Knowledge-Based Systems 66 (2014) 99–111

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Authorship identification from unstructured texts

Chunxia Zhang [a,*], Xindong Wu [b], Zhendong Niu [c], Wei Ding [d]

[a] School of Software, Beijing Institute of Technology, Beijing 100081, China
[b] Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
[c] School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
[d] Computer Science Department, University of Massachusetts Boston, Boston, MA 02125, USA

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form 11 March 2014
Accepted 17 April 2014
Available online 2 May 2014

Keywords:
Semantic association model
Authorship identification
Linear discriminant analysis
Principal components analysis
Feature extraction

ABSTRACT

Authorship identification is a task of identifying authors of anonymous texts given examples of the writing of authors. The increasingly large volumes of anonymous texts on the Internet enhance the great yet urgent necessity for authorship identification. It has been applied to more and more practical applications including literary works, intelligence, criminal law, civil law, and computer forensics. In this paper, we propose a semantic association model about voice, word dependency relations, and non-subject stylistic words to represent the writing style of unstructured texts of various authors, design an unsupervised approach to extract stylistic features, and employ principal components analysis and linear discriminant analysis to identify authorship of texts. This paper provides a uniform quantified method to capture syntactic and semantic stylistic characteristics of and between words and phrases, and this approach can solve the problem of the independence of different dimensions to some extent. Experimental results on two English text corpora show that our approach significantly improves the overall performance over authorship identification.

1. Introduction

Authorship identification is a task of identifying authors of anonymous texts, according to the given examples of the writing of a predefined set of candidate authors [1,2]. The first work on authorship identification was to attribute authorship to the literacy work of the plays of Shakespeare in the nineteen century. In recent years, the increasingly large volumes of anonymous texts, such as online forum messages, emails, blogs, and source codes, enhance the great yet urgent necessity for authorship identification [1]. Authorship identification has been applied to more and more applications including literary works, intelligence, criminal law, civil law, and computer forensics [1–3]. It also plays an important role in many areas such as information retrieval, information extraction and question answering. In the literature, an application case of authorship identification was illustrated by identifying the authors of literary works with unknown or disputed authorship such as *The Federalist Papers* [4]. Another example in intelligence applications is to determine authors of online messages, given known security risks. Moreover, recognizing writers of offensive or threatening messages is discussed in criminal law applications. In addition, an example in computer forensics applications is to judge the identity of programmers of source codes which maybe destroy computers or data [1,5].

The task of authorship identification mainly focuses on two issues: how to extract features of texts to represent the writing styles of different authors [6], and how to select appropriate methods to predict authors of unrestricted texts. The text representation features, called style markers, need to be objective, quantifiable, content independent and un-ambiguously identifiable so that they can be employed to effectively discriminate a variety of authors of different kinds of texts [7].

The stylometric features used in current works can be divided into six types: character, lexical, syntactic, structural, semantic and application-specific features [1]. Character and lexical features use measures of characters, words, or punctuation marks as the textual style [6,8–12], while syntactic features utilize properties about part-of-speeches of words and the phrases of sentences as the style markers of documents [13]. Structural features are characteristics of the document structure such as word length, sentence length, use of indentation, and types of signatures [1,14,15]. In addition, application-specific features are ones related to a specific domain, language, or application [1,15].

Semantic features employed in the existing works include (a) binary semantic features and semantic modification relations [16], (b) synonyms, hypernyms, and causal verbs [17], and (c) func-

* Corresponding author.
E-mail addresses: cxzhang@bit.edu.cn (C. Zhang), xwu@cs.uvm.edu (X. Wu), zniu@bit.edu.cn (Z. Niu), ding@cs.umb.edu (W. Ding).

http://dx.doi.org/10.1016/j.knosys.2014.04.025
0950-7051/

tional features [18]. Binary semantic features consist of number and person features on nouns and pronouns, and tense, aspect, and sub-categorization features on verbs [16]. Semantic modification relations mean the modification relations between words of sentences. For example, "Noun Possr Noun" denotes the relation of a nominal node with a pronominal possessor, while "Noun Locn Noun" shows the relation of a nominal node with a nominal modifier indicating location [16]. Functional features are schemes which express the semantic function of certain words or phrases on some aspects of its preceding content based on the systemic functional grammar [1,18]. For instance, the word "specifically" signifies an "ELABORATION" of the "CONJUNCTION" scheme.

Actually, binary semantic features only capture the syntactic or semantic information about nouns, pronouns and verbs. Semantic modification relations are represented via the sequences of part-of-speeches of words about certain modification relations. Synonyms and hypernyms record the words with the same meanings and the inheritance relations, respectively. Functional features are the modification relations about certain words or phrases. However, those character, lexical, syntactic, and semantic features are constrained by some specific words, phrases, or part-of-speeches.

The above observation motivates us to consider (a) what features are capable of representing the essential semantic structures of sentences, (b) what features are independent of specific words, phrases, and part-of-speeches, (c) what features are independent of contents of different texts, and (d) what features maintain roughly constant across different documents of the same author. To this end, a semantic association model about word dependency relations, voice, and non-subject stylistic words is proposed in this paper to capture the writing style of authors. Word dependencies use the uniform binary typed dependency relations to express all relationships among individual words of sentences, while phrase relations in [19,20] only represent the nesting of multi-word constituents. Meanwhile, word dependencies also provide relations within a predicate–argument structure, while phrase relations in [19,20] cannot give such a kind of information. The predicate–argument structure forms the semantic backbone of a sentence, and most words in the sentence are the auxiliary components of this backbone. Hence, word dependencies provide characteristics of syntactic and semantic levels of sentences. Usually authors use those abstract structural semantic patterns in an unconscious way. Accordingly, such relationships are implicitly embedded in the writings of authors in different topics.

Voice features are to reflect the relationship between a verb of a sentence and a subject participating in the action that the verb describes. Features about non-subject stylistic words are intended to express the characteristics of words that are not related to the contents of texts, since subject words are to reflect the topics and contents of texts, and the intersection between the set of subject words and the set of non-subject stylistic words is usually empty. Therefore, features of word dependencies, voice, and non-subject stylistic words have nothing to do with the content of documents, and are not restricted to specific words, phrases, and part-of-speeches. Features of word dependencies can capture the essential semantic frames or patterns of sentences.

Authorship identification can be formulated as a multi-class categorization problem where the authors act as the class labels [6]. Hence, the second issue of the authorship identification task is the selection of classification methods. The Support Vector Machines (SVM) [21] method is a main classifier used in related works about identifying authorship [7,15,22,23]. Other classification methods include linear discriminant analysis (LDA) [24,25], decision trees [15], neural networks [15], and genetic algorithms [4]. Typically, in authorship identification [12,26], principal components analysis (PCA) [27] is used to reduce the dimensions of features derived from the occurring frequencies of the most frequent words. In addition, in [25,28], LDA is employed to learn the subspace of features used in authorship recognition of digital crime and registers.

In fact, PCA is an optimal linear representation of the data, and maintains the original information of the data to the greatest extent possible, and is not constrained by any parameter [27]. Further, PCA captures the descriptive features for dimension reduction. As a supervised subspace learning approach, LDA is able to generate a linear function which maximizes the difference between classes of data, and minimizes the difference within classes [29–34]. Thus, the goal of LDA is to extract the discriminant features for classification [15]. Currently, it becomes a powerful learning approach, and is popularly used in data classification [15]. Here our emphasis is to employ PCA and LDA to evaluate the discriminant power of the extracted features. In this paper, lexical, syntactic and structural features, and our proposed semantic association model about word dependencies, voice, and non-subject stylistic words will be evaluated on two public English text corpora. Comparative experimental results indicate that, with the help of our proposed features, the overall performance over authorship identification can be improved, and the performance using PCA and LDA reaches the highest accuracy in most cases.

The contributions of our work can be highlighted as follows:

(a) A semantic association model based on word dependency relations, voice, and non-subject stylistic words is proposed to represent the writing style of different authors. Moreover, we develop an unsupervised approach to extract these features. Features of the word dependencies capture the patterns of essential semantic structures of a sentence, namely, the configuration patterns of a predicate–argument structure and its subordinate semantic components. These features can be extracted as sentences with different words or different syntactic patterns may have the same patterns of semantic structures. In parallel, voice features can capture the configuration patterns of a predicate-verb and participants associated with this verb. Features about non-subject stylistic words are not indicators of text contents. Hence, those three types of semantic association features are confined neither to specific lexicons, phrases, and part-of-speeches, nor to specific domains, topics and contents of texts. Experimental results demonstrate that those semantic association features improve the overall performance of authorship identification.

(b) This paper develops a uniform vector space model to represent the abstract semantic patterns of sentences, and it can solve the problem of the independence of different dimensions to some extent. The language model of the context-free grammar is a set of rewriting rules about the grammatical categories and the specific words, which cannot represent the lexical and semantic dependencies between words in a sentence [35]. However, our vector space model is able to describe the characteristics of abstract patterns of semantic collocation relationships between different types of verbs and their different types of auxiliary words. Moreover, features of the word dependencies and voice capture the correlations between lexical and syntactic features.

(c) This paper offers a promising approach for authorship identification. Our experiments on two public corpora demonstrate that the identification performance with our proposed features by using PCA and LDA is better than those of KNN and SVM, better than that of the baseline approach, and also better than those of present features in related works.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents our authorship identification algorithm. Experiments and result analysis are given in

Section 4. We conclude this paper and discuss some issues in Section 5.

2. Related work

In this section, we will briefly review the related work about feature selection and authorship identification. The character features include fixed-length and variable-length n-grams on the character level [9,22,36,37]. The characteristics of character features are independent of languages, and do not need any natural language processing tools [1,37]. In literature, Houvardas and Stamatatos [22] used the most frequent n-gram character sequences of variable length as the text representation features.

