achieves improvements on accuracy. Since the CCE is aiming at improving the discriminating power for terms, the comparisons will be carried with the most of the state-of-the-art term weighting factors. Following with the works of Refs. [6,9–10], the term weighting scheme employed in text representation can be expressed by

$$W_{\text{weight}}(t,D) = W_{\text{importance}}(t,D)W_{\text{discriminator}}(t) \tag{6}$$

where the $W_{\text{importance}}(t,D)$ is the importance part corresponding to different features involved in the scheme, and the TF is being used in this experiment. $W_{\text{discriminator}}(t)$ is the weight of discriminator part, i.e., any one of the aforementioned term weighting factors. Then the experimented schemes are listed in Table 2.

Table 2 Summary of 11 term weighting schemes

No.	Importance method	Discriminator method	Denoted by
1	TF	IDF	TF-IDF
2	TF	IG	TF-IG
3	TF	GR	TF-GR
4	TF	OR	TF-OR
5	TF	χ^2	TF-CHI
6	TF	RF	TF-RF
7	TF	CCE_2	TF-CCE_2
8	TF	CCE_e	TF-CCE_e
9	1	RF	RF
10	1	CCE_2	CCE_2
11	1	CCE_e	CCE_e

The second series of experiments is to measure the benefit brought by the distributional coefficient based on both of the passage strategy and the CCE in SVDRD. While the SVDRD is working in the reduced dimension mode, we compare it with the baselines of TF-IDF and the Dino's [13] described in Table 3. In Table 3, we formulate the term weighting factor probality cdenoted by (PROB) by

$$W_{\text{PROB}}(t) = \max_{k\in[1,N_\text{c}]} \Pr(C_k \mid t) \tag{7}$$

Table 3 The reported combinations for SVDRD

No	Distributional coefficient	Weighting factor	Denoted by
1	DC_SUM	PROB	Dino's in Ref. [13]
2	DC_GI	PROB	DC_GI-PROB
3	DC_GLI	PROB	DC_GLI-PROB
4	DC_LL	PROB	DC_LL-PROB
5	DC_LVL	PROB	DC_LVL-PROB
6	DC_SUM	CCE_0	DC_SUM-CCE
7	DC_GI	CCE_0	DC_GI-CCE
8	DC_GLI	CCE_0	DC_GLI-CCE
9	DC_LL	CCE_0	DC_LL-CCE
10	DC_LVL	CCE_0	DC_LVL-CCE

Apparently, the Dino's is a special case of the SVDRD without employing the passage strategy. It is noteworthy that the Algorithm 1 of document partition is required for distributional coefficient. Following the suggestion of Callan [21], we set $L_\text{P} = 200$ for experiments. Since the original passages is preferred, a number of experiments show that by the proposed scheme is insensitive to the threshold λ, thus $\lambda = 0$ is employed for simplicity.

In the third series experiments, to further evaluate the effectiveness of the NWET and N2WET, we prefer the CCE$_2$ as the representative of the proposed term weighting factors to be involved. The experiments conducted here are similar with the work of Ref. [6], however, we implement the term weighting schemes based on passages while they preferred sentences. The TF-RF [9] which has been reported as one of the excellent term weighting schemes is chosen as the baseline in this experiments. The summarization of the experimented combinations is listed in Table 4. Here, CP is the compactness of the appearances of a word defined by Ref. [6].

Table 4 The summarization of the combinations for the third series experiments

No.	Group of Importance	Discriminator	Combinations
1	TF	RF, CCE$_2$	TF-RF, TF-CCE$_2$
2	CP	RF, CCE$_2$	3 CP features (CP$_{\mathrm{PN}}$, CP$_{\mathrm{FLD}}$, CP$_{\mathrm{PV}}$) × 2 (RF, CCE$_2$)
3	FA	RF, CCE$_2$	4 FA features (FA$_{\mathrm{GI}}$, FA$_{\mathrm{GLI}}$, FA$_{\mathrm{LL}}$, FA$_{\mathrm{LVL}}$) × 2 (RF, CCE$_2$)
4	TF+CP	RF, CCE$_2$	3 (combinations of TF and each CP feature) × 2 (RF, CCE$_2$)
5	TF+FA	RF, CCE$_2$	4 (combinations of TF and each FA feature) × 2 (RF, CCE$_2$)
6	CP+FA	RF, CCE$_2$	3×4 (combinations of one CP feature and one FA feature) × 2 (RF, CCE$_2$)
7	TF+CP+FA	RF, CCE$_2$	3×4 (combinations of TF, one CP feature and one FA feature) × 2 (RF, CCE$_2$)
8	NWET	RF, CCE$_2$	4 NWET features(DC$_{\mathrm{GI}}$, DC$_{\mathrm{GLI}}$, DC$_{\mathrm{LL}}$, DC$_{\mathrm{LVL}}$) × 2 (RF, CCE$_2$)
9	N2WET	RF, CCE$_2$	4 N2WET features(DC$_{\mathrm{GI}}$, DC$_{\mathrm{GLI}}$, DC$_{\mathrm{LL}}$, DC$_{\mathrm{LVL}}$) × 2 (RF, CCE$_2$)

Usually, precision, recall and F$_1$ are popular performance measures. Due to isolating from each other, neither precision and recall are run well to reflect the performance of text categorization in practice. Therefore we prefer to use F$_1$ measure to compute the averaged performance in two way: macro-averaging (maF$_1$) and micro-averaging (miF$_1$). maF$_1$ is computed as the arithmetic mean of category-specific measure over all target categories, while miF$_1$ is defined in terms of the micro-averaged values of precision p and recall γ [3].

$$\left.\begin{aligned} \mathrm{miF}_1 &= \frac{2pr}{p+r} \\ \mathrm{maF}_1 &= \frac{1}{N_{\mathrm{c}}}\sum_{k=1}^{N_{\mathrm{c}}}\mathrm{F}_{1k} \end{aligned}\right\} \tag{8}$$

The SVM classifier used here is the LibLinear 1.6 [29]. All the parameters are set by five-fold cross validations. Instances are normalized before being provided to the LibLinear.

