Expert Systems with Applications 38 (2011) 12708–12716

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A semantic term weighting scheme for text categorization

Qiming Luo [a,b,*], Enhong Chen [a,b], Hui Xiong [c]

[a] School of Computer Science and Technology, University of Science and Technology of China, China
[b] MOE-MS Key Laboratory of Multimedia Computing and Communication of USTC, China
[c] Management Science and Information Systems Department, Rutgers University, USA

ARTICLE INFO

Keywords:
Text categorization
Semantic term weighting
WordNet
TF-IDF

ABSTRACT

Traditional term weighting schemes in text categorization, such as TF-IDF, only exploit the statistical information of terms in documents. Instead, in this paper, we propose a novel term weighting scheme by exploiting the semantics of categories and indexing terms. Specifically, the semantics of categories are represented by senses of terms appearing in the category labels as well as the interpretation of them by WordNet. Also, the weight of a term is correlated to its semantic similarity with a category. Experimental results on three commonly used data sets show that the proposed approach outperforms TF-IDF in the cases that the amount of training data is small or the content of documents is focused on well-defined categories. In addition, the proposed approach compares favorably with two previous studies.

1. Introduction

Automatic text categorization refers to the task of assigning documents to pre-defined categories by computer algorithms Sebastiani (2002). Text categorization is an important research field and is attracting increasing attention due to the growing availability of text documents in electronic forms.

A general process of text categorization is to induce a classifier based on a set of training documents with manually assigned category labels, and then apply the classifier to predict labels for uncategorized documents. The common paradigm of representing a document is the vector space model. Specifically, each document is transformed into a feature vector, where each feature refers to a term occurring in the document and the feature value corresponds to its weight. Different approaches have been introduced for term weighting. These approaches vary in terms of the definition of a term and the determination of term weights. The usual bag-of-words (BOW) approach treats each word as a feature and considers the features independent of each other. Such an approach ignores the syntactic and semantic information in a document, such as word order, multi-word phrases, synonymy, polysemy, and other semantic relationships among words.

Many previous studies are based on the two assumptions Sebastiani (2002), namely no meanings for categories are available and no exogenous knowledge is available. Concerning the second assumption, some recent works have explored strategies to exploit semantic ontologies such as WordNet,[1] Open Directory Project,[2] and Wikipedia[3] to semantically expand the indexing terms. However, approaches concerning the first assumption have received relatively little attention.

In this study, we propose a novel term weighting scheme by exploiting the semantics of categories and indexing terms. We employ WordNet as the dictionary for assigning senses to terms appearing in category labels as well as in documents. Specifically, our term weighting approach can be implemented by the following steps:

1. Determining the semantics of categories based on terms appearing in category labels as well as the interpretations of these terms by WordNet.
2. For each category, estimating the semantic similarity of each term with the category.
3. For each category, combining the semantic similarity of each term with the category and its term frequency in a document to obtain the feature vector of each document.

We make the following contributions in this paper.

1. We propose an automatic approach to determine the semantics of categories based on semantic ontologies.

* Corresponding author at: School of Computer Science and Technology, University of Science and Technology of China, China.
E-mail address: luoq@ustc.edu.cn (Q. Luo).

[1] http://wordnet.princeton.edu/.
[2] http://www.dmoz.org/.
[3] http://www.wikipedia.org/.

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.058

2. We propose a novel semantic term weighting scheme that outperforms TF-IDF in the cases that the amount of training data is small or the content of documents are focused on well defined categories.
3. The proposed approach is an alternative way to exploit semantic ontologies in text categorization, in contrast to existing approaches that semantically expand the indexing terms.

The second contribution is significant since manually assigning labels to documents by experts is a time-consuming and labor-intensive process. Together with the rapid growth of the amount of electronic documents, the amount of labeled training data becomes relatively scarce.

The remainder of the paper is organized as follows. Section 2 presents related works on term-weighting schemes and knowledge-oriented text categorization. In Section 3, we elaborate the proposed semantic term weighting scheme. Section 4 presents the performance evaluation results. Finally, we conclude the paper in Section 5.

2. Related work

Most term weighting schemes in text categorization are based on the statistical information of terms in documents. In a comparative study, Lan, Tan, Su, and Lu (2009) grouped term weighting schemes into two categories: supervised and unsupervised, which differ in whether the category labels of documents are exploited (supervised) or not (unsupervised). They proposed a new supervised scheme. Barak, Dagan, and Shnarch (2009) proposed term weighting schemes based on the distribution of a word in the document. Liu, Loh, and Sun (2009) proposed a probability based term weighting scheme to improve the performance for imbalanced text classification. Chang, Chen, and Liau (2008) proposed a new method to compute the relevance score of each term with respect to each category.

An alterative direction is to exploit WordNet for building term weighting schemes. WordNet is a lexical database for the English language created and maintained at the Cognitive Science Laboratory of Princeton University Miller, Beckwith, Fellbaum, Gross, and Miller (1990). In WordNet, English words are grouped into sets of synonyms called synsets. In addition, WordNet provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose of WordNet is to produce a combination of dictionary and thesaurus that is more intuitively usable and to support automatic text analysis and AI applications.

Some approaches based on exploiting the semantic relationships in WordNet to enrich document representation have been proposed. Scott and Matwin (1998) proposed a representation of documents based on hypernym density. Specifically, a list of all synonyms and hypernym synsets for all nouns and verbs is constructed based on WordNet, and the density of each synset is computed to form a feature vector for each document. The Ripper algorithm is applied at different levels of hypernym generalization. Results show that hypernym density outperforms BOW on data sets with narrowly defined and semantically distant classes, and underperforms BOW on data sets with opposite characteristics. Bloehdorn and Hotho (2004) proposed a hybrid approach for document representation based on the common term stem representation which is enhanced with concepts extracted from WordNet. Evaluation experiments using the AdaBoost algorithm on three text corpora (Reuters-21578, OHSUMED and FAODOC collections) showed consistent improvements over term-based representation. Bloehdorn, Basili, Cammisa, and Moschitti (2006) proposed semantic smoothing kernels for text classification, which implicitly encode a super-concept expansion in a semantic network. Several weighting schemes for the super-concepts are explored using well-known measures of term similarity computed from WordNet. The experimental evaluation on two data sets indicates that the approach consistently improves performance in situations of little training data and data sparseness. Mansuy and Hilderman (2006) performed empirical studies and showed that incorporating WordNet features, utilizing part of speech tags during WordNet expansion, and term weighting schemes have no statistically significant positive effect on the accuracy of the Naive Bayes and SVM classifiers.

Similar approaches to exploit other semantic ontologies have also been proposed. Wang and Domeniconi (2008) proposed to embed background knowledge derived from Wikipedia into a semantic kernel, which is then used to enrich the representation of documents. The experimental evaluation on real data sets demonstrates that the approach successfully achieves improved classification accuracy with respect to the BOW and other methods. Gabrilovich and Markovitch (2007) proposed an approach that automatically maps documents into appropriate concepts from external knowledge repositories to augment the bag of words representation. Specifically, words are mapped to categories in the Open Directory Project based on local contexts. Web crawling is applied to further increase the amount of knowledge. Experimental results over a range of data sets confirm improved performance compared to the bag of words document representation. However, the approach is computationally intensive. For instance, it took nearly 3 days to build the feature generator as the cumulative one-time expenditure.