In early works on authorship identification, researchers mainly used lexical features as style markers. Lexical features contain word frequencies [6,11,26], function word frequencies [10,38], first/s/third person pronoun count, short/long word count [39], vocabulary richness [2,8], and bi-grams and tri-grams on the word level [40]. Vocabulary richness means various measures about the number of different words and the total number of words of texts [1,2,8,24]. Kopple et al. [11] selected the 250 most frequent words to represent the writing style of nineteen century English books. Stamatatos [6] used the 1000 most frequent words as the style markers of texts of the newspaper. The stylistic features in the work of Madigan et al. [38] included function words, words, numbers, signs, and punctuations.

Syntactic feature extraction requires part-of-speech and phrase parsers. Tas and Gorur [2] applied 35 style markers to express the writing styles of articles, which are the vocabulary richness and the measures related to numbers of words, sentences, punctuations, and part-of-speech tags. Luyckx and Daelemans [13] used the frequency distribution of part-of-speeches, verb forms, function words, and content words.

In order to extract the semantic features, researchers utilized WordNet to extract synonyms and hypernyms of words, which are further employed to build features [17]. Typically, Argamon et al. [18] constructed functional features between certain words or phrases as style markers. Gamon [16] used the NLPWin system to extract grammar productions of sentences, binary semantic features, and semantic modification relations.

Structural features consist of sentence lengths, word lengths, phrase lengths, paragraph lengths and so on [14,15]. Application-specific features include features of content-specific keywords, features related to specific text genres such as the use of greetings and farewells, and features concerned with certain natural languages. Zheng et al. [15] selected some content-specific keywords as stylistic features of online messages.

Most works about authorship identification employ classifiers to identify authors of documents [41–44]. In addition, some works adopt a meta-learning, an automaton method, and a language model of context-free grammars to attribute authorship [11,35,45]. Lin and Zhang [45] developed a stochastic finite automaton to identify authorship. This automaton uses sequences of part-of-speeches of function words of sentences to represent the writing characteristics of an author. Those part-of-speeches contain adverb, auxiliary verb, pronoun, preposition, conjunction, interjection, and number. Koppel et al. [11] presented a meta-learning based approach by measuring the difference depth of the accuracy between different sets of features.

3. An authorship identification approach

In this section we will present the framework of our authorship identification algorithm, feature construction, and the classification technique.

The task of the authorship identification can be addressed as follows. Let A be a set of authors, T be a set of texts in which each text is written by an author in A. This collection can be formally described as follows:

$$\begin{cases} A = \{a_1, a_2, \ldots, a_r\} \\ T = \bigcup_{i=1}^{s} T_i, \quad T_i = \{t_{i1}, t_{i2}, \ldots, t_{ij_i}\}, \quad t_{ik} \in a_i, \quad 1 \leqslant k \leqslant j_i, \end{cases} \tag{1}$$

where $t_{ij} \in a_i$ means that a_i is the author of the text t_{ij}. Given an anonymous text t, the authorship identification task is to predict the most plausible author of t in A.

In this paper, we assume that authors have personal traits of language use which can be discovered in their writings [35]. Halteren et al. [46] demonstrated that a set of measurable characteristics of texts can be used to identify a given author. However, the language model of context-free grammars cannot express lexical, syntactic, and semantic relationships between words in a sentence [35]. To remedy this drawback, in this paper, we use the vector space model to represent the writing styles of texts. The goal of developing the vector space model is to capture various measurements of characteristics of texts, since multiple kinds of features can be grouped in a uniform quantified way.

3.1. Overview of our framework

Fig. 1 illustrates the framework of our authorship identification approach. The algorithm consists of four phases: text analysis, feature construction, dimension reduction, and author classification. The test texts are comprised of unstructured natural language texts written by multiple authors. Unstructured texts means textual unstructured data that either does not have a pre-defined data model or is not organized in a pre-defined manner. Examples of unstructured texts include books, journals, and web pages.

The first phase about text analysis includes sentence splitting, tokenizing, part-of-speech tagging, phrase parsing, word dependency relation parsing, pronoun identifying, function word identifying, non-subject stylistic word identifying, voice extracting, and tense extracting. We implement the former five functions based on tools of OpenNLP and Stanford typed dependency parser. In this work, we complete the rest five functions.

The second phase about feature construction is to build structural features, lexical features, syntactic features, features within the semantic association model from documents. After this step, each unstructured document is represented by a feature vector.

The third phase about dimension reduction employes PCA to reduce the number of dimensions of feature vectors and to derive principal stylistic features of documents. The fourth phase about author classification uses LDA to select the most discriminant stylistic features of documents and employ the 1-NN classifier to classify the authorship of documents.

3.2. Feature construction

All features used in our approach are partitioned into ten feature sets, denoted by $F_1, F_2, \ldots, F_{10}$, as shown in Table 1. The symbol "$\surd$" indicates that the corresponding feature set is proposed in this paper, while "$\times$" means that the corresponding feature set has been used in other works. As can be seen from Table 1, the three feature sets F_8, F_9 and F_{10} about voice, non-subject stylistic words, and word dependency relations are proposed in our work. The details about how these feature sets are constructed are given in this subsection.

Definition 1 (*Property predicate*). Let T be a collection of texts, W be a set of characters or words in T. We define a n-ary predicate in

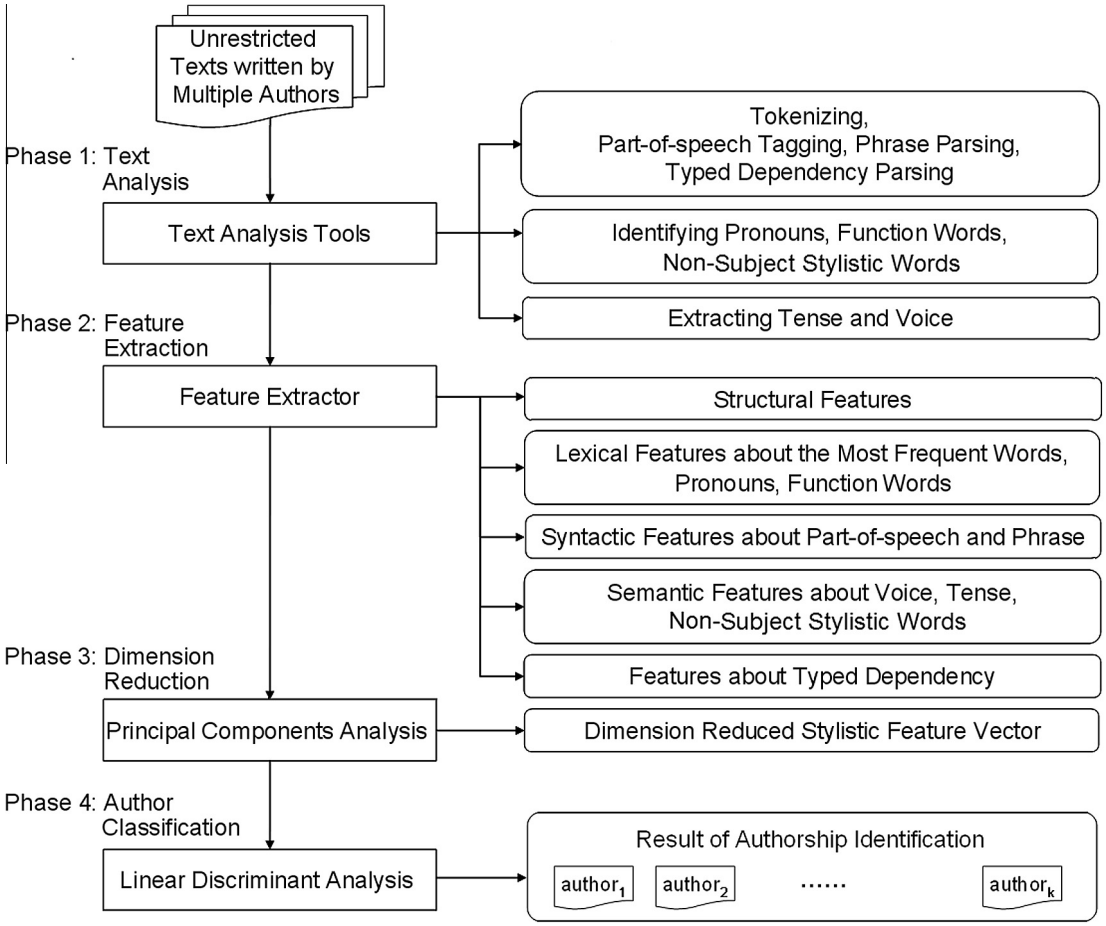

Fig. 1. The framework of our authorship identification algorithm.

Table 1
Feature sets used in our approach.

Category	Feature set	Functionality	Proposed in our work	Features
Structural	F_1	Measures about sentence lengths and words within sentences	×	Sentence lengths measured by numbers of characters and tokens
Lexical	F_2	The most frequent words	×	The frequency distribution of the most frequent words
	F_3	Pronouns	×	The frequency distribution of pronouns
	F_4	Function words	×	The frequency distribution of function words
Syntactic	F_5	Part-of- speeches	×	The frequency distribution of types of part-of-speeches
	F_6	Phrases	×	The frequency distribution of types of phrases
Semantic	F_7	Tense	×	The frequency distribution of types of tense
	F_8	Voice	√	The frequency distribution of types of voice
	F_9	Non-subject stylistic words	√	The frequency distribution of types of non-subject stylistic words
Syntactic and semantic	F_{10}	Word dependency	√	The frequency distribution of types of dependency relations

(2) to denote a property of a writing style of a text among characters or words $x_1, x_2, \ldots, x_n$, and call this predicate as a property predicate:

$$p(x_1, x_2, \ldots, x_n), \quad x_i \in W, \quad i = 1, 2, \ldots, n. \tag{2}$$

Definition 2 (*Feature transformation function*). Let $t \in T$ be a text, P collect the n property predicates of t, and V be a set of features of each property predicate in P. The feature transformation function f is a mapping from P to V:

$$f(p) = v, \tag{3}$$

where $p \in P$ is a property predicate, and $v \in V$ is a feature value of p.