3.3 Effectiveness of the CCE

Table 5 shows the result of SVM on four data sets using the term weighting schemes summarized in Table 2. Rank of each term weighting scheme highlighted by boldface with superscript is given depending on its performance measure followed by corresponding rank (from 1 to 3). It is apparent that in terms of both the miF$_1$ and maF$_1$ performance of the proposed CCE (with two variants) almost wins all the first two prizes except on the Reuters-21578. Despite this, the TF-CCE$_2$ outperforms the best reported term weighting scheme TF-RF. What can be observed from the last three rows is that the binary representations for both CCE$_2$ and CCE$_e$ achieve at least comparable performance to the RF. Therefore the CCE is indeed helpful for text representation.

Table 5 Results of the CCE and the state-of-the-art term weighing schemes on four corpora

Term weighting schemes	Reuters-21578		20Newsgroup		WebKB		Ohsumed	
	miF$_1$	maF$_1$	miF$_1$	maF$_1$	miF$_1$	maF$_1$	miF$_1$	maF$_1$
TF-IDF	0.8388	0.6326	0.7250	0.7242	0.8074	0.7847	0.5859	0.4992
TF-IG	0.8771	0.7017	0.7468	0.7479	0.9114	0.9013	0.6156	0.4755
TF-GR	0.8814	**0.7175^2**	0.7470	0.7481	0.9050	0.8946	0.6255	0.5277
TF-OR	0.8667	0.6917	0.7410	0.7406	0.8156	0.7933	0.6113	0.5104
TF-CHI	0.8818	**0.7186^1**	0.7515	0.7520	0.9071	0.8968	0.6306	0.5259
TF-RF	**0.8868^2**	0.7148	0.8137	0.8152	**0.9396^3**	**0.9273^3**	**0.6606^3**	**0.5464^3**
TF-CCE$_2$	**0.8871^1**	**0.7161^3**	**0.8168^2**	**0.8171^2**	**0.9418^2**	**0.9311^2**	**0.6702^2**	**0.5594^1**
TF-CCE$_e$	**0.8853^3**	0.7116	0.8153	0.8157	**0.9425^1**	**0.9325^1**	**0.6709^1**	**0.5589^2**
RF	0.8771	0.6901	0.8151	0.8162	0.9264	0.9132	0.6391	0.5372
CCE$_2$	0.8767	0.6898	**0.8183^1**	**0.8184^1**	0.9318	0.9193	0.6429	0.5274
CCE$_e$	0.8764	0.6915	**0.8163^3**	**0.8164^3**	0.9339	0.9220	0.6461	0.5335

3.4 Contributions from terms by class tendency

To check whether any benefit is brought by combining terms on class tendency (see Table 3), based on the variants of the proposed CCE in the SVDRD, all the nine groups of the results are listed in Table 6. The Dino's is used as the baseline, for which the miF$_1$ and maF$_1$ are reported. For other combinations, the gain of performance compared to the baseline is reported. Suppose the performance of the nth combination and the baseline are $p_{\text{perf}}(n)$ and $p_{\text{perf}}(\text{base})$, respectively, the gain is calculated as follows:

$$G_{\text{gain}}(n)=\frac{p_{\text{perf}}(n)-p_{\text{perf}}(\text{base})}{p_{\text{perf}}(\text{base})}\times 100\% \tag{9}$$

Table 6 Results of the SVDRD on four data sets

Gain/(%)	Reuters-21578		20Newsgroup		WebKB		Ohsumed	
	miF$_1$	maF$_1$	miF$_1$	maF$_1$	miF$_1$	maF$_1$	miF$_1$	maF$_1$
Dino's[13]	0.8696	0.6349	0.7952	0.7948	**0.8589**^{3}	0.8317	0.6372	0.5347
DC$_{\text{GI}}$-PROB	– 0.2530	0.3465	– 21.1645	– 21.1626	– 2.0026	– 0.8296	– 5.7910	– 3.6656
DC$_{\text{GLI}}$-PROB	**1.1155**^{1}	**3.6384**^{1}	**0.9809**^{2}	**0.9940**^{2}	**1.9560**^{1}	**2.5009**^{1}	1.3497	0.4488
DC$_{\text{LL}}$-PROB	– 0.2530	0.3465	– 21.1645	– 21.1626	– 1.9676	– 0.8296	– 5.8067	– 8.0606
DC$_{\text{LVL}}$-PROB	**0.8625**^{3}	**1.8428**^{3}	– 0.5533	– 0.5536	– 0.5472	– 0.1683	**1.5851**^{3}	0.2244
DC$_{\text{SUM}}$-CCE	0.4140	0.5828	**0.3521**^{3}	**0.3145**^{3}	**0**^{3}	**0.0481**^{3}	0.7533	**1.0847**^{3}
DC$_{\text{GI}}$-CCE	– 0.8625	– 0.5040	– 21.7807	– 21.7036	– 1.7464	– 0.3487	– 5.3045	– 7.4995
DC$_{\text{GLI}}$-CCE	**0.9430**^{2}	**2.6933**^{2}	**1.2953**^{1}	**1.2833**^{1}	**1.4903**^{2}	**1.9478**^{2}	**2.2128**^{1}	**2.6931**^{1}
DC$_{\text{LL}}$-CCE	– 0.8625	– 0.5040	– 21.7807	– 22.6220	– 1.7115	– 0.6733	– 5.3045	– 7.4995
DC$_{\text{LVL}}$-CCE	0.4945	0.3150	– 0.4024	– 0.4529	– 0.5821	– 0.4088	**1.7577**^{2}	**1.6832**^{2}