In contrast, exploiting the semantics of categories has received relatively little attention. Li, Zhao, and Liu (2009) proposed to automatically construct training data by using the knowledge of the category name and WordNet, and showed that the best performance can achieve more than 90% of the baseline SVM classifiers in F1 measure. Barak et al. (2009) proposed to improve the bootstrapping scheme by combining LSA-based similarity with WordNet-based similarity, and showed improvement for Reuter10 but not for 20 Newsgroups.

de Buenaga Rodriguez, Gmez-Hidalgo, and Diaz-Agudo (1997) incorporated WordNet synonyms of the category label into the category model. Specifically, WordNet synsets close in meaning to category labels were manually selected. Terms appearing in these synsets are filtered by a stop list and then stemmed. And terms which do not appear in any training documents are deleted. For each remaining term, a degree of semantic closeness to the category it comes from is heuristically determined. The values of semantic closeness have been taken as the initial weights for categories when applying Rocchio and WidrowHoff algorithms to the training data. Results show that average precision achieves an improvement of 20 points for both algorithms. Although the idea is novel and results are significant, the key processes of the approach, e.g., manual sense disambiguation for terms in category labels, and determining semantic closeness to the category, are ad hoc in nature. In addition, it is only tested on one data set and the classifier algorithms are not among the state-of-the-art. A systematic and practical approach for exploiting the semantics of categories and terms for text categorization remains to be developed.

3. Proposed approach: semantic term weighting

The common way of term weighting is TF-IDF (Salton & Buckley, 1988), as defined in Eq. (1):

$$TF - IDF(t_i, d_j) = count(t_i, d_j) \times log \frac{|corpus|}{count\_doc(t_i, corpus)} \quad (1)$$

where $count(t_i, d_j)$ refers to the frequency of term t_i in document d_j, also known as term frequency (tf); $|corpus|$ refers to the number of

documents in the corpus; $count\_doc(t_i, corpus)$ refers to the number of documents in the corpus that contain the term t_i. TF-IDF was proposed in the information retrieval field. It was based on the intuition that the importance of a term to a document is dependent on its frequency as well as the degree of rareness at the document level in the corpus.

In this study we propose a term weighting scheme where the weight of each term is dependent on its semantic similarity to the category. This entails two required steps: determining the semantics of categories, and computing the semantic similarity of each term to categories.

3.1. Determining the semantics of categories

We assume that the semantics of each category is determined by the senses of a collection of words appearing in the category label as well as the interpretations of them by WordNet. Therefore, it is important to determine the sense of each word in the collection by word sense disambiguation (WSD).

To perform WSD, we apply the general disambiguation framework based on optimization principle stated in Navigli (2009). For a target word w with a set of senses $S^w = \left\{s_1^w, \ldots, s_{ns(w)}^w\right\}$, the most likely sense of w appearing in a context $CW = \{w_1, \ldots, w_n\}$ is determined by:

$$Sense(w) = \arg \max_{1 \leqslant i \leqslant ns(w)} \sum_{w_j \in CW} \max_{1 \leqslant k \leqslant ns(w_j)} Sim\left(s_i^w, s_k^{w_j}\right) \tag{2}$$

where $ns(\cdot)$ refers to the number of senses of a word, and $Sim(\cdot,\cdot)$ refers to the similarity measure of two senses.

The context for the collection of words of a category C is defined to be a set of words ranked among top K ($K = max\{100, N_C\}$ in this study) by χ^2 (CHI) feature selection measure for the category, where N_C refers to the number of unique words in category C. The rational is that these words are most effective in discriminating documents with respect to the category. According to Yang and Pedersen (1997), CHI is one of the most effective feature selection methods, and CHI values are comparable across terms for the same category.

We choose Lin's similarity measure Lin (1998) based on its theoretical foundation and its superior performance in the comparative study Bloehdorn et al. (2006):

$$Sim^{Lin}(s_1, s_2) = \frac{2\log(p(LCA(s_1, s_2)))}{\log(p(s_1)) + \log(p(s_2))} \tag{3}$$

where $LCA(s_1, s_2)$ refers to the lowest common ancestor of senses s_1 and s_2 in the hierarchy of senses. $Sim^{Lin}(s_1, s_2)$ ranges between 0 and 1.

In practice, the general disambiguation framework suffers from the problem that the sum of similarity measures from a large number of irrelevant or marginally relevant words would overwhelm the sum of a few highly relevant words, so a thresholding strategy is introduced.

$$Sim(s_1, s_2) = T\left(Sim^{Lin}(s_1, s_2), \delta_{Lin}, 0\right) \tag{4}$$

where

$$T(x, \delta, \gamma) = \begin{cases} x & \text{if } x > \delta \\ \gamma & \text{if } x \leqslant \delta \end{cases}$$

δ_{Lin} is the threshold for Lin's similarity measure, and it is empirically determined to be 0.82 by optimizing the accuracy of identifying senses of words in the category labels of Reuters-21578 as shown in Fig. 1. Here the correct senses of all words in the category labels have been manually determined by consulting the documents of each category and the available senses in WordNet. Since the range of $Sim^{Lin}(s_1, s_2)$ is [0,1], the threshold value of 0.82 selects only highly relevant words. The Reuters-21578 data sets is written by professional reporters on a large number of categories (115), the determined value can be considered as representative of text data and it is fixed for data sets in this study.

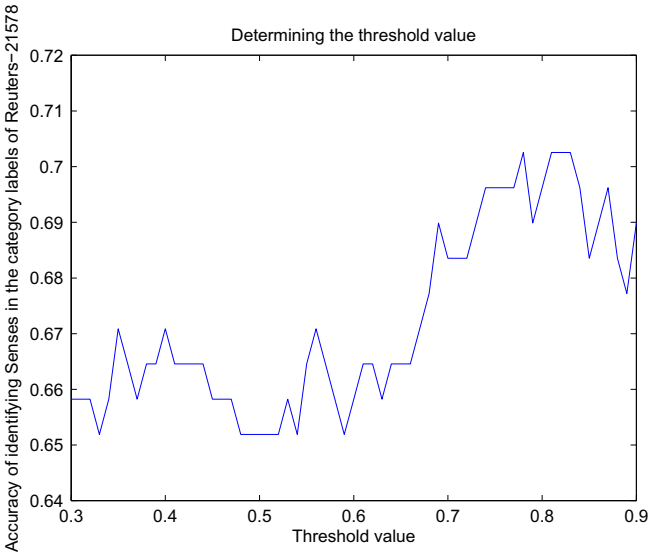

Fig. 1. Determining the threshold value for WSD of category labels.

For each category C, we define W_L^C as the set of words appearing in the label of C:

$$W_L^C = \{w | w \in label\ of\ C\} \tag{5}$$

and W_N^C as the set of words in the WordNet interpretation of each word in W_L^C:

$$W_N^C = \{v | \exists w \in W_L^C, v \in WordNet\ interpretation\ of\ w\} \tag{6}$$

In Tables 1–3, the left column lists the category labels of the three data sets used in this study, while the second column shows the WordNet interpretation of words in category labels. Category labels represented by abbreviations or acronyms have been expanded to their full list of words shown as underlined words in the right column.