Specifically, we let $P = \{p_1(x_{11}, x_{12}, \ldots, x_{1i_1}), \ldots, p_n(x_{n1}, x_{n2}, \ldots, x_{ni_n})\}$, and $V = \{v_1, v_2, \ldots, v_n\}$, where each p_i $(i = 1, 2, \ldots, n)$ is a property predicate of text t, v_i is the feature of p_i, x denotes a character or word, and the subscription (for example, $1i_1$) in each character or word is the index of x. Then, for the kth property predicate in P, we have

$$f(p_k(x_{k1}, x_{k2}, \ldots, x_{ki_k})) = v_k, \quad k = 1, 2, \ldots, n. \tag{4}$$

Definition 3 (*Independent feature*). If $p(x_1, x_2, \ldots, x_n)$ is a 1-ary predicate, then the feature that it maps to is called an independent feature.

Definition 4 (*Associated feature*). If $p(x_1, x_2, \ldots, x_n)$ is a n-ary($n \geqslant 2$) predicate, then the feature that it maps to is called an associated feature.

Definition 5 (*Explicit feature*). If $p(x_1, x_2, \ldots, x_n)$ denotes a property about characters or words, then the feature that it maps to is named as an explicit feature.

Definition 6 (*Implicit feature*). If $p(x_1, x_2, \ldots, x_n)$ expresses a property about texts involving a certain kind of parsing of texts, then the feature that it maps to is called as an implicit feature.

According to the above definitions, features in F_1, F_3, F_5, F_6, F_7, F_8 and F_{10} are associated features, while feature in F_2, F_4 and F_9 are independent features. Furthermore, features in F_2 and F_4 are explicit features, while features in F_1, F_3, F_5, F_6, F_7, F_8, F_9 and F_{10} are implicit features.

3.2.1. Structural level features

Different authors have different preferences to sentence lengths in their writings. For a document written by an author, we build the structural feature set F_1 about sentence length measured by the number of characters and words in a sentence. F_1 includes twenty-two features: average sentence character length lc_{avg} and word length lw_{avg}, maximum sentence character length lc_{max} and word length lw_{max}, minimum sentence character length lc_{min} and word length lw_{min}, average length of the top 10%, 20%, 80%, 90% sentences in terms of character lengths, i.e., lc_{top10}, lc_{top20}, lc_{top80}, and lc_{top90}, average length of the bottom 10%, 20%, 80%, 90% sentences in terms of character lengths, i.e., lc_{bot10}, lc_{bot20}, lc_{bot80}, and lc_{bot90}, average length of the top 10%, 20%, 80%, 90% sentences in terms of word lengths, i.e., lw_{top10}, lw_{top20}, lw_{top80}, and lw_{top90}, and average length of the bottom 10%, 20%, 80%, 90% sentences in terms of word lengths, i.e., lw_{bot10}, lw_{bot20}, lw_{bot80}, and lw_{bot90}. Accordingly, we can list the features in F_1 as follows:

$$\begin{aligned} F_1 = \{&lc_{avg}, lc_{max}, lc_{min}, lc_{top10}, lc_{top20}, lc_{top80}, lc_{top90}, lc_{bot10}, lc_{bot20}, \\ &lc_{bot80}, lc_{bot90}, lw_{avg}, lw_{max}, lw_{min}, lw_{top10}, lw_{top20}, lw_{top80}, \\ &lw_{top90}, lw_{bot10}, lw_{bot20}, lw_{bot80}, lw_{bot90}\} \end{aligned} \tag{5}$$

3.2.2. Lexical level features

Feature sets on the lexical level are composed of the set of the most frequent words, the set of pronouns, and the set of function words. For clarity, we collect them in sets F_2, F_3 and F_4, which are described in (6)–(8) and listed in Table 1. Accordingly, we have

$$F_2 = \{x | FreqFn(x) > \alpha\}, \tag{6}$$

$$F_3 = \{x | PronounFn(x) = 1\}, \tag{7}$$

and

$$F_4 = \{x | FunctionFn(x) = 1\}. \tag{8}$$

In (6), $FreqFn(x)$ denotes the occurring frequency of the word x in a test corpus, and α is a threshold. In (7) and (8), both $PronounFn(x)$ and $FunctionFn(x)$ are Boolean functions, which are computed as follows:

$$PronounFn(x) = \begin{cases} 1, & \text{if } x \text{ is a pronoun word} \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

and

$$FunctionFn(x) = \begin{cases} 1, & \text{if } x \text{ is a function word} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The feature set F_2 records the set of most frequent words in the corpus. In our work, the dimensions of F_2 will be taken as 250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000, respectively. Accordingly, the parameter α in (6) is automatically determined by the dimensions of F_2.

Pronouns are words that substitutes for nouns or noun phrases. Function words are used to describe grammatical relationships between words, and do not have lexical or semantic meanings [1]. Function words include articles (e.g. *an* and *the*), pronouns (e.g. *we* and *our*), auxiliary verbs (e.g. *will* and *should*), particles (e.g. *but* and *since*), expletives (e.g. *Ugh*), and pro-sentences (e.g. *yes*). Two lists of pronouns and function words come from the tool of OpenNLP Shallow Parser. The number of features in F_3 and F_4 are about 20 and about 310, respectively.

The reason that the features of the function words and pronouns are selected in our work can be explained as follows. Actually, these features are content-independent, are not constrained by text topics and genres, and can reflect preferences of different authors to the use of function words and pronouns. Accordingly, we build three sets of the most frequent words, function words and pronouns, and calculate the occurring frequencies of those words in a document as the features of this document.

3.2.3. Syntactic level features

The syntactic feature sets include the feature set F_5 of part-of-speeches and the feature set F_6 of phrase types, as described in (11), (12), and Table 1. That is, we have

$$F_5 = \{\beta | \exists x \in t,\ PosFn(x) = \beta\}, \tag{11}$$

and

$$F_6 = \{\gamma | \exists p \in t,\ PhrasetypeFn(p) = \gamma\}, \tag{12}$$

where function $PosFn(x)$ means the part-of-speech of the word x, and function $PhrasetypeFn(p)$ shows the phrase type of the phrase p.

Specifically, the types of phrases include NP (Noun Phrase), VP (Verb Phrase), ADJP (Adjective Phrase), ADVP (Adverb Phrase), PP (Prepositional Phrase), CONJP (Conjunct Phrase) and so on. The sum of number of features in F_5 and F_6 is about 50. The appearing frequencies of these types of syntactic information are obtained as the features of a document. Almost all languages have the lexical categories (i.e., part of speeches) and phrase categories, and lexical and phrase categories are the common features of those languages. Therefore, those syntactic features are independent of natural languages of documents, and can embody preferences of authors to the use of types of words and phrases [13]. These traits indicate why F_5 and F_6 are selected as the stylistic features.

3.2.4. A semantic association model

We propose a semantic association model to express the writing style of texts, which includes voice features, non-subject stylistic word features and word dependency features, and is intended to capture semantic stylistic characteristics of words and phrases and semantic stylistic relations between words and phrases.

Here, we build four semantic feature sets: the tense feature set F_7, the voice feature set F_8, the feature set F_9 of non-subject stylistic words, and the feature set F_{10} of word dependencies, as described in (13)–(15), (17), and Table 1. That is, we have

$$F_7 = \{\delta | \exists s \in t,\ TenseFn(s) = \delta\}, \tag{13}$$

$$F_8 = \{\varepsilon | \exists s \in t,\ VoiceFn(s) = \varepsilon\}, \tag{14}$$

and

$$F_9 = \{\eta |\ NonSubjectFn(x) = 1 \quad and \quad PosFn(x) = \eta\}, \tag{15}$$

where function $TenseFn(s)$ is the tense of the sentence s, function $VoiceFn(s)$ is the voice of s, and $NonSubjectFn(x)$ is a Boolean function. In addition, $NonSubjectFn(x)$ is calculated as follows:

$$NonSubjectFn(x) = \begin{cases} 1, & \text{if } x \text{ is a non-subject stylistic word} \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The tense feature set includes all kinds of verb tenses in English such as the simple present tense, the past perfect continuous tense, and the future perfect tense. The voice feature set contains two types of voices: active and passive. The number of features in F_7 is 12. The reason that F_7 and F_8 are chosen as stylistic features lies in that they are independent of specific words, phrases, and contents of texts.

Tenses and voices of sentences are identified based on the parsing results of word dependencies of sentences. First, we identify the predicate verbs and voices of sentences from the word dependency relations: *nsubj*, *csubj*, *nsubjpass*, and *csubjpass*. Then we recognize the tenses of sentences according to the syntactic characteristics of different tenses.

The non-subject stylistic words are to describe properties, states, grammatical relationships between words, but not objects and actions. Thus, those words are not closely related to specific topics of texts. In linguistics, those words can be adjectives, adverbs, pronouns, determiners, particles, prepositions, conjunctions or interjections, except for nouns and verbs. Those part-of-speeches (not including nouns and verbs) compose the feature set F_9. The feature set F_9 is a subset of F_5 in a sense of form. Totally, the number of features in F_9 is about 25. In statistic, the frequencies of features in F_9 reflect the using frequencies of the non-topic words independent of the topics and contents of texts. That is, the statistical characteristics of the features in F_9 are intended to capture the writing styles of authors across texts with different topics. In contrast, the frequencies of features in F_5 express the distribution of part-of-speeches of words in texts. We identify the occurring frequencies of all features in F_7, F_8 and F_9 as the features of a document. In summary, F_7, F_8 and F_9 can exhibit preferences of authors about the use of tenses, voices, and non-subject stylistic words.

