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.

© 2014 Elsevier B.V. All rights reserved.

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/© 2014 Elsevier B.V. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/402332

Download Persian Version:

https://daneshyari.com/article/402332

Daneshyari.com