Obviously, the best performances are consistently obtained by either the DC$_{\text{GLI}}$-PROB or the DC$_{\text{GLI}}$-CCE. Taking Table 5 into account, comparable results with the SVDRD are achieved without reducing the dimension. Furthermore, one of the major reasons which inspired us to employ it in text categorization is its excellent performance in efficiency.

Fig.1 reports the time comparisons of TF-CCE$_2$ and DC$_{\text{GLI}}$-CCE which are tried while the maximum number of terms marked under each corpus name is selected. The exact time costs are shown above each histogram. Except for handling the Ohsumed, the time required by the DC$_{\text{GLI}}$-CCE are less than one-fifth of the TF-CCE$_2$. Obviously, it is the way of combining different terms on class tendency that reduces the dimensionality to the size of categories, usually much lower than that of the sparse vector of features in traditional method, which leads to this significant improvement. An interest funding is that the vector of the Ohsumed with TF-CCE$_2$ is much more sparser than the others since the time consumed by the SVDRD is close to half of the TF-CCE$_2$. In fact, this is an inherent characteristic of the medicine journal. Even so, compared with TF-CCE$_2$, the accuracy of DC$_{\text{GLI}}$-CCE reaches 87.88% to 88.71% for Reuters-21578, 80.55% to 81.68% for 20Newsgroup, 87.17% to 94.18% for WebKB and 65.13% to 67.02% for the Ohsumed (see Tables 5 and 6). The achieved performance shows its comparability of the SVDRD in compared with the high performance CCE, especially for the large-scale data.

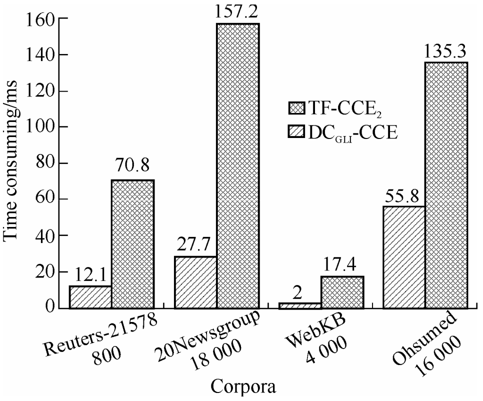

Fig 1 The comparisons of TF-CCE$_2$ and DC$_{\text{GLI}}$-CCE in time consuming

3.5 Contributions from the weighted combination of term in passages

Table 7 summarizes the performances obtained by the combinations reported in Table 4. Since the TF-RF which is one of the best term weighting schemes outperforms the TF-IDF significantly [10], almost all of the combinations fail to improve the baseline on Reuters-21578. However, while compared with the TF-IDF (see Table 5), we get contradictory results. In spite of different term weighting factors are evaluated, this phenomenon is similar to the results of Ref. [6] because documents with multiple labels are common for Reuters-21578. Meanwhile, in the case, a better performance reached by TF-FA$_{\text{GI}}$-RF indicates that almost all of the important information are locating at the head or the beginning of documents. Therefore, the combination of distributional features and either RF or CCE_2 is not recommended in text categorization. It is worth stressing that most of the combinations outperform the TF-IDF significantly on the rest three corpora. In addition, several interesting observations can gathered from Table 7.

Table 7 Results of the combinations of the distributional features in Ref. [6], NWET, N2WET and CCE, RF on the corpora (passages)

