A Profile-Based Authorship Attribution Approach to Forensic Identification in Chinese Online Messages

Jianbin Ma[1(✉)], Bing Xue[2(✉)], and Mengjie Zhang[2]

[1] College of Information Science and Technology,
Agricultural University of Hebei, Baoding 071001, China
majianbin@hebau.edu.cn
[2] School of Engineering and Computer Science,
Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
{Bing.Xue,Mengjie.Zhang}@ecs.vuw.ac.nz

Abstract. With the popularity of Internet technologies and applications, inappropriate or illegal online messages have become a problem for the society. The goal of authorship attribution for anonymous online messages is to identify the authorship from a group of potential suspects for investigation identification. Most previous contributions focused on extracting various writing-style features and employing machine learning algorithms to identify the author. However, as far as Chinese online messages are concerned, they contain not only Chinese characters but also English characters, special symbols, emoticons, slang, etc. It is challenging for word segmentation techniques to segment Chinese online messages correctly. Moreover, online messages are usually short. The performance for short samples would be decreased greatly using traditional machine learning algorithms. In this paper, a profile-based authorship attribution approach for Chinese online messages is firstly provided. N-gram techniques are employed to extract frequency sequences, and the category frequency feature selection method is used to filter common frequent sequences. The profile-based method is used to represent the suspects as category profiles. The illegal messages are attributed to the most likely authorship by comparing the similarity between unknown illegal online messages and suspects' profiles. Experiments on BBS, Blog, and E-mail datasets show that the proposed profile-based authorship attribution approach can identify the authors effectively. Compared with two instance-based benchmark methods, the proposed profile-based method can obtain better authorship attribution results.

Keywords: Profile · Authorship attribution · N-gram · Chinese · Online messages · Forensic

1 Introduction

The number of Internet users in the world reach 3.2 billion in 2015 [19]. Nowadays, Internet is an important information source in people's daily life. People

M. Chau et al. (Eds.): PAISI 2016, LNCS 9650, pp. 33–52, 2016.
DOI: 10.1007/978-3-319-31863-9_3

can communicate by various mediums such as E-mail, BBS, Blog, and Chat Rooms. Unfortunately, these online communication mediums are being misused for inappropriate or illegal purposes. It is common to find out fraud information, antisocial information, terroristic threatening information, etc. [25]. Terrorists make use of Internet to post messages for radicalization and recruitment of youth. In China, there is an institute called "12321" in charge of inappropriate or junk information complaint and treatment. Taking the statistics from the "12321" institute in September 2015 as an example, the institute received complaints about 9,438 junk mails, 28,407 illegal websites, and 22,481 illegal or junk text messages on mobile phones [1]. So, the inappropriate or illegal online information has strongly disturbed people's daily life.

The obvious characteristics of cybercrime is anonymous and borderless. In order to escape from detecting, criminals always forge their personal information or send information anonymously [8]. Moreover, by the help of pervasive network technology, criminals can hide in any corder at any time to commit crime. So, it is hard to identify the real authorship of inappropriate or illegal web information.

Computer forensic has been used as evidence since the mid-1980s [6], which is a science for finding legal evidence in computers or digital storage medias. Forensic investigation methods try to dig various types of evidence for courts. Fingerprints, blood, hair, witness testimony, and shoe prints are the traditional incriminating types of digital investigation methods [14]. But the traditional forensic investigation methods are inapplicable to cybercrime investigation. There is limited information in the crime scenes. Only some electronic text messages are available. However, like footprint and handwriting, authors' writing-style features can be mined by analyzing the authors' writing habits on Internet medias.

The purpose of authorship attribution is to attribute the authorship of unknown writings according to writing-style analysis on the author's known works [25]. There are some typical studies such as authorship attribution on Shakespeare's works [12,26] and "The Federalist Papers" [17,27]. Since 2000, authorship attribution on E-mail, BBS and Chatting room information for forensic purpose have drew researchers' attention [3,8,11,20,35]. Current research focus on two aspects: feature extraction and selection methods, and authorship attribution algorithms. Identifying the authorship by analyzing the writing-style features from Chinese online messages is difficult due to complicated linguistic characteristics, which are analyzed as follows.

(1) Chinese language has no natural delimiters between words. The word segmentation is a key technique to process Chinese natural language. Nowadays, some word segmentation softwares are available. However, word segmentation softwares are difficult to segment correctly for some neologies such as slang on the online messages. The terms TMD and 886 are commonly used slang in Chinese online messages. It is difficult to segment such kinds of slang by word segmentation softwares. Further, Chinese online messages sometimes contain English words, which are important features to mine authors' writing-style features. So, feature extraction methods that are language independent need to be investigated and developed.