For a word w with a set of senses $S^w = \left\{s_1^w, \ldots, s_{ns(w)}^w\right\}$, we define

$$S_L(i) = \sum_{w_j \in W_L^C} \max_{1 \leqslant k \leqslant ns(w_j)} Sim\left(s_i^w, s_k^{w_j}\right) \tag{7}$$

and

$$S_N(i) = \sum_{w_j \in W_N^C} \max_{1 \leqslant k \leqslant ns(w_j)} Sim\left(s_i^w, s_k^{w_j}\right) \tag{8}$$

The sense of word w is then determined by

$$Sense(w) = \arg \max_{1 \leqslant i \leqslant ns(w)} \left(T(S_L(i), 0, -\infty) + S_N(i)\right) \tag{9}$$

The application of the thresholding function is to enforce that words appearing in category labels are more important than words appearing in the WordNet interpretation of labels in the computation of the objective function.

3.2. Determining the semantic similarity of each term with categories

For categorizing documents with respect to a category, the weight of each term should be correlated to its semantic similarity with the category. While it is possible to obtain the senses of all words in a document by applying WSD approaches, it is computationally expensive and achieving high accuracy in WSD remains a

Table 1
WordNet interpretations of words in the category label of the Reuters-21578 data set.

Category label	Interpretations of words in the category label by WordNet
grain	Grain: foodstuff prepared from the starchy grains of cereal grasses
wheat	Wheat: grains of common wheat
corn	Corn: tall annual cereal grass bearing kernels on large ears: widely cultivated in America in many varieties
earn	Earnings: the excess of revenues over outlays in a given period of time (including depreciation and other non-cash expenses); forecast: a prediction about how something (as the weather will develop)
acq	Acquisition: the act of contracting or assuming or acquiring possession of something; merger: an occurrence that involves the production of a union
ship	Shipping: conveyance provided by the ships belonging to one country or industry
trade	Trade: the commercial exchange (buying and selling on domestic or international markets of goods and services)
crude	Crude oil: a dark oil consisting mainly of hydrocarbons
money-fx	Foreign exchange: the system by which one currency is exchanged for another; money: the official currency issued by a government or national bank
interest	Interest rate: the percentage of a sum of money charged for its use

Table 2
WordNet interpretations of words in the category label of the 20 newsgroups data set.

Category label	Interpretations of words in the category label by WordNet
comp.os.ms-windows.misc	Operating system: (computer science software that controls the execution of computer programs and may provide various services); window: (computer science a rectangular part of a computer screen that contains a display different from the rest of the screen); windows: (trademark an operating system with a graphical user interface)
rec.sport.baseball	Baseball: a ball game played with a bat and ball between two teams of nine players
rec.autos	Automobile: a motor vehicle with four wheels
rec.motorcycles	Motorcycle: a motor vehicle with two wheels and a strong frame
talk.politics.misc	Politics: the study of government of states and other political units
sci.crypt	Cryptography: act of writing in code or cipher
soc.religion.christian	Christian: a religious person who believes Jesus is the Christ and who is a member of a Christian denomination; Religion: an institution to express belief in a divine power
comp.graphics	Computer graphic: an image that is generated by a computer; computer graphics: the pictorial representation and manipulation of data by a computer
alt.atheism	Atheism: a lack of belief in the existence of God or gods
comp.sys.mac.hardware	Hardware: (computer science the mechanical, magnetic, electronic, and electrical components making up a computer system); macintosh: a waterproof raincoat made of rubberized fabric
sci.electronics	Electronics: the branch of physics that deals with the emission and effects of electrons and with the use of electronic devices
rec.sport.hockey	Hockey: a game played on an ice rink by two opposing teams of six skaters each who try to knock a flat round puck into the opponents' goal with angled sticks
talk.politics.guns	Gun: a weapon that discharges a missile at high velocity (especially from a metal tube or barrel); politics: the study of government of states and other political units
sci.med	Medicine: the learned profession that is mastered by graduate training in a medical school and that is devoted to preventing or alleviating or curing diseases and injuries
comp.sys.ibm.pc.hardware	Hardware: (computer science the mechanical, magnetic, electronic, and electrical components making up a computer system); personal computer: a small digital computer based on a microprocessor and designed to be used by one person at a time
sci.space	Aerospace: the atmosphere and outer space considered as a whole
comp.windows.x	Operating system: (computer science software that controls the execution of computer programs and may provide various services); window: (computer science a rectangular part of a computer screen that contains a display different from the rest of the screen); windows: (trademark an operating system with a graphical user interface); x: the cardinal number that is the sum of nine and one
misc.forsale	Sale: the general activity of selling
talk.politics.mideast	Mideast: the area around the eastern Mediterranean; politics: social relations involving authority or power
talk.religion.misc	Religion: an institution to express belief in a divine power

Table 3
WordNet interpretations of words in the category label of the WebKB data set.

Category label	Interpretations of words in the category label by WordNet
course	Course: education imparted in a series of lessons or meetings
faculty	Faculty: the body of teachers and administrators at a school
project	Project: any piece of work that is undertaken or attempted
student	Student: a learned person (especially in the humanities)

research challenge. In this study, we follow Bloehdorn et al. (2006) and adopt the first sense baseline, which is often hard to beat Navigli (2009). For example, three word sense disambiguation strategies were explored in Bloehdorn and Hotho (2004) and the data shown in Tables 1 and 2 indicated that choosing first sense resulted in close to best performance, with the difference in F1 values no larger than 1 basis point.

We explore two strategies to determine the semantic similarity $s(w,C)$ of each term w with the category C. We use S^C to denote the senses of words appearing in the category label as well as the interpretation of them by WordNet.

The first strategy is to compute the similarity of the first sense of each term with each sense in S^C and take the maximum value.

$$s(w,C) = \arg \max_{1 \leqslant k \leqslant |S^C|} Sim(s_1^w, s_k^C) \tag{10}$$

The second is to follow Eq. (10) for all words except for the top K words ranked by CHI for each category. For each category, these top K words are important and unique for the category, and determining their senses within the context of the category is both feasible and desirable. For each such word, we take the maximum value among the similarity values of all senses of the word with each sense in S^C:

$$s(w,C) = \arg \max_{1 \leqslant i \leqslant ns(w), 1 \leqslant k \leqslant |S^C|} Sim(s_i^w, s_k^C) \tag{11}$$

The underlying assumption of using maximization in Eq. (11) is that since the words are relevant to the category, it is reasonable to assume that the sense of each word is the one most similar to the semantics of the category. Therefore, the maximization strategy is only applicable to the top K words, not to the entire set of words in the corpus.

Then for categorization with respect to category C, the importance of each word w is assumed to be linearly correlated with $s(w,C)$.

$$Importance(w) = s(w,C) + \theta \tag{12}$$

θ serves as a smoothing factor to account for words that do not exist in WordNet. In this study θ is empirically determined to be 0.3, based on cross validation on the Reuters training data. And it is fixed for all three data sets. The underlying assumption is that a term with higher value of semantic similarity with a category should be more important for categorization and therefore deserves larger weight. For terms that do not exist in WordNet, $s(w,C)$ is set to 0 and $Importance(w)$ is equal to θ.