The word dependencies construct a uniform relationship model for various relationships among any two words of sentences [19,20]. For example, in the hierarchal categories of word dependency relations, arguments consist of subject, agent, and complement. The argument of a subject includes nsubj (nominal subject), nsubjpass (passive nominal subject), csubj (clausal subject), and csubjpass (passive clausal subject). The argument of a complement is segmented into acomp (adjectival complement), attr (attributive), ccomp (clausal complement with internal subject), xcomp (clausal complement with external subject), complm (comlementizer), obj (object), mark (marker), and rel (relative). In addition, obj is further divided into dobj (direct object), iobj (indirect object), and pobj (object of preposition) [19].

The feature set F_{10} of word dependencies includes all word dependency relations between words in a sentence, which can be described as follows:

$$F_{10} = \{r |\ \exists\, x_i, x_j \in W_s,\ \exists\, r \in R,\ r(x_i, x_j) \text{ holds}\}, \tag{17}$$

where W_s is a set of words in a sentence, R is a set of word dependency relations, that is, R={nsubj, nsubjpass, csubj, csubjpass, agent,. . ., attr, ccomp, xcomp, complm}. The number of features in F_{10} is about 50. As an example, a sentence in the Reuters Corpus Volume 1 [47] and its parsing result of the word dependency relations is illustrated in Fig. 2. In this sentence, there are four predict verbs: *said*, *take*, *pocket*, and *deliver*. The argument *nsubj* of *said* is *Goldberg*, while the arguments *nsubj*, *dobj*, *ccomp* of take are *retailers*, *payment*, and *said*. The argument *dobj* of *pocket* is *money*, and *odobj* of *deliver* is *service*. These nine words constitute the semantic backbone of this sentence. Dependency relations *conj_or*, *conj_but*, and *conj_and* belong to the dependency relation *conjunct*, and *aux* belongs to *auxiliary*. Other dependency relations *amod*, *det*, *dep*, *prep_for*, and *prep_of* are members of the dependency relation *modifier*. All words involved in the *conjunct*, *auxiliary*, and *modifier* relations are the auxiliary components for predicate verbs and argument words.

With this example, we have illustrated that how the word dependency relations describe the relations within the predicate–argument structures of sentences. Actually, the word dependencies represent the relationships between pairs of words in a sentence on the syntactic and semantic levels. They also describe the patterns of the fundamental and the essential semantic structures of sentences. In our work, we employ the occurring frequencies of various types of dependency relations as the features of a document. Hence, these dependency features are irrelevant to specific words in sentences, independent of text topics and contents, and able to reveal preferences of authors about the use of semantic structures of sentences.

3.3. Feature reduction and authorship identification algorithm

After feature construction, PCA is introduced to reduce dimensions of feature vectors. Actually, PCA is able to reduce a possibly correlated high dimensional stylistic feature set into an uncorrelated lower dimensional feature set [27], and replaces the original

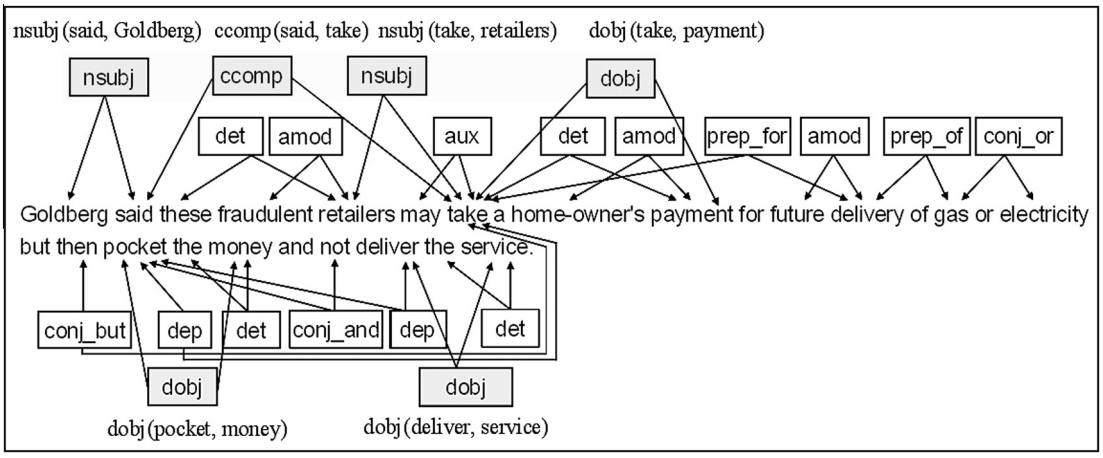

Fig. 2. A sentence and its parsing result of the word dependency.

features by the new uncorrelated features as *principal components*. Hence, the new features retain the characteristics of the original ones to the fullest extent possibly [27,34].

Further, LDA is employed to find a linear function of features to identify documents into different groups [48]. Finally, 1-nearest neighbor classifier is utilized to predict class labels to the groups, i.e., authors of those groups of documents.

Specifically, a text t_k is described by feature vector $\mathbf{v}_k = [v_{k1}, \ldots, v_{km}]^T \in \mathbb{R}^m$ after constructing features, where m is the total number of features in $F_1, F_2, \ldots$, and F_{10}. The task of PCA can be formulated as the following optimization problem:

$$\max_{\mathbf{W}\mathbf{W}^T=\mathbf{I}} \mathrm{tr}(\mathbf{W}^T\mathbf{U}\mathbf{W}), \tag{18}$$

where $\mathbf{W}$ is the project matrix, $\mathbf{U}$ is the covariance matrix of the features, $\mathbf{I}$ is an identity matrix, and $\mathrm{tr}(\cdot)$ stands for the trace operator of a matrix.

Algorithmically, problem (18) can be solved via the eigenvalue decomposition of $\mathbf{U}$. To reduce the dimensionality, $\mathbf{W}$ are usually determined by only containing the eigenvectors of $\mathbf{U}$ corresponding to the non-zero eigenvalues. After $\mathbf{W}$ is learned, each feature vector $\mathbf{v}_k$ is finally transformed as a new vector:

$$\tilde{\mathbf{v}}_k = \mathbf{W}^T\mathbf{v}_k. \tag{19}$$

After $\mathbf{v}_k$ is transformed by (19), LDA will be employed to learn a linear transformation matrix $\mathbf{X}$ such that the optimal separability is achieved in a subspace. For $\tilde{\mathbf{v}}_k$, it can be projected as a new vector $\mathbf{y}_k$ in subspace $\mathbb{R}^d$:

$$\mathbf{y}_k = \mathbf{X}^T\tilde{\mathbf{v}}_k, \tag{20}$$

where d is the reduced dimensionality. To learn the projection matrix $\mathbf{X}$, the within-class matrix and between-class matrix are defined as follows:

$$\mathbf{S}_w = \frac{1}{n}\sum_{i=1}^{r}\sum_{\mathbf{v}\in a_i}(\mathbf{v}^T - \mathbf{w}_i)(\mathbf{v}^T - \mathbf{w}_i)^T, \tag{21}$$

and

$$\mathbf{S}_b = \frac{1}{n}\sum_{i=1}^{r} n_i(\mathbf{w} - \mathbf{w}_i)(\mathbf{w} - \mathbf{w}_i)^T. \tag{22}$$

In (21) and (22), n is the number of training samples, r is the number of candidate authors, n_i is the number of documents in the corpus written by the author a_i, $\mathbf{w}_i$ is the centroid of the ith class, and $\mathbf{w}$ is the global centroid. Given $\mathbf{X}$, these two scatter matrices in a d-dimensional subspace can be calculated as follows:

$$\mathbf{S}_w^{(d)} = \mathbf{X}^T\mathbf{S}_w\mathbf{X}, \quad \mathbf{S}_b^{(d)} = \mathbf{X}^T\mathbf{S}_b\mathbf{X}. \tag{23}$$

The optimal rule in classic LDA is to minimize the within-class scatter and maximize the between-class scatter. Equivalently, LDA can be formulated as the following optimization problem:

$$\max_{\mathbf{X}\in\mathbb{R}^{m\times d}} \mathrm{tr}\left(\left(\mathbf{S}_w^{(d)}\right)^{-1}\cdot\mathbf{S}_b^{(d)}\right) = \max_{\mathbf{X}\in\mathbb{R}^{m\times d}} \mathrm{tr}\left(\left(\mathbf{X}^T\mathbf{S}_w\mathbf{X}\right)^{-1}\cdot\left(\mathbf{X}^T\mathbf{S}_b\mathbf{X}\right)\right), \tag{24}$$

where $\mathrm{tr}(\cdot)$ stands for the trace operator of matrix.

The optimal solution $\mathbf{X}$ to problem (24) can be obtained from the d eigenvectors of matrix $\left(\mathbf{S}_w^{(d)}\right)^{-1}\cdot\mathbf{S}_b^{(d)}$, associated to its d largest eigenvalues. After $\mathbf{X}$ is learned via problem (24), Eq. (20) will be used to transform each $\tilde{\mathbf{v}}_k$ into a d dimensional subspace with the maximum linear separability.

Finally, we point out that the above classic LDA algorithm may suffers from the small sample size problem when dealing with high dimensional data. In other words, the within-class scatter matrix $\mathbf{S}_w$ may become singular. This will make LDA difficult to be performed. Thus, in computation, a regularized algorithm will be implemented to avoid the possible singularity of $\mathbf{S}_w$. That is, the within-class scatter matrix $\mathbf{S}_w$ will be replaced by $\mathbf{S}_w + \lambda\mathbf{I}$, where λ is a regularization parameter, and $\mathbf{I}$ is an identity matrix.

Now we give our algorithm of authorship identification in Algorithm 1.