Gain/(%)	Reuters-21578		20Newsgroup		WebKB		Ohsumed	
	miF_1	maF_1	miF_1	maF_1	miF_1	maF_1	miF_1	maF_1
TF-RF	0.886 8	0.714 8	0.813 7	0.815 2	0.939 6	0.927 3	0.660 6	0.546 4
CP_{FLD}-RF	– 0.530 0	– 1.734 8	– 1.523 9	– 1.447 5	– 1.521 9	– 1.585 2	– 3.209 2	– 2.635 4
CP_{FLD}-CCE_2	– 0.563 8	– 1.734 8	– 1.142 9	– 1.079 5	– 1.670 9	– 1.811 7	– 1.861 9	– 0.128 1
CP_{PN}-RF	– 0.203 0	– 1.455 0	– 0.331 8	– 0.294 4	– 1.788 0	– 2.027 4	– 1.044 5	– 1.940 0
CP_{PN}-CCE_2	– 0.248 1	– 1.287 1	0.110 6	0.024 5	– 1.255 9	– 1.380 4	0.514 7^{++}	2.580 5^{++}
CP_{PV}-RF	– 2.345 5	– 4.630 7	– 5.087 9	– 5.078 5	– 3.341 8	– 4.054 8	– 9.521 6	– 14.549 8
CP_{PV}-CCE_2	– 2.266 6	– 4.280 9	– 4.227 6	– 4.219 8	– 3.458 9	– 4.227 3	– 9.128 1	– 11.237 2
FA_{GI}-RF	– 0.281 9	– 1.510 9	0.700 5^{++}	0.674 7^{++}	0.234 1	0.442 1*	0.272 5	0.494 1*
FA_{GI}-CCE_2	– 0.327 0	– 1.427 0	0.897 1^{++}	0.821 9^{++}	0.266 1	0.603 9^{++}	1.650 0^{++}	4.923 1^{++}
FA_{GLI}-RF	– 0.563 8	– 2.336 3	**1.179 8**^{3}	1.042 7^{++}	0.308 6*	0.420 6	0.787 2^{++}	1.647 1^{++}
FA_{GLI}-CCE_2	– 0.530 0	– 4.308 9	**1.573 1**^{1}	**1.423 0**^{1}	0.340 6*	0.571 6^{++}	**2.300 9**^{2}	5.838 2^{++}
FA_{LL}-RF	– 0.687 9	– 3.147 7	0.467 0*	0.368 0*	– 0.947 2	– 0.992 1	– 1.544 1	– 1.573 9
FA_{LL}-CCE_2	– 0.654 0	– 3.203 7	0.589 9^{++}	0.466 1*	– 0.798 2	– 0.711 7	0.378 4*	2.269 4^{++}
FA_{LVL}-RF	– 0.563 8	– 2.112 5	0.479 3*	0.453 9*	0.074 5	0.248 0	– 1.619 7	– 2.983 2
FA_{LVL}-CCE_2	– 0.733 0	– 1.902 6	0.934 0^{++}	0.785 1^{++}	0.308 6*	0.539 2^{++}	0.242 2	0.970 0^{++}
TF-CP_{FLD}-RF	– 0.969 8	– 3.763 3	– 4.645 4	– 4.551 0	– 1.745 4	– 1.800 9	– 5.146 8	– 6.259 2
TF-CP_{FLD}-CCE_2	– 0.845 7	– 2.224 4	– 3.981 8	– 3.888 6	– 2.128 6	– 2.049 0	– 4.072 1	– 3.385 8
TF-CP_{PN}-RF	– 0.733 0	– 2.867 9	– 3.121 5	– 3.079 0	– 1.628 4	– 1.876 4	– 4.541 3	– 4.941 4
TF-CP_{PN}-CCE_2	– 0.890 8	– 2.784 0	– 2.543 9	– 2.465 7	– 1.596 4	– 1.703 9	– 3.345 4	– 2.745 2
TF-CP_{PV}-RF	– 1.014 9	– 1.916 6	– 3.723 7	– 3.667 8	– 5.811 0	– 6.524 3	– 5.646 4	– 9.919 5
TF-CP_{PV}-CCE_2	– 1.014 9	– 1.678 8	– 2.998 6	– 2.993 1	– 5.395 9	– 5.952 8	– 3.421 1	– 5.142 8
TF-FA_{GI}-RF	**0.203 0**^{1}	**0.097 9**^{3}	– 0.073 7	– 0.085 9	0.840 8^{++}	1.294 1^{++}	– 0.393 6	– 0.457 5
TF-FA_{GI}-CCE_2	0.000 0	– 0.251 8	0.258 1	0.196 3	0.840 8^{++}	1.337 2^{++}	0.847 7^{++}	4.227 7^{++}
TF-FA_{GLI}-RF	0.033 8	– 0.153 9	0.774 2^{++}	0.809 6^{++}	0.840 8^{++}	1.164 7^{++}	1.089 9^{++}	3.294 3^{++}
TF-FA_{GLI}-CCE_2	0.000 0	– 0.251 8	0.934 0^{++}	0.969 1^{++}	**0.957 9**^{3}	**1.445 1**^{3}	**2.179 8**^{3}	**5.856 5**^{3}
TF-FA_{LL}-RF	**0.078 9**^{2}	**0.125 9**^{2}	– 0.098 3	– 0.036 8	– 0.266 1	– 0.215 7	0.772 0^{++}	2.214 5^{++}
TF-FA_{LL}-CCE_2	0.033 8	– 0.685 5	0.405 6*	0.441 6*	– 0.117 1	0.075 5	2.134 4^{++}	5.819 9^{++}
TF-FA_{LVL}-RF	0	– 0.139 9	0.331 8*	0.368 0*	0.766 3^{++}	0.938 2^{++}	0.560 1^{++}	0.237 9