The proposed term weighting scheme TFSW (stands for term frequency semantic weighting) for a word w in document d with respect to category C is then defined by:

$$\begin{aligned} TFSW(w,d) &= count(w,d)(s(w,C) + \theta) \\ &\propto count(w,d)\left(\frac{s(w,C)}{\theta} + 1\right) \end{aligned} \tag{13}$$

The feature vector for each document is normalized before applying classifier algorithms, so the scaling in the above equation holds. For terms which do not exist in WordNet, the quotient in the parenthesis of the rightmost formula would be 0 and *TFSW* defaults to *TF*.

4. Experiments

4.1. Data sets

The effectiveness of the proposed approach is evaluated on three commonly used data sets: Reuters-21578, 20 Newsgroups data set, and WebKB. Documents are processed with a stop list and no stemming is performed.

Reuters-21578 data set[4] contains 21578 articles collected from the Reuters newswire. We follow the ModApte split to obtain training and testing data sets. After removing documents without labels or without body, the training data set contains 7775 documents, and the test data set contains 3019 documents. The 20 Newsgroups data set,[5] collected by Ken Lang, contains about 20,000 articles evenly divided among 20 USENET discussion groups. The training and test data sets are obtained by 60%/40% split based on temporal order, provided by the web page. The WebKB data set[6] consists of web pages collected from computer science departments of many universities. We follow the common practice of using four categories: course, faculty, student, and project. To prevent web pages of the same university from appearing in both training and testing data sets, we define the split in the following way: the training data set consists of the 3153 pages in the "misc" subdirectory while the testing data set consists of 1041 pages in the subdirectories of the four named universities.

For comparing categorization performance of the different term weighting schemes, we use the standard measures: precision, recall and F1 values with macro-averaging as well micro-averaging Sebastiani (2002). We employ the linear kernel support vector machine (SVM) as the classifier inducer, and the implementation is based on Joachims (1998). The soft margin parameter c has a significant impact on the performance for the Reuters data set, as has been observed by Bloehdorn et al. (2006). We have found that the performance peaks around $c = 5$ for the first two data sets when TF-IDF weighting is used. So c is set to 5 for the first two data sets, and the default value 1 for the third one.

4.2. Results and analysis

Table 4 shows the results of proposed TFSW1 and TFSW2 weighting schemes against TF-IDF on the three data sets. On the Reuters-21578 data set with 90 categories, although there is only about 1% improvement for Micro-F1 values, the Macro-F1 values increase for nearly 10% and 20%. This indicates the proposed term weighting schemes are more effective to improve the performance of small categories. In summary, the proposed TFSW1 and TFSW2 weighting schemes outperform TF-IDF on the Reuters and WebKB data sets, but underperform TF-IDF on the 20 Newsgroups data set.

In order to simulate the situation of small amount of training data, we follow Bloehdorn et al. (2006) and randomly sample the full training data set with different sizes represented by percentages. For each size, we repeat the sampling ten times to reduce the effect of sampling variations. The average performance figures of four term weighting schemes (TF, TF-IDF, TFSW1 and TFSW2) at each size are reported in Tables 5–10 for the three data sets. TFSW1 and TFSW2 denote the proposed semantic weighting scheme with the two strategies for computing $s(w,C)$. We use paired T-test to evaluation of the statistical significance of the difference. The superscript after the inequality signs denotes the significance level of 5%, 1%, 0.5% and 0.1% for 4, 3, 2 and 1, correspondingly.

We can draw the following conclusions from the figures:

1. The proposed term weighting schemes TFSW1 and TFSW2 significantly outperform TF and TF-IDF when the amount of training data is small, which holds for all three data sets.
2. On the Reuters-21578 data set, TFSW1 and TFSW2 consistently outperform TF and TF-IDF regardless of the amount of training data.
3. On the 20 Newsgroups data set, TFSW1 outperform TF when the amount of training data is less than 20% and TFSW2 outperform TF when the amount of training data is less than 30%. Both TFSW1 and TFSW2 outperform TF-IDF when the amount of training data is less than 80%.
4. On the WebKB data set, both TFSW1 and TFSW2 outperform TF on Micro-F1 when the amount of training data is less than 3% or larger than 40%. The results on Macro-F1 are mixed. Both TFSW1 and TFSW2 outperform TF-IDF regardless of the amount of training data.
5. TF outperforms TF-IDF most of the time, and the difference is larger with smaller amount of training data, which can be attributed to the unreliable values of IDF.
6. TFSW2 outperforms TFSW1 most of the time. This shows that determining the senses of a set of important words from context helps improve performance.

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[5] http://people.csail.mit.edu/jrennie/20newsgroups/.
[6] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/.

Table 4
Results of TFSW1 and TFSW2 against TF-IDF on the three data sets.

Data set		TF	TF-IDF	TFSW1	TFSW2
Reuters 90 category	Micro-F1	0.8713	0.8738	0.8822 (+1.25%, +0.96%)	0.8857 (+1.65%, +1.36%)
	Macro-F1	0.4923	0.5299	0.5404 (+9.77%, +1.98%)	0.5938 (+20.62%, +12.06%)
Reuters 10 category	Micro-F1	0.9262	0.9268	0.9313 (+0.55%, +0.49%)	0.9313 (+0.55%, +0.49%)
	Macro-F1	0.8579	0.8596	0.8744 (+1.94%, +1.72%)	0.8752 (+2.03%, +1.81%)
20 News-groups	Micro-F1	0.7291	0.7349	0.7236 (-0.75%, −1.54%)	0.7259 (-0.44%, −1.22%)
	Macro-F1	0.7218	0.7279	0.7158 (-0.83%, −1.66%)	0.7171 (-0.65%, −1.48%)
WebKB	Micro-F1	0.8398	0.8250	0.8572 (+2.07%, +3.90%)	0.8533 (+1.61%, +3.43%)
	Macro-F1	0.8173	0.7981	0.8284 (+1.37%, +3.80%)	0.8223 (+0.62%, +3.03%)

Percentages of relative improvements over TF and TF-IDF are shown in parentheses.

Table 5
Results of TFSW1 on Reuters-21578 top 10 categories.