Algorithm 1. Authorship identification from unstructured texts

Input: The unstructured texts T of anonymous authors.
Output: the author of each text in T.
1: **for** $t_i, i = 1, 2, \ldots, n$
2: Make text analysis of t_i. That is, split t_i into sentences, tokenize words and phrases, tag part-of-speech, and identify pronouns, function words, and non-subject stylistic words.
3: Build word dependencies of sentences, and extract voices and tenses of sentences.
4: **end for**
5: Construct structural and lexical stylistic feature set F, $F = F_1 \cup F_2 \cup F_3 \cup F_4$.
6: Build syntactic and semantic feature set G, $G = F_5 \cup F_6 \cup F_7$.
7: Construct the semantic association model H of texts, $H = F_8 \cup F_9 \cup F_{10}$.
8: **for** $t_i, i = 1, 2, \ldots, n$ **do**
9: Extract stylistic features in $F \cup G \cup H$ from t_i.
10: Build the feature vector v of t_i.
11: **end for**
12: Apply PCA to reduce the number of dimensions of feature vectors of T.
13: Use LDA and 1-NN classifier to classify the authorship of each text.

4. Experiments

In this section, we report the experimental results of our approach. In addition, we will also analyze the influence of the number of feature dimensions and the training size on overall accuracy. Finally, we discuss the reasons why our approach works.

4.1. Corpora and evaluation

We use two text corpora in English in our experiments: English books and the Reuters Corpus Volume 1 (RCV1). The collection of 21 English books published in nineteenth century has been used in Koppel et al. [11]. These books were written by ten different authors. The RCV1 corpus is a publicly document collection about news stories [47]. It is the benchmark test corpus for text classification tasks, and recently has been utilized in authorship identification [22,23]. The RCV1 corpus includes documents with four main topics: CCAT (Corporate and Industrial), ECAT (Economics), CAT (Government and Social) and MCAT (Markets) [22,23,47].

As a text categorization measure, recognition accuracy is used to evaluate our experimental results. We use KNN, SVM and LDA to compare the performance of different feature sets. For SVM, the popularly-used package of LIBSVM [49] was used in our experiments. Note that there is a regularization parameter C in SVM. We use fivefold cross validation approach to select it from the candidate set $\{10^{-2}, 10^{-1}, \ldots, 10^4\}$. In experiments, the regularized LDA is performed with the regularization parameter λ, which is tuned via fivefold cross validation approach from the candidate set $\{0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0\}$. The selected parameters will be finally used in experiments.

4.2. Experimental results

The first group of experiments are conducted on the 21 English books. In experiments, each book is split into files with an approximate length of about 5000 bytes. Here, 50% files are randomly selected from the files of books of each author for training and the rest for testing. The average recognition rates and the standard deviations are reported over 20 random splits. The baseline approach is a combination of a character n-gram feature set F_{cg} (shown in (25)) and SVM without PCA feature extraction, where n is a positive integer, μ is a threshold. Here, F_{cg} is extracted as follows:

$$F_{cg} = \{x | FreqFn(x) > \mu,\ x \text{ is a character } n-\text{gram sequence}\} \tag{25}$$

To analyze the influences of the feature sets F_1, F_2, F_3, F_4, F_5, F_6, F_7 used in the existing works and the feature sets F_8, F_9, F_{10} proposed by us in this work, we build the following three combined features sets CF_1, CF_2 and CF_3:

$$CF_1 = \{F_1, F_2, F_3, F_4, F_5, F_6, F_7\}, \tag{26}$$

$$CF_2 = \{F_1, F_2, F_3, F_4, F_5, F_6, F_7\} \cup \{F_8, F_9, F_{10}\}, \tag{27}$$

and

$$CF_3 = \{F_1, F_2, F_3, F_4, F_5, F_6, F_7\} \cup \{F_8, F_{10}\}. \tag{28}$$

Table 2 lists the identification accuracy of KNN, SVM and LDA by using the features F_{cg1}, F_{cg2}, F_{cg3}, CF_1, CF_2 and CF_3. Here, F_{cg1}, F_{cg2} and F_{cg3} are 3-gram, 4-gram and 5-gram character feature sets. The dimensions of F_{cg1}, F_{cg2}, F_{cg3} and F_2 are all 2000. As seen from Table 2, with SVM, F_{cg1} generates the accuracy 97.52%, which is higher than those of F_{cg2} and F_{cg3}. Our feature set CF_2 reaches the 98.48% accuracy by using the LDA approach. In KNN, SVM and LDA, the accuracy of CF_3 and CF_2 is higher than those of CF_1. In addition, the identification accuracies of SVM without PCA by using F_{cg1}, F_{cg2}, and F_{cg3} are 97.02%, 96.24%, and 95.18% respectively. We can see that the performance of SVM with PCA in Table 2 is higher than that of SVM without PCA by using F_{cg1}, F_{cg2}, and F_{cg3}, respectively. The experimental results in Table 2 demonstrate that (1) LDA achieves higher identification accuracy than KNN and SVM on six feature sets; (2) our proposed features when combined with the features used in the existing works improve the accuracy of those existing features; and (3) The performance with our proposed features by using LDA reaches the highest accuracy.

In addition, we conducted the experiments about "leave-one-book-out" test referred in [11]. That is, for each book B in the corpus of the 21 English books, all books but B are employed for training while B is used for testing in the experiments. To this end, we run the following steps for each book to obtain the authorship identified result. (a) For book B, identify its author by our approach with the feature sets F_{cg1}, F_2 and CF_2, respectively. Here, the dimensions of the features in F_{cg1} and F_2 are all 2000. (b) Compute the numbers of the test samples belonging to this book (namely book B) which are now classified as the work of the first, the second, ..., and the tenth author by using the features in F_{cg1}. For clarity, we denote these numbers by $N_{1,1}, N_{1,2}, \ldots, N_{1,10}$. Analogously, we count $N_{2,1}, N_{2,2}, \ldots, N_{2,10}$ according to the experimental results obtained with F_2, and $N_{3,1}, N_{3,2}, \ldots, N_{3,10}$ with CF_2. (c) For $i = 1, 2, \ldots, 10$, calculate $N_i = N_{1,i} + N_{2,i} + N_{3,i}$. If N_i $(1 \leqslant i \leqslant 10)$ is the maximum, then the ith author is identified as the author of this book.

In the work of Koppel et al. [11], they obtained the following results that 19 of 20 same-author pairs and 181 of 189 different-author pairs are correctly classified. Here, the same-author pair of *Emily Bronte* and *Wuthering Heights* is excluded since the book Wuthering Heights is the author's only work, this pair cannot be tested. Within our 20 independent authorship identification experiments in which the experiment of testing the book Wuthering Heights is eliminated, 19 books are correctly classified. Moreover, we only employ the features in CF_1, CF_2, and CF_3 respectively (without considering the features in F_{cg1} and F_2) to identify the authors of the books, and 17, 16, and 17 books are correspondingly correctly identified.

The second group of experiments are conducted on the RCV1 corpus. In experiments, we choose the top 50 authors in terms of the number of documents with the CCAT topic. In each set of documents written by an author, the first 100 documents are selected, among which 50% randomly selected documents are used for training and the rest for testing. Table 3 reports the accuracy results of KNN, SVM and LDA by using F_{cg1}, F_{cg2}, F_{cg3}, CF_1, CF_2 and CF_3, respectively. In addition, the identification accuracies of SVM without PCA by using F_{cg1}, F_{cg2}, and F_{cg3} are 78.15%, 76.42%, and 74.72% respectively. Here the dimension of the features in F_{cg1}, F_{cg2}, F_{cg3} and F_2 are all 2000. Compared together within F_{cg1}, F_{cg2} and F_{cg3}, we see F_{cg1} generates the highest accuracy 78.67% with SVM. In contrast, with our extracted features in CF_2, we obtained the highest accuracy of 84.80% with LDA.

In order to investigate the influence of different authors on the accuracy results, the fifty authors are divided into five sets As_1, As_2, As_3, As_4, and As_5. To reduce the influence of the total size of texts of different authors, we perform the following partition criteria: (1) As_i (i = 1, 2, 3, 4, 5) has ten authors, and (2) The five sets S_1, S_2, S_3, S_4, S_5 have an approximately equal size, where S_i denotes all documents of the CCAT topic in the RCV1 corpus which were written by authors in As_i. Further, 5000 documents in our experiments are segmented into five data sets D_1, D_2, D_3, D_4, D_5, and documents in D_i were written by authors belonging to As_i. Thus D_i consists of 1000 documents of ten authors.

Table 4 lists the identification accuracy of KNN, SVM and LDA on D_1, D_2, D_3, D_4 and D_5 by using F_{cg1}, F_{cg2}, and F_{cg3}. The dimensions of F_{cg1}, F_{cg2}, and F_{cg3} are all 2000. The identification accuracies of SVM without PCA by using F_{cg1}, F_{cg2}, and F_{cg3} on D_1 are 91.35%, 89.42%, and 89.28%, respectively. The accuracies of the same method on D_2 are 87.86%, 88.42%, 85.86%, respectively; those on D_3 are 79.02%, 79.09%, 77.65%, respectively; those on D_4 are 76.97%, 76.62%, 74.89%, respectively; and those on D_5 are 83.79%, 83.86%, 82.88%, respectively. In the experiments via SVM with F_{cg1}, F_{cg2} and F_{cg3}, the highest accuracies on D_1, D_2, D_3, D_4, and D_5 are 91.82%, 88.99%, 79.73%, 77.68%, and 84.62%, respectively.

Table 2
The identification accuracy with KNN, SVM and LDA on English books.

	KNN (%)	SVM (%)	LDA (%)
F_{cg1} (3-gram)	79.63 ± 1.15	97.52 ± 0.38	97.65 ± 0.50
F_{cg2} (4-gram)	67.98 ± 1.32	96.79 ± 0.45	97.45 ± 0.34
F_{cg3} (5-gram)	48.77 ± 3.05	95.78 ± 0.60	95.84 ± 0.50
CF_1	68.70 ± 1.30	95.76 ± 0.63	98.32 ± 0.18
CF_2 (our)	71.29 ± 1.44	95.87 ± 0.71	98.48 ± 0.17
CF_3 (our)	69.75 ± 1.32	95.96 ± 0.62	98.45 ± 0.15

Table 3
The identification accuracy with KNN, SVM and LDA on the RCV1 corpus.