TF-FA_{LVL}-CCE_2	– 0.078 9	– 0.405 7	0.503 9^{++}	0.515 2^{++}	0.840 8^{++}	1.132 3^{++}	1.498 6^{++}	3.623 7^{++}
CP_{PN}+FA_{GI}-RF	– 0.484 9	– 1.846 7	– 1.179 8	– 1.226 7	– 0.383 1	– 0.312 7	– 3.648 2	– 4.831 6
CP_{PN}+FA_{GI}-CCE_2	– 0.608 9	– 2.126 5	– 0.577 6	– 0.637 9	– 0.298 0	– 0.086 3	– 2.240 4	– 0.219 6
CP_{PN}+FA_{GLI}-RF	– 0.281 9	– 1.832 7	– 0.049 2	– 0.134 9	– 0.532 1	– 0.463 7	– 1.665 2	– 2.068 1
CP_{PN}+FA_{GLI}-CCE_2	– 0.281 9	– 1.105 2	0.245 8	0.294 4	– 0.340 6	– 0.194 1	– 0.514 7	1.921 7^{++}
CP_{PN}+FA_{LL}-RF	– 0.248 1	– 1.427 0	– 0.442 4	– 0.502 9	– 1.479 4	– 1.639 2	– 1.498 6	– 0.787 0
CP_{PN}+FA_{LL}-CCE_2	– 0.484 9	– 1.455 0	– 0.147 5	– 0.122 7	– 1.213 3	– 1.304 9	– 0.363 3	2.434 1^{++}
CP_{PN}+FA_{LVL}-RF	– 0.327 0	– 1.804 7	– 0.307 2	– 0.368 0	0	0.097 1	– 1.998 2	– 2.964 9
CP_{PN}+FA_{LVL}-CCE_2	– 0.484 9	– 1.678 8	0.135 2	0.012 3	– 0.191 6	– 0.053 9	– 0.620 6	2.269 4^{++}
CP_{FLD}+FA_{GI}-RF	– 0.733 0	– 2.070 5	– 3.244 4	– 3.152 6	– 1.447 4	– 1.229 4	– 5.449 6	– 6.112 7
CP_{FLD}+FA_{GI}-CCE_2	– 1.014 9	– 2.406 3	– 2.298 1	– 2.343 0	– 1.064 3	– 0.830 4	– 4.026 6	– 3.459 0
CP_{FLD}+FA_{GLI}-RF	– 0.563 8	– 1.832 7	– 1.880 3	– 1.791 0	– 0.947 2	– 0.733 3	– 3.663 3	– 4.355 8
CP_{FLD}+FA_{GLI}-CCE_2	– 0.654 0	– 1.441 0	– 1.253 5	– 1.165 4	– 0.872 7	– 0.711 7	– 2.785 3	– 2.360 9
CP_{FLD}+FA_{LL}-RF	– 0.608 9	– 1.594 9	– 1.868 0	– 1.778 7	– 1.713 5	– 1.736 2	– 3.557 4	– 4.063 0
CP_{FLD}+FA_{LL}-CCE_2	– 0.484 9	– 1.455 0	– 1.192 1	– 1.104 0	– 1.628 4	– 1.671 5	– 2.361 5	– 1.207 9
CP_{FLD}+FA_{LVL}-RF	– 0.530 0	– 1.692 8	– 2.015 5	– 1.913 6	– 0.681 1	– 0.550 0	– 3.981 2	– 5.929 7
CP_{FLD}+FA_{LVL}-CCE_2	– 0.733 0	– 1.720 8	– 1.265 8	– 1.189 9	– 0.830 1	– 0.690 2	– 2.845 9	– 2.836 7
CP_{PV}+FA_{GI}-RF	– 0.327 0	– 1.133 2	0.565 3^{++}	0.453 9*	– 0.457 6	– 0.021 6	– 0.787 2	– 1.464 1
CP_{PV}+FA_{GI}-CCE_2	– 0.406 0	– 1.608 8	0.835 7^{++}	0.821 9^{++}	– 0.223 5	0.194 1	0.454 1*	2.434 1^{++}
CP_{PV}+FA_{GLI}-RF	– 0.969 8	– 2.644 1	– 0.749 7	– 0.748 3	– 1.064 3	– 0.959 8	– 3.148 7	– 5.600 3
CP_{PV}+FA_{GLI}-CCE_2	– 1.048 7	– 2.853 9	– 0.331 8	– 0.466 1	– 1.096 2	– 1.078 4	– 1.150 5	– 0.658 9
CP_{PV}+FA_{LL}-RF	– 1.172 8	– 2.798 0	– 0.786 5	– 0.785 1	– 1.937 0	– 2.167 6	– 4.162 9	– 5.856 5
CP_{PV}+FA_{LL}-CCE_2	– 1.048 7	– 2.784 0	– 0.417 8	– 0.539 7	– 1.670 9	– 1.962 7	– 2.437 2	– 2.873 4
CP_{PV}+FA_{LVL}-RF	– 1.657 6	– 4.155 0	– 4.965 0	– 4.968 1	– 2.958 7	– 3.375 4	– 5.888 6	– 5.929 7
CP_{PV}+FA_{LVL}-CCE_2	– 1.499 8	– 3.819 3	– 4.043 3	– 4.023 6	– 2.884 2	– 3.256 8	– 5.843 2	– 5.472 2
TF-CP_{PN}+FA_{GI}-RF	– 1.014 9	– 2.923 9	– 5.616 3	– 5.642 8	– 1.670 9	– 1.445 1	– 8.113 8	– 10.779 6
TF-CP_{PN}+FA_{GI}-CCE_2	– 1.014 9	– 2.630 1	– 4.756 1	– 4.771 8	– 1.447 4	– 1.304 9	– 6.978 5	– 7.137 6
TF-CP_{PN}+FA_{GLI}-RF	– 0.845 7	– 2.406 3	– 4.289 1	– 4.183 0	– 1.330 4	– 1.369 6	– 6.479 0	– 8.180 8
TF-CP_{PN}+FA_{GLI}-CCE_2	– 0.935 9	– 2.504 2	– 3.453 4	– 3.348 9	– 0.947 2	– 0.776 4	– 4.828 9	– 4.630 3