(2) Chinese online messages contain complicated linguistic elements, which include not only Chinese characters but also English characters, special symbols and emoticons. The authors usually write freely. There are a lot of extra blank spaces or blank lines. Moreover, there are symbols such as "∼∼∼", "...." and "!!!", emoticons such as ☺ and ☹, and english characters such as "byebye" and "bye" in Chinese online messages. These characters, symbols and emoticons can be extracted as writing-style features to identify the authorship of online messages. So, authorship attribution methods on Chinese online messages need more effective feature extraction and selection methods to make better use of the online message information.
(3) In general, online messages are short. The texts of online messages contain few words. The classification accuracy for short text would be decreased greatly in traditional machine learning algorithms, such as K-nearest neighbors (KNN) and support vector machines (SVMs) [7].

From the above analysis, we can see that authorship attribution methods on Chinese online messages should be investigated, which should be language independent, can analyze extensive linguistic elements including Chinese characters, English characters, digits, symbols, emoticons, and are suitable for dealing with short online message samples. Profile is used to represent the training texts per author. Stamatatos, et al. (2009) had compared the profile-based and instance-based methods [32], and described that the important advantages of the profile-based methods were that the profile-based methods might produce a more reliable representation when the training samples are short texts such as E-mail messages and online forum messages. Keselj et al. (2003) [23] and Peng et al. (2003) [28] presented authorship attribution methods based on character-level n-gram, which was language independent. So, in this paper, we employ n-gram techniques which is language independent to extract frequent sequences from extensive linguistic elements in Chinese online messages. The profile-based method is used to represent the suspects as category profiles.

One factor that influences authorship attribution accuracy is class imbalance. Class imbalance is caused by uneven distribution of the training samples over the candidate classes. Some candidate classes have more training samples, while some other classes have fewer training samples. Moreover, the text lengths of training samples are different. Some candidate classes have longer texts than other candidate classes. Most previous authorship attribution approaches work well when the training samples are balanced, namely, equal number of training samples for each candidate class and the almost same length of training samples for each candidate calss. However, in most cases, the training samples over the candidate classes are not balanced. In such a situation, most of classifiers are biased toward the majority class [22]. Class imbalance problem will greatly reduce authorship attribution accuracy.

Class imbalance can cause frequent sequence imbalance in profiles. The number of frequent sequences in majority classes is more than that of minority classes. Further, the feature value of the frequent sequence in majority classes is larger than that of minority classes based on the n-gram based frequent sequence

extraction and selection method (shown in Sect. 3.2). So, in this paper, a frequent sequence standardization method is introduced to solve class imbalance problem.

1.1 Goals

In this paper, the overall goal is to propose a profile-based authorship attribution approach for Chinese online messages to effectively identify the author from a list of suspects and provide convictive evidence for cybercrime investigation. We will focus on the following four objectives in order to achieve the overall goal.

Objective 1: Employ n-gram techniques to extract frequent sequences from extensive linguistic elements including Chinese characters, English characters, digits, symbols, etc., and use the category frequency features selection method to filter common frequent sequences that do not have distinguished ability.

Objective 2: Develop a profile-based method to represent the suspects to category profiles, and present a similarity computing method to attribute unknown illegal messages to most likely authorship.

Objective 3: Develop a frequent sequence standardization method to solve class imbalance problem, and investigate whether the method is effective.

Objective 4: Propose a profile-based authorship attribution approach based on the above methods, and investigate whether this approach can obtain effective experimental results, and achieve better performance than two instance-based benchmark methods.

1.2 Organization

The rest of the paper is organized as follows. Section 2 reviews the previous contributions. Section 3 presents the proposed profile-based authorship attribution method. Section 4 describes the experiment design. Section 5 provides the experimental results and discussions. Section 6 are the conclusions and future work.

2 Related Works

Authorship attribution was used to attribute the authorship of literatures. The pioneering authorship attribution methods traced back to Mosteller and Wallace (1964) [27] who tried to attribute the authorship of "The Federalist Papers". Since the late 1990s, the vast amount of electronic texts (E-mail, Blog, Online forum, etc.) have appeared on the Internet. Authorship attribution studies have been used on forensic investigation [2,8,11,20]. In this section, we review n-grams writing-style features, authorship attribution method for forensic identification, and two typical writing-style feature representation methods.

2.1 N-Grams Writing-Style Features

In the views of computational linguistics, an n-gram is a contiguous sequence of n items extracted from a sequence of text. N-grams at the character-level were widely applied to authorship attribution. There were some successful applications. Forsyth and Homes (1996) [15] found that bigrams and charater n-grams achieved better performance than lexical features in authorship attribution. Keselj et al. (2003) [23] and Peng et al. (2003) [28] presented an authorship attribution method using character-level n-gram, which was language independent. Sun et al. (2012) [33] proposed an online writeprint identification framework using variable length character n-gram to represent the author's writing-style. The items of an n-gram can be characters and words according to the previous application. However, more extensive linguistic elements in Chinese online messages including Chinese characters, English characters, digits, symbols, emoticons should be extracted. In this paper, these linguistic elements are termed n-gram frequent sequences.