	KNN (%)	SVM (%)	LDA (%)
F_{cg1} (3-gram)	67.30 ± 0.78	78.67 ± 0.59	78.44 ± 0.69
F_{cg2} (4-gram)	64.00 ± 0.80	77.07 ± 0.76	77.23 ± 0.70
F_{cg3} (5-gram)	61.56 ± 0.93	75.41 ± 0.74	75.95 ± 0.94
CF_1	29.76 ± 0.65	69.55 ± 0.75	84.06 ± 0.69
CF_2 (our)	30.64 ± 0.65	67.94 ± 0.81	84.80 ± 0.72
CF_3 (our)	30.56 ± 0.78	67.83 ± 0.77	84.76 ± 0.74

Table 4
The identification accuracy on five data sets using n-gram character feature set.

No. of data sets	Feature sets	KNN (%)	SVM (%)	LDA (%)
D_1	F_{cg1} (3-gram)	82.92 ± 1.77	91.82 ± 0.94	90.76 ± 1.52
D_1	F_{cg2} (4-gram)	79.94 ± 1.16	90.16 ± 1.49	89.43 ± 1.64
D_1	F_{cg3} (5-gram)	76.40 ± 1.57	89.84 ± 1.15	89.30 ± 1.44
D_2	F_{cg1} (3-gram)	79.07 ± 1.50	88.37 ± 0.92	87.22 ± 1.98
D_2	F_{cg2} (4-gram)	76.00 ± 3.02	88.99 ± 1.20	87.63 ± 1.47
D_2	F_{cg3} (5-gram)	72.85 ± 2.18	86.53 ± 1.84	85.47 ± 1.94
D_3	F_{cg1} (3-gram)	64.48 ± 1.98	79.64 ± 1.54	80.54 ± 1.96
D_3	F_{cg2} (4-gram)	60.34 ± 2.84	79.73 ± 1.74	80.02 ± 1.83
D_3	F_{cg3} (5-gram)	56.40 ± 1.94	78.31 ± 1.80	78.60 ± 2.09
D_4	F_{cg1} (3-gram)	64.57 ± 2.61	77.68 ± 2.14	79.40 ± 2.49
D_4	F_{cg2} (4-gram)	59.29 ± 2.64	77.34 ± 2.18	78.32 ± 2.44
D_4	F_{cg3} (5-gram)	56.45 ± 3.08	75.64 ± 2.62	76.67 ± 2.77
D_5	F_{cg1} (3-gram)	73.42 ± 3.62	84.53 ± 2.30	85.02 ± 2.73
D_5	F_{cg2} (4-gram)	68.44 ± 3.70	84.62 ± 2.75	84.32 ± 2.63
D_5	F_{cg3} (5-gram)	66.91 ± 4.78	83.61 ± 2.29	83.17 ± 3.12

Table 5 reports the identification accuracy of KNN, SVM and LDA on D_1, D_2, D_3, D_4 and D_5 by using the features in CF_1, CF_2, and CF_3, where the dimensionality of the features in F_2 is 2000. As can be seen from Table 5, the accuracies obtained with KNN on five data sets by using CF_2 or CF_3 are higher than those of the same method by using CF_1. This fact also holds on the data sets D_1, D_2, D_4 and D_5 for SVM, and on the data sets D_1, D_2 and D_3 for LDA. We see the performance with our proposed features by using LDA on D_1, D_2, D_3, D_4, and D_5 achieves the accuracies of 95.30%, 91.37%, 87.11%, 78.4%, and 81.48%, respectively, which are higher than those obtained by the baseline method on D_1, D_2, D_3 and D_4. Those accuracies reach the highest value on D_1, D_2 and D_3 in Table 5.

Finally, to illustrate the statistical difference between our extracted features in CF_2 and CF_3 against the existing features in CF_1, F_{cg1}, F_{cg2}, and F_{cg3}, we did the paired student's t test on these two data sets (namely, the 21 English Book data set and the RCV1 data set). The hypothesis we test here is "the classification (mean) accuracy obtained by LDA with the features in CF_2 (or CF_3) is greater than that obtained with the other features in CF_1 (or F_{cg1}, F_{cg2}, and F_{cg3})". The statistic test is performed on the whole data set. Each test is run on two accuracy sequences, which are obtained from the 20 splits with our features and the existing features. Table 6 reports the results of the statistical tests. In each entity, "1" means that the hypothesis is correct (true) with probability 0.95, and "0" means that "the hypothesis is wrong (false)" with probability 0.95. For example, on the *English Book* data set (see Table 2), the decision "98.48 (CF_2 (Our)) > 97.65 (F_{cg1})" is correct with probability 0.95. For another example, on the *RCV1* data set (see Table 3), the decision "84.76 (CF_3 (Our)) > 78.44 (F_{cg1})" is correct with probability 0.95. In summary, from Table 6, we see the decision that "our algorithm achieves higher classification accuracy" is correct on the data sets.

Table 5
The identification accuracy on five data sets split from the RCV1 corpus.

No. of data sets	Feature sets	KNN (%)	SVM (%)	LDA (%)
D_1	CF_1	52.09 ± 1.53	84.12 ± 1.76	94.56 ± 1.13
D_1	CF_2 (our)	53.00 ± 2.18	83.89 ± 2.22	95.26 ± 1.18
D_1	CF_3 (our)	52.44 ± 2.00	84.50 ± 2.24	95.30 ± 1.19
D_2	CF_1	40.7 ± 2.31	73.97 ± 2.28	90.68 ± 1.58
D_2	CF_2 (our)	42.27 ± 2.02	73.40 ± 2.40	91.14 ± 1.54
D_2	CF_3 (our)	41.67 ± 2.01	74.06 ± 2.56	91.37 ± 1.56
D_3	CF_1	33.58 ± 1.63	66.35 ± 2.45	86.68 ± 1.45
D_3	CF_2 (our)	33.79 ± 1.35	66.08 ± 2.67	87.05 ± 1.55
D_3	CF_3 (our)	33.74 ± 1.37	66.07 ± 2.64	87.11 ± 1.64
D_4	CF_1	32.98 ± 2.09	54.03 ± 2.27	79.32 ± 2.46
D_4	CF_2 (our)	33.89 ± 2.30	53.71 ± 2.29	78.06 ± 2.70
D_4	CF_3 (our)	33.71 ± 2.52	54.22 ± 2.42	78.4 ± 2.7
D_5	CF_1	33.51 ± 1.49	56.31 ± 3.07	83.02 ± 1.85
D_5	CF_2 (our)	35.45 ± 1.75	56.23 ± 2.40	80.94 ± 1.84
D_5	CF_3 (our)	34.79 ± 2.01	56.42 ± 2.61	81.48 ± 1.88

Table 6
Results of the statistical significance test.

	$CF_2 > CF_1$	$CF_3 > CF_1$	$CF_2 > F_{cg1}$	$CF_3 > F_{cg1}$
Books	0	0	1	1
RCV1	1	1	1	1
	$CF_2 > F_{cg2}$	$CF_3 > F_{cg2}$	$CF_2 > F_{cg3}$	$CF_3 > F_{cg3}$
Books	1	1	1	1
RCV1	1	1	1	1

4.3. The influence of parameters

The influence of different sizes of F_{cg1}, F_{cg2}, F_{cg3} (i.e., the 3-gram, 4-gram, and 5-gram character feature set) is investigated in our experiments. The dimensions of F_{cg1}, F_{cg2}, F_{cg3} are set as 250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and

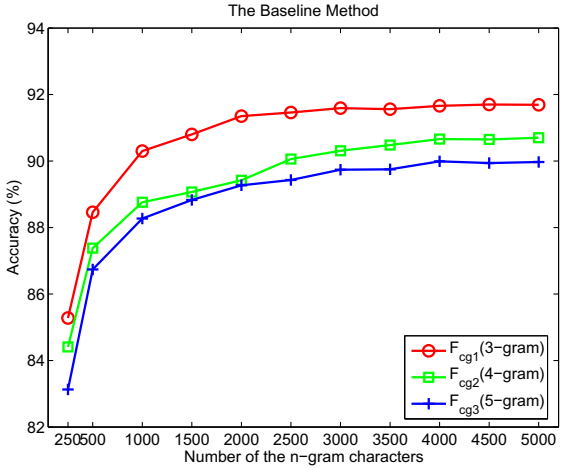

Fig. 3. The identification accuracy of different feature dimensions for the baseline method.

5000, respectively. Fig. 3 shows the accuracy curves of F_{cg1}, F_{cg2}, F_{cg3} using the baseline method. The curves in Fig. 3 show that the accuracy of F_{cg1} is higher than those of F_{cg2} and F_{cg3} in all eleven cases.

The feature set F_2 of the most frequent words in (6) is selected here to analyze the performance influence of different feature dimensions. The dimensions of F_2 in the feature sets CF_1, CF_2, CF_3 are set as 250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000, respectively. Fig. 4a–c shows that the accuracy curves of the feature sets CF_1, CF_2, and CF_3 using KNN, SVM and LDA, respectively. The curves in Fig. 4a–c show that (1) the accuracy of CF_2 is higher than that of CF_1 using KNN and LDA in all eleven cases, and this fact holds true for SVM in about half cases; and (2) the accuracy of CF_3 is higher than that of CF_1 by using LDA in all cases, and this fact holds true for KNN and SVM in most cases. Furthermore, we observe that the accuracy of CF_2 or CF_3 is higher than that of CF_1 by using KNN and LDA in all cases, and by using SVM in most cases. As a whole, the accuracy of our proposed features with LDA reaches the highest value in all cases. This means that LDA is more suitable than KNN and SVM to solve the problem of authorship identification.

The influence of different sizes of training data is also investigated in our experiments. We use 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of training documents to conduct the authorship identification task. Fig. 5a–c shows the accuracy curves of the features in CF_1, CF_2, CF_3 by using KNN, SVM and LDA, respectively. Here, the dimensions of F_2 is 2000. The curves in Fig. 5 indicate that the accuracy of CF_2 and CF_3 is higher than that of CF_1 by using KNN in all cases, while this fact holds for SVM and LDA in most cases. Within eight of the nine cases demonstrated in Fig. 5, the accuracy of our proposed features by using LDA reaches the highest value.