Sample size (%)	TF		TFIDF		TFSW1		Against TF		Against TFIDF	
	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
1	0.6855	0.3194	0.1249	0.0357	0.7224	0.4223	5.38%>^{4}	32.22%>^{4}	478.38%>^{4}	1082.91%>^{4}
2	0.7688	0.4429	0.2214	0.0711	0.7935	0.5313	3.21%>^{4}	19.96%>^{4}	258.40%>^{4}	647.26%>^{4}
3	0.8155	0.5501	0.2826	0.1070	0.8317	0.6223	1.99%>^{4}	13.12%>^{4}	194.30%>^{4}	481.59%>^{4}
4	0.8355	0.6115	0.3221	0.1377	0.8519	0.6878	1.96%>^{4}	12.48%>^{4}	164.48%>^{4}	399.49%>^{4}
5	0.8509	0.6564	0.3552	0.1708	0.8636	0.7172	1.49%>^{4}	9.26%>^{4}	143.13%>^{4}	319.91%>^{4}
6	0.8612	0.6832	0.3867	0.2003	0.8707	0.7346	1.10%>^{4}	7.52%>^{4}	125.16%>^{4}	266.75%>^{4}
7	0.8687	0.7088	0.4144	0.2285	0.8779	0.7605	1.06%>^{4}	7.29%>^{4}	111.85%>^{4}	232.82%>^{4}
8	0.8755	0.7306	0.4387	0.2522	0.8822	0.7726	0.77%>^{4}	5.75%>^{4}	101.09%>^{4}	206.34%>^{4}
9	0.8797	0.7437	0.4635	0.2770	0.8880	0.7892	0.94%>^{4}	6.12%>^{4}	91.59%>^{4}	184.91%>^{4}
10	0.8825	0.7518	0.4829	0.2995	0.8907	0.7968	0.93%>^{4}	5.99%>^{4}	84.45%>^{4}	166.04%>^{4}
20	0.9033	0.8042	0.6271	0.4572	0.9073	0.8271	0.44%>^{4}	2.85%>^{4}	44.68%>^{4}	80.91%>^{4}
30	0.9121	0.8281	0.7187	0.5642	0.9139	0.8407	0.20%>^{1}	1.52%>^{4}	27.16%>^{4}	49.01%>^{4}
40	0.9163	0.8376	0.7875	0.6510	0.9189	0.8495	0.28%>^{3}	1.42%>^{4}	16.69%>^{4}	30.49%>^{4}
50	0.9193	0.8430	0.8369	0.7177	0.9212	0.8537	0.21%>^{2}	1.27%>^{4}	10.07%>^{4}	18.95%>^{4}
60	0.9227	0.8512	0.8725	0.7707	0.9246	0.8598	0.21%>^{1}	1.01%>^{4}	5.97%>^{4}	11.56%>^{4}
70	0.9242	0.8547	0.8985	0.8102	0.9256	0.8628	0.15%>^{2}	0.95%>^{4}	3.02%>^{4}	6.49%>^{4}
80	0.9257	0.8565	0.9155	0.8380	0.9279	0.8657	0.24%>^{4}	1.07%>^{4}	1.35%>^{4}	3.31%>^{4}
90	0.9271	0.8602	0.9244	0.8541	0.9293	0.8699	0.24%>^{3}	1.13%>^{4}	0.53%>^{4}	1.85%>^{4}
100	0.9262	0.8578	0.9268	0.8596	0.9313	0.8744	0.55%	1.94%	0.49%	1.72%

We use paired *T*-test to evaluation of the statistical significance of the difference. The superscript after the inequality signs denotes the significance level of 5%, 1%, 0.5% and 0.1% for 4, 3, 2 and 1, correspondingly.

Table 6
Results of TFSW2 on Reuters-21578 top 10 categories.

Sample size (%)	TF		TFIDF		TFSW2		Against TF		Against TFIDF	
	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
1	0.6855	0.3194	0.1249	0.0357	0.7501	0.5085	9.42%>^{4}	59.20%>^{4}	500.56%>^{4}	1324.37%>^{4}
2	0.7688	0.4429	0.2214	0.0711	0.8226	0.6348	7.00%>^{4}	43.33%>^{4}	271.54%>^{4}	792.83%>^{4}
3	0.8155	0.5501	0.2826	0.1070	0.8540	0.7043	4.72%>^{4}	28.03%>^{4}	202.19%>^{4}	558.22%>^{4}
4	0.8355	0.6115	0.3221	0.1377	0.8675	0.7466	3.83%>^{4}	22.09%>^{4}	169.33%>^{4}	442.19%>^{4}
5	0.8509	0.6564	0.3552	0.1708	0.8756	0.7630	2.90%>^{4}	16.24%>^{4}	146.51%>^{4}	346.72%>^{4}
6	0.8612	0.6832	0.3867	0.2003	0.8815	0.7764	2.36%>^{4}	13.64%>^{4}	127.95%>^{4}	287.62%>^{4}
7	0.8687	0.7088	0.4144	0.2285	0.8847	0.7879	1.84%>^{4}	11.16%>^{4}	113.49%>^{4}	244.81%>^{4}
8	0.8755	0.7306	0.4387	0.2522	0.8898	0.7984	1.63%>^{4}	9.28%>^{4}	102.83%>^{4}	216.57%>^{4}
9	0.8797	0.7437	0.4635	0.2770	0.8934	0.8106	1.56%>^{4}	9.00%>^{4}	92.75%>^{4}	192.64%>^{4}
10	0.8825	0.7518	0.4829	0.2995	0.8955	0.8151	1.47%>^{4}	8.42%>^{4}	85.44%>^{4}	172.15%>^{4}
20	0.9033	0.8042	0.6271	0.4572	0.9100	0.8406	0.74%>^{4}	4.53%>^{4}	45.11%>^{4}	83.86%>^{4}
30	0.9121	0.8281	0.7187	0.5642	0.9168	0.8514	0.52%>^{4}	2.81%>^{4}	27.56%>^{4}	50.90%>^{4}
40	0.9163	0.8376	0.7875	0.6510	0.9214	0.8590	0.56%>^{4}	2.55%>^{4}	17.00%>^{4}	31.95%>^{4}
50	0.9193	0.8430	0.8369	0.7177	0.9232	0.8616	0.42%>^{4}	2.21%>^{4}	10.31%>^{4}	20.05%>^{4}
60	0.9227	0.8512	0.8725	0.7707	0.9258	0.8656	0.34%>^{3}	1.69%>^{4}	6.11%>^{4}	12.31%>^{4}
70	0.9242	0.8547	0.8985	0.8102	0.9280	0.8687	0.41%>^{4}	1.64%>^{4}	3.28%>^{4}	7.22%>^{4}
80	0.9257	0.8565	0.9155	0.8380	0.9296	0.8715	0.42%>^{4}	1.75%>^{4}	1.54%>^{4}	4.00%>^{4}
90	0.9271	0.8602	0.9244	0.8541	0.9306	0.8733	0.38%>^{4}	1.52%>^{4}	0.67%>^{4}	2.25%>^{4}
100	0.9262	0.8578	0.9268	0.8596	0.9313	0.8752	0.55%	2.03%	0.49%	1.81%

When the amount of training data is small, the statistical information of words are not reliable to highlight important words for characterizing each category. Semantic information of categories and words help to identify words relevant to the semantics of each

Table 7
Results of TFSW1 on 20 newsgroups.