2.2 Authorship Attribution Method for Forensic Investigation

De (2000, 2001) [8–10] extracted a set of linguistic and structural features, and employed SVMs algorithm to attribute the authorship of E-mail documents. Iqbal et al. (2008, 2010) [20,21] mined frequent patterns for authorship attribution in E-mail forensic investigation. However, in the pre-processing phase, the spaces, punctuations, special characters and blank lines which are important information that can be used to mine authors' writing-styles are removed. Zheng et al. (2003, 2006) [35,36] presented an authorship attribution method that extracted a comprehensive set of syntactical features, lexical features, structural features, and content-specific features, and used inductive learning algorithms to build classification models. Chen (2008) [3] extracted a rich set of writing-style features and developed the Writeprints technique for identification and similarity detection on online messages. Ding et al. (2015) [11] proposed a visualizable evidence-driven approach based on an End-to-End Digital Investigation framework to visualize and corroborate the linguistic evidence supporting output attribution results.

2.3 Writing-Style Feature Representation Methods

The purpose of authorship attribution techniques is to form an attribution model by analyzing the writing-style features from the training corpus. Then, the attribution model attributes text samples of unknown authorship to a candidate author. The authorship attribution methods can be divided into two classes, namely, profile-based methods and instance-based methods. Profile-based methods extract a general style (called the author's profile) for each author from available training texts. Instance-based methods treat each text sample in the training set as an instance and extract a separate style for each text sample. Classification algorithms are often used to develop an attribution model in instance-based methods.

Profile-Based Authorship Attribution Methods. Keselj et al. (2003) [23] proposed an authorship attribution method that used n-grams to form author profiles. Iqbal et al. (2008) [21] extracted author profiles called write-prints based on frequent patterns extraction method. Estival et al. (2007) [13] presented an author profiling method with the application to Arabic E-mail authorship attribution. Stamatatos (2009) [32] had compared the profile-based and instance-based methods. He described that the important advantage of profile-based methods were that the profile-based methods might produce a more reliable representation when only short texts were available for training such as E-mail messages and online forum messages. Further, the computation time of profile-based methods was lower than that of instance-based methods.

Instance-Based Authorship Attribution Methods. In instance-based methods, machine learning algorithms such as neural networks, Bayesian, decision trees, KNN and SVMs were widely used to train an attribution model. Merriam and Matthews (1994) [26] employed neural network classifiers in the authorship attribution study. Kjell (1994a) [24] employed Bayesian and neural networks as classifiers. Hoorn et al. (1999) [18] extracted letter sequences for authorship analysis of three Dutch poets using neural networks. Holmes (1998) [16] used a genetic rule based learner for "The Federalist papers" authorship attribution problem by comparing the effects of vocabulary richness, and word frequency analysis.

Most authorship attribution studies focus on various pre-defined writing-style features and test different feature sets on effect of experimental results. Due to Chinese online messages' complicated linguistic characteristics (shown in Sect. 1), effective feature extraction, selection, and representation methods that are suitable for Chinese online messages authorship attribution should be investigated. Keselj et al. (2003) [23] and Peng et al. (2003) [28] proposed a language independent authorship attribution method using character-level n-gram language models. However, the method is restricted to character level, and merely applied to literature documents. They did not consider integrating extensive linguistic elements including Chinese characters, English characters, digits, symbols, emoticons, etc. So, in this paper, we propose a profile-based authorship attribution approach for Chinese online messages. Frequent sequences that are combinations of syntactic features, lexical features, and structural features are extracted by n-gram techniques from extensive linguistic elements. The frequent sequences are represented as category profiles. A similarity computing method is employed to attribute unknown illegal messages to most likely authorship.

3 The Proposed Profile-Based Authorship Attribution Approach

3.1 Overview

The process of the proposed profile-based authorship attribution approach for Chinese online messages, as shown in Fig. 1, can be divided into the following five steps.