4.4. Algorithm analysis

The reasons that our approach works are given as follows. (1) The semantic association model actually captures the semantic stylistic characteristics "of and between" words. Features of the word dependencies reflect the configuration patterns of semantic structures of a sentence, namely, the constitutive laws of the predicate–argument structures and their associated semantic components. As an example of those laws, the number and the properties of the semantic components attached to different types of predicate verbs are relatively stable. Hence, those patterns are not restricted to specific lexicons, phrases, and part-of-speeches. Furthermore, sentences with different words or syntactic patterns may have the same patterns of semantic structures. (2) Features about non-subject stylistic words also capture the characteristics of non-topic words of texts. Thus, those features are not related

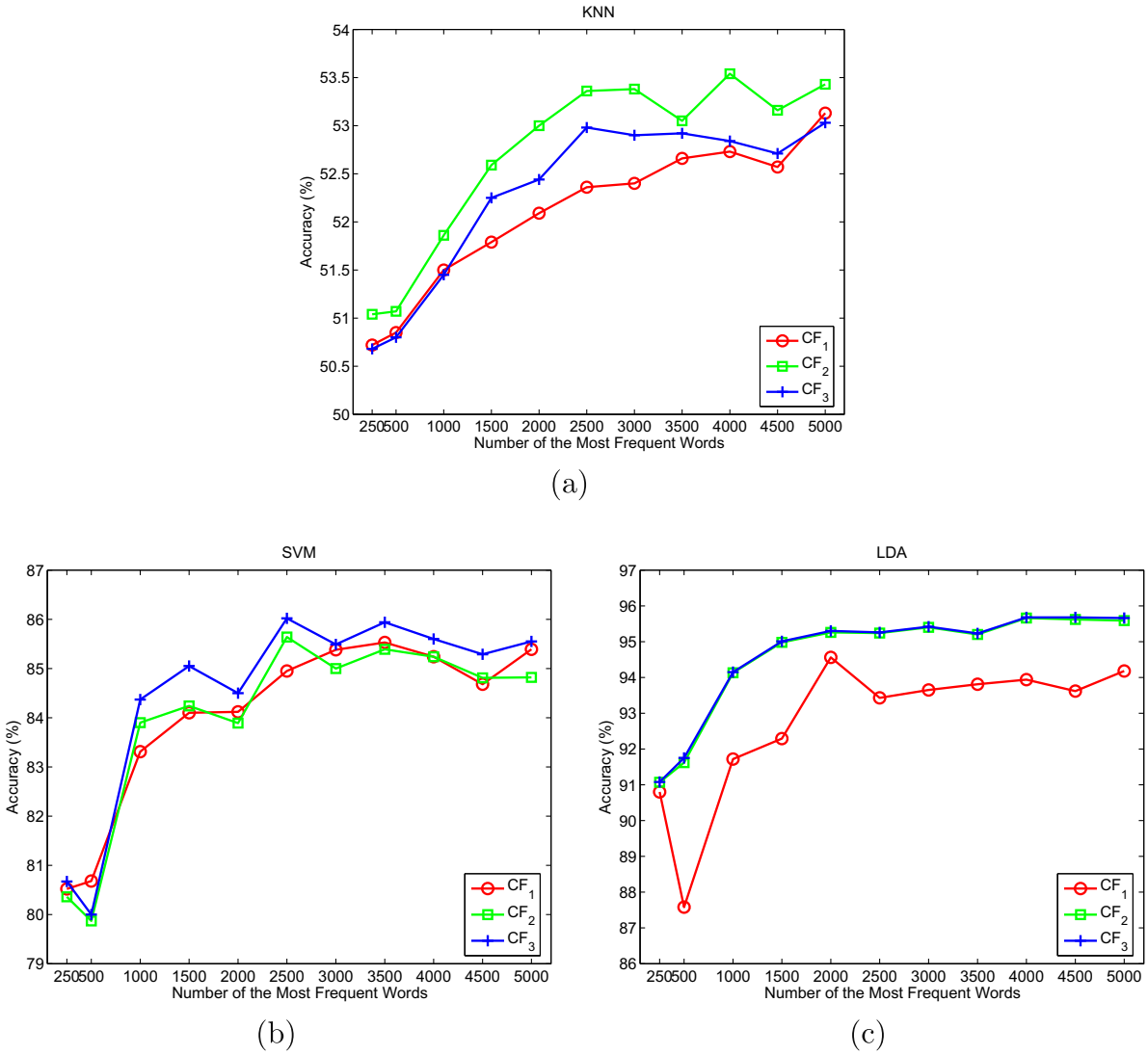

Fig. 4. The identification accuracy of different feature dimensions.

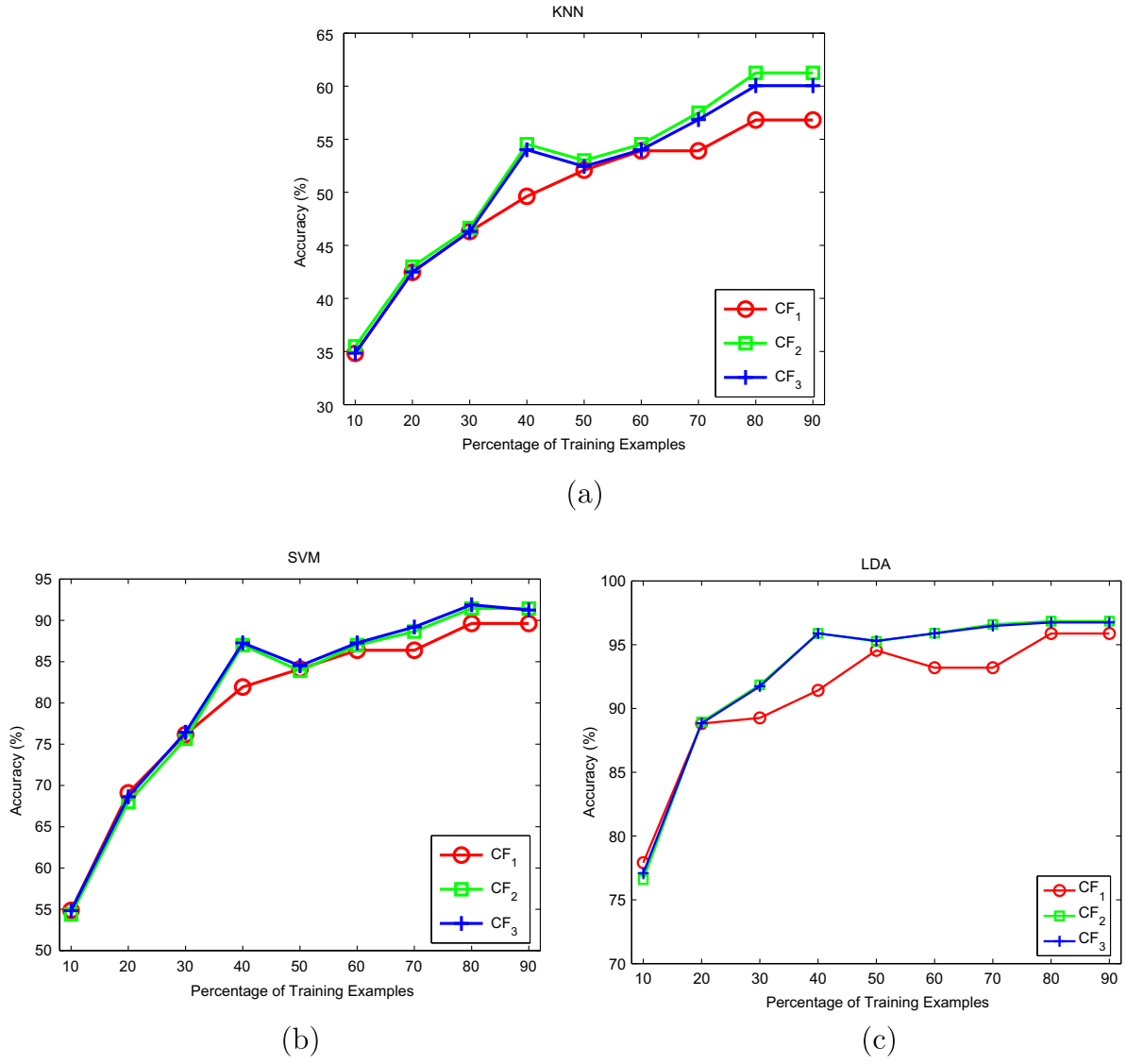

Fig. 5. The identification accuracy of different training size.

to a specific domain, topic, or genre. (3) The problems of the vector space model are the independence of dimensions and the very high number of dimensions [35]. First, the word dependency features and the voice features solve the problem of the independence of different dimensions to some extent. Those two kinds of features reflect the correlations between words which are included in the feature sets of the most frequent words, pronouns, function words, and non-subject stylistic words. Second, with the help of dimensionality reduction via PCA, the LDA approach handles the high-dimensional feature space well in our experiments. In conclusion, our approach has the scalability and portability, since the methods of feature representation, feature extraction and authorship identification in this paper can be applied to texts of any language, domain, topic and genre.

Finally, we analyze the computational complexity of our approach. The computational complexity of the LDA we used is about $O(ndt + t^3)$, where n is the number of samples, d is the number of features and $t = min(n, d)$.[1] In practice, the computational complexity of SVM for training scales in between $O(n)$ and $O(n^{2.3})$ [51], where n is the number of the training samples. In summary, as t is always less than n, the time complexity of our approach is slightly less than those developed via SVMs.