Sample size (%)	TF		TFIDF		TFSW1		Against TF		Against TFIDF	
	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
1	0.0372	0.0351	0.0003	0.0003	0.1364	0.1136	266.67%>^{4}	223.65%>^{4}	45366.67%>^{4}	37766.67%>^{4}
2	0.1175	0.1091	0.0007	0.0007	0.2416	0.2116	105.62%>^{4}	93.95%>^{4}	34414.29%>^{4}	30128.57%>^{4}
3	0.1994	0.1855	0.0019	0.0018	0.3063	0.2756	53.61%>^{4}	48.57%>^{4}	16021.05%>^{4}	15211.11%>^{4}
4	0.2681	0.2515	0.0037	0.0036	0.3546	0.3247	32.26%>^{4}	29.11%>^{4}	9483.78%>^{4}	8919.44%>^{4}
5	0.3247	0.3080	0.0065	0.0063	0.3932	0.3661	21.10%>^{4}	18.86%>^{4}	5949.23%>^{4}	5711.11%>^{4}
6	0.3675	0.3504	0.0105	0.0103	0.4290	0.4022	16.73%>^{4}	14.78%>^{4}	3985.71%>^{4}	3804.85%>^{4}
7	0.4027	0.3848	0.0148	0.0145	0.4529	0.4279	12.47%>^{4}	11.20%>^{4}	2960.14%>^{4}	2851.03%>^{4}
8	0.4308	0.4136	0.0219	0.0215	0.4733	0.4500	9.87%>^{4}	8.80%>^{4}	2061.19%>^{4}	1993.02%>^{4}
9	0.4550	0.4384	0.0293	0.0288	0.4913	0.4696	7.98%>^{4}	7.12%>^{4}	1576.79%>^{4}	1530.56%>^{4}
10	0.4770	0.4604	0.0369	0.0362	0.5063	0.4859	6.14%>^{4}	5.54%>^{4}	1272.09%>^{4}	1242.27%>^{4}
20	0.5921	0.5787	0.1463	0.1424	0.5987	0.5835	1.11%>^{4}	0.83%>^{3}	309.23%>^{4}	309.76%>^{4}
30	0.6407	0.6289	0.2817	0.2741	0.6407	0.6281	0.00%∼	−0.13%∼	127.44%>^{4}	129.15%>^{4}
40	0.6675	0.6571	0.4037	0.3939	0.6633	0.6523	−0.63%<^{4}	−0.73%<^{4}	64.31%>^{4}	65.60%>^{4}
50	0.6848	0.6754	0.5076	0.4979	0.6813	0.6712	−0.51%<^{4}	−0.62%<^{4}	34.22%>^{4}	34.81%>^{4}
60	0.6967	0.6879	0.5847	0.5754	0.6923	0.6827	−0.63%<^{4}	−0.76%<^{4}	18.40%>^{4}	18.65%>^{4}
70	0.7067	0.6984	0.6476	0.6391	0.7033	0.6944	−0.48% <^{4}	−0.57%<^{4}	8.60%>^{4}	8.65%>^{4}
80	0.7152	0.7072	0.6931	0.6851	0.7115	0.7029	−0.52%<^{4}	−0.61%<^{4}	2.65%>^{4}	2.60%>^{4}
90	0.7227	0.7150	0.7220	0.7147	0.7179	0.7097	−0.66%<^{4}	−0.74%<^{4}	−0.57%<^{4}	−0.70%<^{4}
100	0.7291	0.7218	0.7349	0.7279	0.7236	0.7158	−0.75%	−0.83%	−1.54%	−1.66%

Table 8
Results of TFSW2 on 20 newsgroups.

Sample size (%)	TF		TFIDF		TFSW2		Against TF		Against TFIDF	
	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
1	0.0372	0.0351	0.0003	0.0003	0.1758	0.1490	372.58%>^{4}	324.50%>^{4}	58500.00%>^{4}	49566.67%>^{4}
2	0.1175	0.1091	0.0007	0.0007	0.2892	0.2554	146.13%>^{4}	134.10%>^{4}	41214.29%>^{4}	36385.71%>^{4}
3	0.1994	0.1855	0.0019	0.0018	0.3469	0.3154	73.97%>^{4}	70.03%>^{4}	18157.89%>^{4}	17422.22%>^{4}
4	0.2681	0.2515	0.0037	0.0036	0.3901	0.3601	45.51%>^{4}	43.18%>^{4}	10443.24%>^{4}	9902.78%>^{4}
5	0.3247	0.3080	0.0065	0.0063	0.4247	0.3972	30.80%>^{4}	28.96%>^{4}	6433.85%>^{4}	6204.76%>^{4}
6	0.3675	0.3504	0.0105	0.0103	0.4520	0.4262	22.99%>^{4}	21.63%>^{4}	4204.76%>^{4}	4037.86%>^{4}
7	0.4027	0.3848	0.0148	0.0145	0.4755	0.4509	18.08%>^{4}	17.18%>^{4}	3112.84%>^{4}	3009.66%>^{4}
8	0.4308	0.4136	0.0219	0.0215	0.4917	0.4690	14.14%>^{4}	13.39%>^{4}	2145.21%>^{4}	2081.40%>^{4}
9	0.4550	0.4384	0.0293	0.0288	0.5077	0.4862	11.58%>^{4}	10.90%>^{4}	1632.76%>^{4}	1588.19%>^{4}
10	0.4770	0.4604	0.0369	0.0362	0.5221	0.5018	9.45%>^{4}	8.99%>^{4}	1314.91%>^{4}	1286.19%>^{4}
20	0.5921	0.5787	0.1463	0.1424	0.6052	0.5904	2.21%>^{4}	2.02%>^{4}	313.67%>^{4}	314.61%>^{4}
30	0.6407	0.6289	0.2817	0.2741	0.6448	0.6322	0.64%>^{4}	0.52%>^{3}	128.90%>^{4}	130.65%>^{4}
40	0.6675	0.6571	0.4037	0.3939	0.6659	0.6548	−0.24%∼	−0.35%<^{1}	64.95%>^{4}	66.24%>^{4}
50	0.6848	0.6754	0.5076	0.4979	0.6831	0.6729	−0.25%<^{2}	−0.37%<^{3}	34.57%>^{4}	35.15%>^{4}
60	0.6967	0.6879	0.5847	0.5754	0.6956	0.6859	−0.16%<^{1}	−0.29%<^{3}	18.97%>^{4}	19.20%>^{4}
70	0.7067	0.6984	0.6476	0.6391	0.7045	0.6954	−0.31%<^{3}	−0.43%<^{4}	8.79%>^{4}	8.81%>^{4}
80	0.7152	0.7072	0.6931	0.6851	0.7128	0.7038	−0.34%<^{2}	−0.48%<^{3}	2.84%>^{4}	2.73%>^{4}
90	0.7227	0.7150	0.7220	0.7147	0.7202	0.7116	−0.35%<^{3}	−0.48%<^{4}	−0.25%∼	−0.43%<^{2}
100	0.7291	0.7218	0.7349	0.7279	0.7259	0.7171	−0.44%	−0.65%	−1.22%	−1.48%

Table 9
Results of TFSW1 on WebKB.