Continued

Gain/(%)	Reuters-21578		20Newsgroup		WebKB		Ohsumed	
	miF$_1$	maF$_1$	miF$_1$	maF$_1$	miF$_1$	maF$_1$	miF$_1$	maF$_1$
TF-RF	0.886 8	0.714 8	0.813 7	0.815 2	0.9396	0.927 3	0.660 6	0.5464
TF-CP$_{\text{PN}}$+FA$_{\text{LL}}$-RF	– 0.845 7	– 3.133 7	– 4.3013	– 4.195 3	– 1.979 6	– 2.113 7	– 5.585 8	– 7.302 3
TF-CP$_{\text{PN}}$+FA$_{\text{LL}}$-CCE$_2$	– 1.048 7	– 3.749 3	– 3.416 5	– 3.336 6	– 1.670 9	– 1.671 5	– 3.996 4	– 3.129 6
TF-CP$_{\text{PN}}$+FA$_{\text{LVL}}$-RF	– 0.845 7	– 2.658 1	– 4.215 3	– 4.121 7	– 1.255 9	– 1.261 7	– 6.660 6	– 7.064 4
TF-CP$_{\text{PN}}$+FA$_{\text{LVL}}$-CCE$_2$	– 0.935 9	– 2.895 9	– 3.367 3	– 3.299 8	– 0.798 2	– 0.603 9	– 4.632 2	– 4.410 7
TF-CP$_{\text{FLD}}$+FA$_{\text{GI}}$-RF	– 1.251 7	– 2.742 0	– 7.889 9	– 7.899 9	– 2.352 1	– 2.178 4	– 8.658 8	– 10.267 2
TF-CP$_{\text{FLD}}$+FA$_{\text{GI}}$-CCE$_2$	– 1.014 9	– 2.266 4	– 6.574 9	– 6.611 9	– 2.054 1	– 1.736 2	– 6.978 5	– 6.881 4
TF-CP$_{\text{FLD}}$+FA$_{\text{GLI}}$-RF	– 1.138 9	– 3.105 8	– 6.267 7	– 6.133 5	– 2.011 5	– 1.919 6	– 7.023 9	– 8.601 8
TF-CP$_{\text{FLD}}$+FA$_{\text{GLI}}$-CCE$_2$	– 1.093 8	– 2.951 9	– 5.026 4	– 4.906 8	– 1.745 4	– 1.552 9	– 5.510 1	– 5.453 9
TF-CP$_{\text{FLD}}$+FA$_{\text{LL}}$-RF	– 1.048 7	– 3.175 7	– 6.120 2	– 5.998 5	– 2.543 6	– 2.685 2	– 6.115 7	– 7.174 2
TF-CP$_{\text{FLD}}$+FA$_{\text{LL}}$-CCE$_2$	– 1.138 9	– 2.881 9	– 4.940 4	– 4.833 2	– 2.426 6	– 2.383 3	– 4.859 2	– 3.825 0
TF-CP$_{\text{FLD}}$+FA$_{\text{LVL}}$-RF	– 1.251 7	– 4.057 1	– 5.886 7	– 5.753 2	– 1.819 9	– 1.747 0	– 6.933 1	– 8.308 9
TF-CP$_{\text{FLD}}$+FA$_{\text{LVL}}$-CCE$_2$	– 1.093 8	– 2.770 0	– 4.792 9	– 4.698 2	– 1.937 0	– 1.714 7	– 5.449 6	– 5.014 6
TF-CP$_{\text{PV}}$+FA$_{\text{GI}}$-RF	– 0.078 9	– 0.615 6	– 1.204 4	– 1.263 5	– 4.023 0	– 3.634 2	– 2.528 0	– 3.843 3
TF-CP$_{\text{PV}}$+FA$_{\text{GI}}$-CCE$_2$	– 0.360 8	– 1.035 3	– 0.675 9	– 0.613 3	– 4.023 0	– 4.087 1	– 1.029 4	0.439 2
TF-CP$_{\text{PV}}$+FA$_{\text{GLI}}$-RF	– 0.530 0	– 1.482 9	– 1.376 4	– 1.324 8	– 4.406 1	– 4.227 3	– 1.665 2	– 1.262 8
TF-CP$_{\text{PV}}$+FA$_{\text{GLI}}$-CCE$_2$	– 0.608 9	– 1.343 0	– 0.970 9	– 1.055 0	– 4.257 1	– 4.421 4	– 0.363 3	0.164 7
TF-CP$_{\text{PV}}$+FA$_{\text{LL}}$-RF	– 0.608 9	– 1.748 7	– 1.511 6	– 1.447 5	– 4.565 8	– 4.831 2	– 1.725 7	– 3.806 7
TF-CP$_{\text{PV}}$+FA$_{\text{LL}}$-CCE$_2$	– 0.563 8	– 1.594 9	– 1.229 0	– 1.177 6	– 4.257 1	– 4.151 8	– 0.333 0	0.420 9*
TF-CP$_{\text{PV}}$+FA$_{\text{LVL}}$-RF	– 0.608 9	– 1.944 6	– 3.650 0	– 3.618 7	– 5.662 0	– 6.179 2	– 3.527 1	– 5.453 9
TF-CP$_{\text{PV}}$+FA$_{\text{LVL}}$-CCE$_2$	– 0.654 0	– 2.182 4	– 3.060 1	– 3.066 7	– 5.204 3	– 5.715 5	– 1.952 8	– 1.006 6
NWET$_{\text{GI}}$-RF	**0.063 8**^{3}	– 1.986 6	0.823 4^{++}	0.736 0^{++}	**1.138 8**^{1}	**1.477 4**^{2}	0.938 5^{++}	3.257 7^{++}
NWET$_{\text{GI}}$-CCE$_2$	– 0.045 1	– 0.433 7	0.958 6^{++}	0.969 1^{++}	**1.064 3**^{2}	**1.639 2**^{1}	**2.376 6**^{1}	**6.423 9**^{1}
NWET$_{\text{GLI}}$-RF	– 0.248 1	– 1.091 2	– 0.356 4	– -0.429 3	– 0.074 5	– 0.043 1	– 0.756 9	– 1.226 2
NWET$_{\text{GLI}}$-CCE$_2$	– 0.078 9	**0.405 7**^{1}	0.073 7	– 0.024 5	0.074 5	0.280 4	0.136 2	3.001 5^{++}
NWET$_{\text{LL}}$-RF	– 0.248 1	– 2.126 5	– 0.565 3	– 0.637 9	– 0.564 1	– 0.603 9	– 1.347 3	– 1.647 1
NWET$_{\text{LL}}$-CCE$_2$	– 0.203 0	– 0.825 4	– 0.331 8	– 0.404 8	– 0.031 9	0.064 7	– 0.408 7	1.390 9^{++}
NWET$_{\text{LVL}}$-RF	– 0.248 1	– 1.161 2	– 0.454 7	– 0.527 5	0.042 6	0.086 3	– 0.529 8	– 1.372 6
NWET$_{\text{LVL}}$-CCE$_2$	– 0.248 1	– 1.231 1	– 0.061 4	– 0.024 5	0.191 6	0.312 7*	0.590 4^{++}	3.587 1^{++}
N2WET$_{\text{GI}}$-RF	– 0.654 0	– 2.546 2	1.155 2^{++}	**1.128 6**^{2}	– 0.117 1	– 0.053 9	0.075 7	0.805 3^{++}
N2WET$_{\text{GI}}$-CCE$_2$	– 0.687 9	– 2.028 5	**1.241 2**^{2}	**1.079 5**^{3}	0.234 1	0.323 5*	1.498 6^{++}	**6.222 5**^{2}
N2WET$_{\text{GLI}}$-RF	– 0.281 9	– 1.371 0	0.995 5^{++}	0.969 1^{++}	– 0.383 1	– 0.377 4	0.726 6^{++}	0.860 2^{++}
N2WET$_{\text{GLI}}$-CCE$_2$	– 0.248 1	– 1.441 0	1.069 2^{++}	1.042 7^{++}	0.180 9	0.118 6	2.073 9^{++}	4.117 9^{++}
N2WET$_{\text{LL}}$-RF	– 0.248 1	– 1.468 9	0.786 5^{++}	0.772 8^{++}	– 1.064 3	– 1.164 7	0.499 5*	0.585 7^{++}
N2WET$_{\text{LL}}$-CCE$_2$	– 0.248 1	– 0.867 4	0.946 3^{++}	0.932 3^{++}	– 0.798 2	– 0.862 7	1.786 3^{++}	3.989 8^{++}
N2WET$_{\text{LVL}}$-RF	– 0.451 1	– 1.958 6	0.712 8^{++}	0.686 9^{++}	– 0.340 6	– 0.399 0	0.196 8	– 0.329 4
N2WET$_{\text{LVL}}$-CCE$_2$	– 0.406 0	– 1.678 8	0.934 0^{++}	0.760 5^{++}	– 0.074 5	– -0.043 1	1.347 3^{++}	2.653 7^{++}