5. Conclusion

More and more attention has been paid on authorship identification in recent years for the sake of information security, copyright dispute, and public security and so on. In this paper, a semantic association model about word dependencies, voices, and non-subject stylistic words is proposed to represent the writing styles of unrestricted texts of different authors, and an unsupervised approach is designed to extract stylistic features. A classification technique of PCA, LDA, and KNN is employed to identify authorship of documents in order to build an optimal set of stylistic features for expressing styles of texts. The word dependencies can describe the essential semantic patterns of sentences, and features of word dependencies, voice, and non-subject stylistic words are independent of specific words, phrases, and part-of-speeches. Hence, all those features can be content-independent

[1] The training can be implemented in linear time complexity (details can be found in [50]).

and relatively easy to be kept across different documents of an author. We use PCA to select the descriptive features for dimension reduction, LDA to extract the descriptive features for subspace learning, and KNN to assign the authorship label. Comparative experimental results on two data sets show that the proposed features in combination with the classification method in this paper achieve a significant improvement of performance for the authorship identification task.

The main difficulties or challenges of authorship identification lie in the issues about language, genre, topic, stylistic features, and available documents. Factors affecting the accuracy of authorship identification mainly include the number of candidate authors, the size of each text, and the amount and types of training texts [1,35]. The difficulties are explained in detail as follows. (1) In theory, each author may have his or her own particular explicit and implicit writing styles. In practice, it is difficult for computers to extract stylistic characteristics of different types of texts, and determine authors of these texts. (2) In some cases such as criminal law and forensic applications, there only exists a small quantity of training documents. For all documents of an author, it is difficult to select texts with the appropriate number and size so as to adequately capture the writing styles of this author [1]. (3) In other application cases, the distribution of training texts over the candidate authors may be very different, or the true author of an anonymous text may not be contained in a set of predefined candidate authors [1].

In the future, we would like to develop models to address the above issues. In addition, we would also like to address the issues about the representation of pragmatic features and discourse writing features, and authorship identification for texts with small sizes such as microblogs.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant No. 61272361, 61370137, 61272169).

References

[1] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inform. Sci. Technol. 60 (3) (2009) 538–556.
[2] T. Tas, A.K. Gorur, Author identification for Turkish texts, J. Arts Sci. 7 (2007) 151–161.
[3] F. Iqbal, H. Binsalleeh, B.C. Fung, M. Debbabi, A unified data mining solution for authorship analysis in anonymous textual communications, Inform. Sci. 231 (2013) 98–112.
[4] D. Holmes, R. Forsyth, The federalist revisited: new directions in authorship attribution, Liter. Linguist. Comput. 10 (2) (1995) 111–127.
[5] G. Frantzeskou, S. MacDonell, E. Stamatatos, S. Gritzalis, Examining the significance of high-level programming features in source code author classification, J. Syst. Softw. 81 (2008) 447–460.
[6] E. Stamatatos, Authorship attribution based on feature set subspacing ensembles, Int. J. Artif. Intell. Tools 15 (5) (2006) 823–838.
[7] L. Dinu, M. Popescu, A. Dinu, Authorship identification of Romanian texts with controversial paternity, in: Proceedings of the Sixth International Language Resources and Evaluation, 2008, pp. 3392–3397.
[8] D. Hoover, Another perspective on vocabulary richness, Comput. Human. 37 (2003) 151–178.
[9] V. Keselj, F. Peng, N. Cercone, C. Thomas, N-gram-based author profiles for authorship attribution, in: Proceedings of Conference of the Pacific Association for Computational Linguistics, 2003, pp. 255–264.
[10] M. Koppel, J.S. Shlomo, A.E. Messeri, Authorship attribution with thousands of candidate authors, in: Proceedings of SIGIR, 2006, pp. 659–660.
[11] M. Koppel, J. Schler, E. Bonchek-Dokow, Measuring differentiability: unmasking pseudonymous authors, J. Machine Learn. Res. 8 (2007) 1261–1276.
[12] J.F. Burrows, Not unless you ask nicely: the interpretative nexus between analysis and information, Liter. Linguist. Comput. 7 (2) (1992) 91–109.
[13] K. Luyckx, W. Daelemans, Shallow text analysis and machine learning for authorship attribution, in: Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands, 2005, pp. 149–160.
[14] J. Grieve, Quantitative authorship attribution: an evaluation of techniques, Liter. Linguist. Comput. 22 (3) (2007) 251–270.
[15] R. Zheng, J. Li, H. Chen, Z. Huang, A framework for authorship identification of online messages: writing-style features and classification techniques, J. Am. Soc. Inform. Sci. Technol. 57 (3) (2006) 378–393.
[16] M. Gamon, Linguistic correlates of style: authorship classification with deep linguistic analysis features, in: Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 611–617.
[17] P.M. McCarthy, G.A. Lewis, D.F. Dufty, D.S. McNamara, Analyzing writing styles with coh-metrix, in: Proceedings of the Florida Artificial Intelligence Research Society International Conference, 2006, pp. 764–769.
[18] S. Argamon, C.P. Whitelaw, P. Chase, et al., Stylistic text classification using functional lexical features, J. Am. Soc. Inform. Sci. Technol. 58 (6) (2007) 802–822.
[19] M.D. Marneffe, B. MacCartney, C.D. Manning, Generating typed dependency parses from phrase structure parses, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation, 2006, pp. 449–454.
[20] M.D. Marneffe, C.D. Manning, Stanford Typed Dependencies Manual, 2008.
[21] V.N. Vapnik, The Nature of Statistical Learning Theory, second ed., Springer Verlag, New York, USA, 2000.
[22] J. Houvardas, E. Stamatatos, N-gram feature selection for authorship identification, in: Proceedings of the International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2006, pp. 77–86.
[23] E. Stamatatos, Using text sampling to handle the class imbalance problem, Inform. Process. Manage. 44 (2) (2008) 790–799.
[24] A. Honore, Some simple measures of richness of vocabulary, Assoc. Liter. Linguist. Comput. Bullet. 7 (2) (1979) 172–177.
[25] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis, D. Tambouratzis, Discriminating the registers and styles in the modern Greek language – part 2: extending the feature vector to optimize author discrimination, Liter. Linguist. Comput. 19 (2) (2004) 221–242.
[26] J. Binongo, Who wrote the 15th book of oz? An application of multivariate analysis to authorship attribution, Chance 16 (2) (2003) 9–17.
[27] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, New York, USA, 1986.
[28] C.E. Chaski, Who's at the keyboard? Authorship attribution in digital evidence investigations, Int. J. Digital Evidence 4 (1) (2005) 1–13.
[29] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.
[30] D. Foley, J. Sammon, An optimal set of discriminant vectors, IEEE Trans. Comput. 3 (1975) 281–289.
[31] L. Duchene, S. Leclerq, An optimal transformation for discriminant and principal component analysis, IEEE Trans. Pattern Anal. Machine Learn. 10 (6) (1988) 978–983.
[32] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, A new lda based face recognition system which can solve the small sample size problem, Pattern Recognit. 33 (10) (2000) 1713–1726.
[33] R. Duda, P. Hart, D. Stork, Pattern Classification, second ed., John Wiley and Sons, New York, USA, 2000.
[34] J. Yu, Q. Tian, T. Rui, T.S. Huang, Integrating discriminant and descriptive information for dimension reduction and classification, IEEE Trans. Circuits Syst. Video Technol. 17 (3) (2007) 372–377.
[35] P. Juola, Authorship attribution, Foundat. Trends Inform. Retriev. 1 (3) (2006) 233–234.
[36] F. Peng, D. Shuurmans, V. Keselj, S. Wang, Language independent authorship attribution using character level language models, in: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003, pp. 267–274.
[37] E. Stamatatos, Ensemble-based author identification using character n-grams, in: Proceedings of the 3rd International Workshop on Text-based Information Retrieval, 2006, pp. 41–46.
[38] D. Madigan, A. Genkin, D. David, et al., Author identification on the large scale, in: Proceedings of the Meeting of the Classification Society of North America, 2005, pp. 1–6.
[39] J. Karlgren, D. Cutting, Recognizing text genres with simple metrics using discriminant analysis, in: Proceedings of the 15th International Conference on Computational Linguistics, 1994, pp. 1–6.
[40] H. Halteren, Author verification by linguistic profiling: an exploration of the parameter space, ACM Trans. Speech Language Process. 4 (1) (2007) 1–17.
[41] Y. Sun, S. Todorovic, S. Goodison, Local-learning-based feature selection for high-dimensional data analysis, IEEE Trans. Pattern Anal. Machine Intell. 32 (9) (2010) 1610–1626.
[42] G. Brown, A. Pocock, M. Zhao, M. Lujan, Conditional likelihood maximization: a unifying framework for information theoretic feature selection, J. Machine Learn. Res. 13 (2012) 27–66.
[43] B. Quanz, J. Huan, M. Mishra, Knowledge transfer with low-quality data: a feature extraction issue, IEEE Trans. Knowl. Data Eng. 24 (10) (2012) 1789–1802.
[44] M. Wasikowski, X. Chen, Combating the small sample class imbalance problem using feature selection, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1388–1400.
[45] T.Y. Lin, S. Zhang, Anautomata based authorship identification system, in: Workshops with the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2008, pp. 134–142.
[46] H. van Halteren, R.H. Baayen, F.A. Tweedie, M. Haverkort, A. Neijt, New machine learning methods demonstrate the existence of a human stylome, J. Quant. Linguist. 12 (1) (2005) 65–77.

[47] D. Lewis, Y. Yang, T. Rose, F. Li, Rcv1: a new benchmark collection for text categorization research, J. Machine Learn. Res. 5 (2004) 361–397.
[48] A. Cord, F. Bach, D. Jeulin, Texture classification by statistical learning from morphological image processing: application to metallic surfaces, J. Microcopy 239 (2010) 159–166.
[49] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001. <http://www.csie.ntu.edu.tw/cjlin/limbos>.
[50] D. Cai, X. He, J. Han, Training linear discriminant analysis in linear time, in: IEEE 24th International Conference on Data Engineering, 2008, pp. 209–217.
[51] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods – Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999, pp. 185–208.