Sample size (%)	TF		TFIDF		TFSW1		Against TF		Against TFIDF	
	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
1	0.2812	0.1963	0.0009	0.0004	0.3309	0.2397	17.67% >^{3}	22.11%>^{4}	36666.67%>^{4}	59825.00%>^{4}
2	0.5258	0.4010	0.0026	0.0012	0.5783	0.4478	9.98% >^{2}	11.67%>^{4}	22142.31%>^{4}	37216.67%>^{4}
3	0.6409	0.4936	0.0057	0.0028	0.6642	0.5147	3.64% >^{3}	4.27%>^{3}	11552.63%>^{4}	18282.14%>^{4}
4	0.6831	0.5448	0.0108	0.0059	0.6849	0.5365	0.26%∼	−1.52%∼	6241.67%>^{4}	8993.22%>^{4}
5	0.7047	0.5810	0.0183	0.0110	0.7081	0.5813	0.48%∼	0.05%∼	3769.40%>^{4}	5184.55%>^{4}
6	0.7197	0.6148	0.0232	0.0147	0.7244	0.6077	0.65%∼	−1.15%∼	3022.41%>^{4}	4034.01%>^{4}
7	0.7369	0.6497	0.0304	0.0197	0.7420	0.6460	0.69%∼	−0.57%∼	2340.79%>^{4}	3179.19%>^{4}
8	0.7496	0.6722	0.0402	0.0268	0.7500	0.6595	0.05%∼	−1.89%<^{1}	1765.67%>^{4}	2360.82%>^{4}
9	0.7581	0.6860	0.0502	0.0347	0.7619	0.6754	0.50%∼	−1.55%∼	1417.73%>^{4}	1846.40%>^{4}
10	0.7690	0.7008	0.0605	0.0429	0.7732	0.6872	0.55%∼	−1.94%<^{1}	1178.02%>^{4}	1501.86%>^{4}
20	0.8041	0.7646	0.1637	0.1250	0.8036	0.7524	−0.06%∼	−1.60%<^{3}	390.90%>^{4}	501.92%>^{4}
30	0.8170	0.7896	0.2750	0.2205	0.8189	0.7821	0.23%∼	−0.95%<^{3}	197.78%>^{4}	254.69%>^{4}
40	0.8247	0.8024	0.3988	0.3373	0.8339	0.8021	1.12%>^{4}	−0.04%∼	109.10%>^{4}	137.80%>^{4}
50	0.8309	0.8094	0.5261	0.4618	0.8398	0.8083	1.07%>^{4}	−0.14%∼	59.63%>^{4}	75.03%>^{4}
60	0.8333	0.8124	0.6295	0.5726	0.8438	0.8134	1.26%>^{4}	0.12%∼	34.04%>^{4}	42.05% >^{4}
70	0.8363	0.8169	0.7088	0.6585	0.8475	0.8200	1.34%>^{4}	0.38%∼	19.57%>^{4}	24.53% >^{4}
80	0.8366	0.8161	0.7686	0.7299	0.8477	0.8190	1.33%>^{4}	0.36%∼	10.29%>^{4}	12.21% >^{4}
90	0.8392	0.8157	0.8065	0.7759	0.8513	0.8224	1.44%>^{4}	0.82%>^{3}	5.55%>^{4}	5.99%>^{4}
100	0.8398	0.8172	0.8250	0.7981	0.8572	0.8284	2.07%	1.37%	3.90%	3.80%

Table 10
Results of TFSW2 on WebKB.

Sample size (%)	TF		TFIDF		TFSW2		Against TF		Against TFIDF	
	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
1	0.2812	0.1963	0.0009	0.0004	0.3719	0.2780	32.25%>^{4}	41.62%>^{4}	41222.22%>^{4}	69400.00%>^{4}
2	0.5258	0.4010	0.0026	0.0012	0.5790	0.4490	10.12%>^{3}	11.97%>^{4}	22169.23%>^{4}	37316.67%>^{4}
3	0.6409	0.4936	0.0057	0.0028	0.6520	0.5046	1.73%∼	2.23%∼	11338.60%>^{4}	17921.43%>^{4}
4	0.6831	0.5448	0.0108	0.0059	0.6719	0.5298	−1.64%∼	−2.75%<^{1}	6121.30%>^{4}	8879.66%>^{4}
5	0.7047	0.5810	0.0183	0.0110	0.6973	0.5712	−1.05%<^{1}	−1.69%∼	3710.38%>^{4}	5092.73%>^{4}
6	0.7197	0.6148	0.0232	0.0147	0.7131	0.5983	−0.92%∼	−2.68%<^{1}	2973.71%>^{4}	3970.07%>^{4}
7	0.7369	0.6497	0.0304	0.0197	0.7299	0.6338	−0.95%∼	−2.45%<^{1}	2300.99%>^{4}	3117.26%>^{4}
8	0.7496	0.6722	0.0402	0.0268	0.7414	0.6522	−1.09%<^{1}	−2.98%<^{1}	1744.28%>^{4}	2333.58%>^{4}
9	0.7581	0.6860	0.0502	0.0347	0.7495	0.6605	−1.13%<^{1}	−3.72%<^{2}	1393.03%>^{4}	1803.46%>^{4}
10	0.7690	0.7008	0.0605	0.0429	0.7590	0.6709	−1.30%<^{1}	−4.27%<^{4}	1154.55%>^{4}	1463.87%>^{4}
20	0.8041	0.7646	0.1637	0.1250	0.7986	0.7471	−0.68%<^{1}	−2.29%<^{4}	387.84%>^{4}	497.68%>^{4}
30	0.8170	0.7896	0.2750	0.2205	0.8166	0.7774	−0.05%∼	−1.55%<^{4}	196.95%>^{4}	252.56%>^{4}
40	0.8247	0.8024	0.3988	0.3373	0.8308	0.7978	0.74%>^{3}	−0.57%<^{1}	108.32%>^{4}	136.53%>^{4}
50	0.8309	0.8094	0.5261	0.4618	0.8393	0.8055	1.01%>^{4}	−0.48%∼	59.53%>^{4}	74.43%>^{4}
60	0.8333	0.8124	0.6295	0.5726	0.8456	0.8137	1.48%>^{4}	0.16%∼	34.33%>^{4}	42.11%>^{4}
70	0.8363	0.8169	0.7088	0.6585	0.8489	0.8173	1.51%>^{4}	0.05%∼	19.77%>^{4}	24.12%>^{4}
80	0.8366	0.8161	0.7686	0.7299	0.8507	0.8198	1.69%>^{4}	0.45%∼	10.68%>^{4}	12.32%>^{4}
90	0.8392	0.8157	0.8065	0.7759	0.8518	0.8207	1.50%>^{4}	0.61%>^{1}	5.62%>^{4}	5.77%>^{4}
100	0.8398	0.8172	0.8250	0.7981	0.8533	0.8223	1.61%	0.62%	3.43%	3.03%

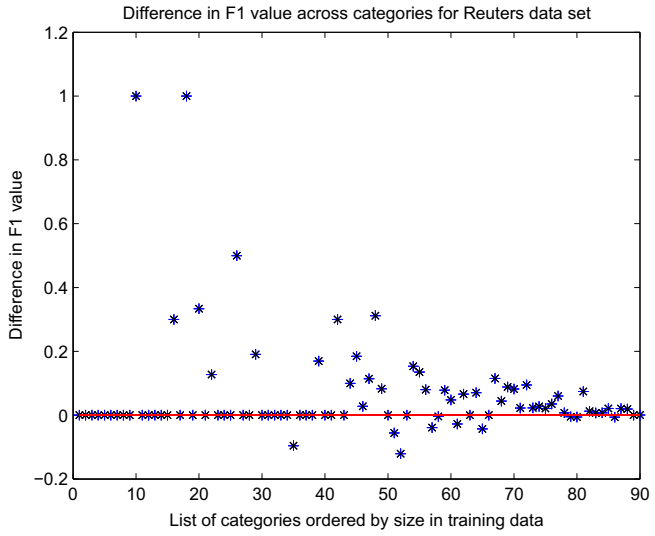

Fig. 2. Difference in F1 value across categories for Reuters data set.

category. As a result, the proposed semantic weighting schemes TFSW1 and TFSW2 significantly outperform statistical weighting schemes TF and TF-IDF.