Note: ++ and * denote the performance is significantly better than the baseline at 0.5 and 0.3 significance level, respectively. The positive gains without markers denote the comparable performance. The gains of rank 1 to 3 are highlighted by bold font with superscripts.

1) The combinations of NWET or N2WET features and CCE$_2$ perform better than the combinations of any other distributional feature [6] and RF frequently, even though some of these performances are unstable.

2) Actually, some of the combinations perform stably, i.e., TF-FA$_{\text{GLI}}$-RF, TF-FA$_{\text{GLI}}$-CCE$_2$, TF-FA$_{\text{GI}}$-CCE$_2$, NWET$_{\text{GI}}$-RF, NWET$_{\text{GI}}$-CCE$_2$, NWET$_{\text{GLI}}$-CCE$_2$, N2WET$_{\text{GI}}$-CCE$_2$ and N2WET$_{\text{GLI}}$-CCE$_2$. Compared with TF-RF, without considering Reuters-21578, their gains range from 0.2341% to 6.4239%. More importantly, the effectiveness of both DC$_{\text{GI}}$ and DC$_{\text{GLI}}$ indicate that, for the evaluated corpora, the location of valuable information is usually close to the front of a document. Furthermore, together with Tables 4–7, the proposed CCE with its variants consistently outperform others schemes.

3) Although so many combinations are reported in Table 7, for the practical use, the NWET and N2WET are the first two features preferred for text categorization. But the combination of CP features and FA features, no matter whether combining with TF features or not, is not recommended.

4 Conclusions

In text categorization, the previous studies rely on the appearance or frequency to characterize a term for text representation. However, in three types of scenarios discussed in Sect. 1, these methods can hardly capture enough information and result in ineffective performance. Therefore, this paper pays special attention to explore several novel ways of integrating the contributions from different terms with respect to class tendency and the contributions for each term with different locations in a document. As a supplementary method, in this study, a self-adaptive strategy for document partition and a group of novel term weighting factors namely CCE are proposed. According to the three series of experiments, we find that the SVDRD indeed significantly reduce the computational cost and reach comparable performances in comparison with the state-of-the-art methods while, for most cases, the method of combining the contributions with respect to each appearance with different locations for a term achieves the best performance. Apparently, an appropriate way of capturing the document structure is at least as important as feature selection and classifiers for the further improvement on text categorization.