However, with sufficient amount of training data, the characteristics of data sets play an role in determining the winner. When the documents in a category have focused content and the category labels are effective in summarizing such content, the proposed semantic weighting scheme wins. As stated in Barak et al. (2009), Reuters articles were written by professional journalists, and the writing style is formal and focused. Documents in the 20 Newsgroups data set were written by ordinary users in a casual and loose style. The categories of newsgroups are not well organized and many of them share similar content. For example, five of them are comp.* discussion groups, three of them discuss religion, and three of them discuss politics. Indeed, "talk.religion.misc" and "talk.politics.misc" are among the worst performing categories for TFSW2 in comparison to TF-IDF. Documents in the WebKB data set somehow fall in between with respect to these characteristics.

Fig. 2 shows the difference in F1 value of TFSW2 against TFIDF across categories for the Reuters data set, where the horizontal axis denotes the 90 categories listed in increasing order of size. It is observed that the difference is non-negative for the majority of categories with varying sizes, which substantiates the effectiveness of the proposed approach. The amount of improvement is in general larger for smaller categories.

Another factor that leads to the underperformance on the Newsgroups data set can be attributed to the limited coverage of WordNet. As mentioned in Mansuy and Hilderman (2006), WordNet is a general ontology for the English language without domain-specific knowledge. For example, words such as "machintosh" for the category "comp.sys.mac.hardware", and names of baseball teams for the category "rec.sport.baseball" are mis-interpreted or missing. These words are actually crucial for determining the categories for documents.

4.3. Performance comparison against previous studies

We compare the performance of the proposed approach against two previous studies exploiting WordNet as external knowledge to enhance text categorization.

Table 11
Performance comparison against Bloehdorn et al. (2006).

Sample size		2%	3%	4%	5%
Bloehdorn et al. (2006)	Baseline	0.45	0.51	0.54	0.57
	Best result	0.53	0.57	0.61	0.62
	Gain	18%	12%	13%	9.8%
TFSW2	Baseline	0.44	0.55	0.61	0.66
	Result	0.63	0.70	0.75	0.76
	Gain	43%	27%	23%	15%

Table 12
Performance comparison against Bloehdorn and Hotho (2004).

		Micro-F1	Macro-F1
Bloehdorn and Hotho (2004)	Baseline	0.8421	0.7275
	Best result	0.8589	0.7514
	Gain	2.00%	3.29%
TFSW2	Baseline	0.8848	0.7010
	Result	0.8969	0.7662
	Gain	1.37%	9.30%

Table 11 compares the gain of Macro-F1 over baseline by TFSW2 against the best results in Table 1 of Bloehdorn et al. (2006). We following the same way of generating random samples of several sizes from the Reuters data set and evaluating performance on the top 10 categories. Due to sampling variation and subtle differences in preprocessing, the baseline figures in this study are not quite the same. The table shows that the gain percentages with four samples sizes of the Reuters data set by TFSW2 are significantly larger than the corresponding figures in Bloehdorn et al. (2006).

Table 12 compares the gain of Macro-F1 and Micro-F1 values over baseline against the best results in Table 1 of Bloehdorn and Hotho (2004). We following the same way of evaluating performance on the top 50 categories of the Reuters data set. Again, due to differences in preprocessing and classifier algorithms, the baseline figures in this study are not quite the same. The table shows that the gain percentage on Macro-F1 of TFSW2 is significantly larger, while the gain percentage on Micro-F1 is smaller. The latter can be partly attributed to the relatively high value of our baseline. In fact, our baseline figure of 0.8848is higher than both the baseline figure of 0.8421 and improved value 0.8589 in Bloehdorn and Hotho (2004).

4.4. Computational cost

The first step of automatically determining the semantics of categories has a cost of $O\left(\sum_{w\in \texttt{Categorylabels}} ns(w)\sum_{w_j\in CW(w)} ns(w_j)\ \texttt{Cost}\{Sim^{Lin}\}\right)$ This step can be performed off-line before categorization. Furthermore, if the domain experts can provide the semantics based on WordNet as part of problem definition, this step can be skipped. The second step of determining the semantic similarity of each term with categories has a cost of $O\left(\sum_C\sum_w |S^C||S^w|\ \texttt{Cost}\{Sim^{Lin}\}\right)$ for TFSW2 and $O\left(\sum_C\sum_w |S^C|\texttt{Cost}\{Sim^{Lin}\}\right)$ for TFSW1. Empirical measurements of executing C++ code on the three data sets indicate that for each category the extra time taken by TFSW2 over TF-IDF ranges between 0.5 and 3.5 seconds on a PC with Intel Pentium E2140 CPU and 1.5 GB memory.

5. Concluding remarks

In this paper, we have proposed a novel semantic term weighting scheme for text categorization. Specifically, the semantics of categories are represented by senses of words appearing in the category labels as well as the interpretations of them by WordNet. And the weight of a term is correlated to its semantic similarity with a category. Experimental results on three commonly used data sets show that the proposed approach significantly outperforms TF-IDF when the amount of training data is small or the content of documents is focused on well defined categories. In addition, the proposed approach compares favorably with two previous studies. To further improve the performance, we plan to employ other ontologies with wider coverage (such as Wikipedia) for expressing the senses of words and category labels. We also plan to explore other ways of representing the semantics of categories and other similarity measures.

Acknowledgement

This work was supported by grants from MOE-MS Key Laboratory of Multimedia Computing and Communication of USTC (Grant No. 07122807), Natural Science Foundation of China (Grant No. 61073110, 60775037), the Key Program of National Natural Science Foundation of China (Grant No. 60933013), and Research Fund for the Doctoral Program of Higher Education of China (20093402110017).

References

Barak, L., Dagan, I., & Shnarch, E. (2009). Text categorization from category name via lexical reference. In *NAACL '09: Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics, companion volume: Short papers* (pp. 33–36).

Bloehdorn, S., & Hotho, A. (2004). Boosting for text classification with semantic features. In *Proceedings of the MSW 2004 workshop at the 10th ACM SIGKDD conference* (pp. 70–87).

Bloehdorn, S., Basili, R., Cammisa, M., & Moschitti, A. (2006). Semantic kernels for text classification based on topological measures of feature similarity. In *ICDM '06: Proceedings of the sixth international conference on data mining* (pp. 808–812).

Chang, Y. C., Chen, S. M., & Liau, C. J. (2008). Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method. *Expert Systems with Applications, 34*, 1948–1953.

de Buenaga Rodriguez, M., Gmez-Hidalgo, J. M., & Diaz-Agudo, B. (1997). Using WordNet to complement training information in text categorization. In *Selected papers from the second international conference on recent advances in natural language processing (RANLP 1997)* (pp. 353–364).

Gabrilovich, E., & Markovitch, S. (2007). Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research, 8*, 2297–2345.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning (ECML)* (pp. 137–142).

Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*, 721–735.

Li, J., Zhao, Y., & Liu, B. (2009). Fully automatic text categorization by exploiting wordnet. In *AIRS* (pp. 1–12).

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning* (pp. 296–304).

Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications, 36*, 690–701.

Mansuy, T., & Hilderman, R. J. (2006). A characterization of wordnet features in boolean models for text classification. In *AusDM '06: Proceedings of the fifth Australasian conference on data mining and analytics* (pp. 103–109).

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *Journal of Lexicography, 3*, 235–244.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*, 513–523.

Scott, S., & Matwin, S. (1998). Text classification using Word Net hypernyms. In *Proceedings of the COLING/ACL workshop on usage of WordNet in natural language processing systems*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*, 1–47.

Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference* (pp. 713–721).

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML* (pp. 412–420).