Acknowledgements

The authors would like to thank Lan M. for valuable suggestions and the handling editor and anonymous reviewers for helping to greatly improve the paper. This work was supported by the Hi-Tech Research and Development Program of China (2009AA01Z430), the National Natural Science Foundation of China (60972077, 60821001), the National S&T Major Program (2010ZX03003-003-01), the Fundamental Research Funds for the Central Universities (BUPT2011RC0210), and the Science and Technology on Electronic Control Laboratory.

References

1. Chen J, Huang H, Tian S, et al. Feature selection for text classification with naive bayes. Expert Systems with Applications, 2009, 36(3): 5432–5435
2. Tan S. Neighbor-weighted *k*-nearest neighbor for unbalanced text corpus. Expert Systems with Applications, 2005, 28(4): 667–671
3. Joachims T. Learning to classify text using support vector machines: Methods, Theory and Algorithms. Dordrecht, Netherlands: Kluwer Academic Publishers, 2002
4. Leopold E, Kindermann J. Text categorization with support vector machines. how to represent texts in input space?. Machine Learning, 2002, 46(1/2/3): 423–444
5. Xue H, Chen S C, Yang Q. Structural regularized support vector machine: a framework for structural large margin classifier. IEEE Transactions on Neural Networks, 2011, 22(4): 573–587
6. Xue X B, Zhou Z H. Distributional features for text categorization. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(3): 428–442
7. Qi X G, Davison B D. Web page classification: features and algorithms. ACM Computing Surveys, 2009, 41(2): 1–31
8. Lan M, Tan C, Low H, et al. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. Proceedings of the 14th International Conference on World Wide Web (WWW’05), May 10–14, 2005, Chiba, Japan. New York, NY, USA: ACM, 2005: 1032–1033
9. Lan M, Tan C L, Su J et al. Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 721–735
10. Altınçay H, Erenel Z. Analytical evaluation of term weighting schemes for text categorization. Pattern Recognition Letters, 2010, 31(11): 1310–1323
11. Quan X J, Liu W Y, Qiu B T. Term weighting schemes for question categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 1009–1021
12. Isa D, Lee L H, Kallimani V. A polychotomizer for case-based reasoning beyond the traditional bayesian classification approach. Journal of Computer and Information Science, 2008, 1(1): 57–68
13. Isa D, Lee L H, Kallimani V P, et al. Text document preprocessing with the bayes formula for classification using the support vector machine. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(9): 1264–1272
14. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 1988, 24(5): 513–523
15. Joachims T. Text categorization with support vector machines: learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning (ECML’98), Apr 21–24, 1998, Chemnitz, Germany. LNCS 1398. Berlin, Germany: Springer-Verlag, 1998: 137–142
16. Debole F, Sebastiani F. Supervised term weighting for automated text categorization. Proceedings of the 20th Annual ACM Symposium on Applied Computing (SAC’03), Mar 9–12, 2003, Melbourne, FL, USA. New York, NY, USA: ACM, 2003: 784–788
17. Ping Y, Zhou Y J, Yang Y X, et al. A novel term weighting scheme with distributional coefficient for text categorization with support vector machine. Proceedings of the IEEE 2nd Youth Conference on Information, Computing and Telecommunications (YCICT’10), Nov 28–30, 2010, Beijing, China. Piscataway, NJ, USA: IEEE, 2010: 182–185
18. Ko Y, Park J, Seo J. Improving text categorization using the importance of sentences. Information Processing and Management, 2004, 40(1): 65–79
19. Kim J, Kim M J. An evaluation of passage-based text categorization. Journal of Intelligent Information Systems, 2004, 23(1): 47–65
20. Tseng C Y, Sung P C, Chen M S. Cosdes: a collaborative spam detection system with a novel e-mail abstraction scheme. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 669–682
21. Callan J P. Passage retrieval evidence in document retrieval. Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR’94), Jul 3–6, 1994, Dublin, Ireland. New York, NY, USA: ACM, 1994: 302–310
22. Lertnattee V, Theeramunkong T. Effect of term distributions on centroid-based text categorization. Information Sciences, 2004, 158(1): 89–115
23. Guan H, Zhou J Y, Guo M Y. A class-feature-centroid classifier for text categorization. Proceedings of the 18th International Conference on World Wide Web (WWW’09), Apr 20–24, 2009, Madrid, Spain. New York, NY, USA: ACM, 2009: 201–210
24. Soucy P, Mineau G W. Beyond TFIDF weighting for text categorization in the vector space model. Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI’05), Jul 30–Aug 5, Edinburgh, UK. Menlo Park, CA, USA: AAAI Press, 2005: 1130–1135
25. Isa D, Kallimani V P, Lee L H. Using the self organizing map for clustering of text documents. Expert Systems with Applications, 2009, 36(5): 9584–9591
26. Porter M. An algorithm for suffix stripping. Program, 1980, 14(3): 130–137
27. Lang K. NewsWeeder: learning to filter netnews. Proceedings of the 12th International Conference on Machine Learning (ICML’95), Jul 9–12, 1995, Tahoe City, CA, USA. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1995: 331–339
28. Graven M, DiPasquo D, Freitag D, et al. Learning to extract symbolic knowledge from the World Wide Web. Proceedings of the 15th National Conference for Artificial Intelligence (AAAI’98), Jul 26–30, 1998, Madison, WI, USA. Cambridge, MA, USA: MIT Press, 1998: 509–516
29. Fan R E, Chang K W, Hsieh C J, et al. Liblinear: a library for large linear classification. Journal of Machine Learning Research, 2008, 9: 1871–1874

(Editor: WANG Xu-ying)