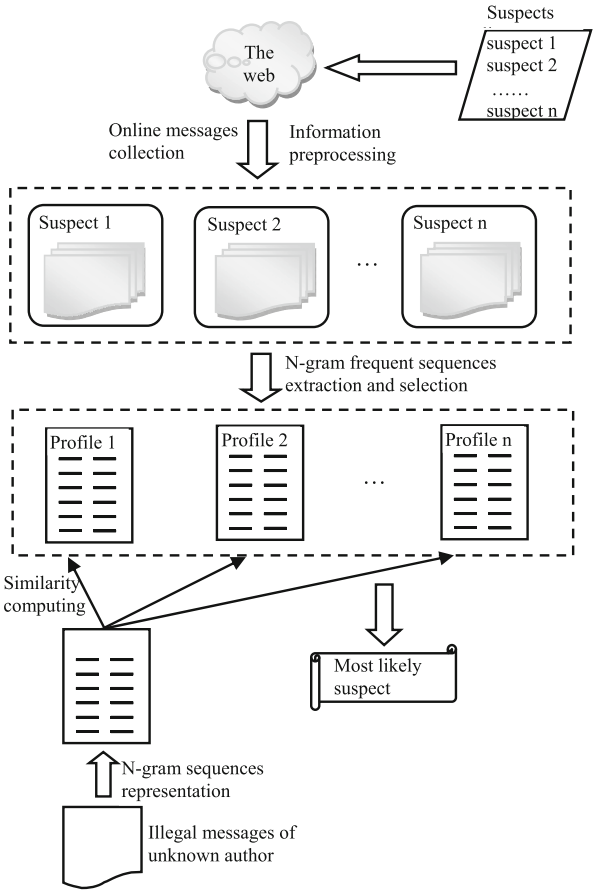

Fig. 1. The process of profile-based authorship attribution approach for Chinese online messages

Step 1. Preliminary investigation

Let's suppose there are illegal online messages that cause bad effect, and the Police Department intends to investigate who wrote these messages on the Internet. By preliminary investigation, investigators could narrow down the suspect list.

Step 2. Online messages collection

Each suspect's online message samples should be collected. These samples are suspect's known works. The goal here is to collect samples as many as possible. These online message samples are used as the training set.

Step 3. Online messages pre-processing

Some useless information such as advertisements, pictures, multimedia information should be removed. Text information and emoticons are reserved. In our

research, emoticons are represented by digital numbers (shown in Sect. 3.2). The blank spaces, blank lines, punctuations, and special symbols contain important information that can be used for mining suspects' writing-styles. This information is reserved too.

Step 4. Feature extraction and selection

In this paper, n-gram techniques are employed to extract frequent sequences from extensive linguistic elements in Chinese online messages. Category frequency is used to filter common frequent sequences. The profile-based method is adopted to represent the suspects as category profiles (details in Sect. 3.2).

Step 5. Authorship attribution

The unknown illegal online messages are represented by n-gram sequences. A similarity computing method is presented to compute the similarity between unknown illegal online messages and category profiles. Then, the messages are attributed to the most likely authorship that has the maximum similarity.

3.2 Author Profile Representation Method

There are some terms used in this paper. We define the terms as follows.

Definition 1 (Sequence element): A sequence element is one of the minimum linguistic elements in Chinese online messages that include Chinese characters, English characters, digits, symbols, etc.

Definition 2 (Frequent sequence): A frequent sequence is combinations of sequence elements and the number of the frequent sequences exceed the given support threshold.

Definition 3 (Profile): A profile is a virtual digital representation of a suspect's identity. In this paper, a profile is represented by a suspect's frequent sequences which are extracted and selected from the suspect's training set.

Let us suppose there is a suspect list $S = \{S_1, S_2, \cdots, S_n\}$, n is the number of suspects. There is an illegal message d that the author is unknown. The decision function is thus asked to map newly incoming illegal message d in one suspect from the suspect list (S), according to its content.

Profile-based classifiers derive a description of each target class (S_i) in terms of a category profile (C_i), usually a vector of features. These vectors are extracted from a training set $D = \{(d_1, y_1), (d_2, y_2), \cdots, (d_m, y_m)\}$ that is pre-categorized under C_i, where d_i is the feature vector of the ith sample, y_i is its label (i.e., category), and m is the number of samples. The profile-based classifiers can be referred as *category-centered* classification, which is thus the evaluation of similarity between unknown document d and different profiles (one for each class) [5].

Given a set of features $\{t_1, t_2, \cdots, t_u\}$ describing an online message $d_h \in D$, u is the number of features. The online message is represented as a feature vector, i.e.: $\boldsymbol{d_h} = \left\{\left(t_1, w_{t_1}^h\right), \left(t_2, w_{t_2}^h\right), \cdots, \left(t_u, w_{t_u}^h\right)\right\}$, where $w_{t_k}^h$ represents the feature value t_k for online message d_h. Category profiles of suspects are represented

as vectors of features, i.e.: $\boldsymbol{C_i} = \left\{\left(t_1, w^i_{t_1}\right), \left(t_2, w^i_{t_2}\right), \cdots, \left(t_u, w^i_{t_u}\right)\right\}$, where $w^i_{t_k}$ represents the feature value t_k in category profile C_i. In this paper, n-gram techniques are employed to extract frequent sequences as the writing-style features.

N-Gram Based Frequent Sequence Extraction and Selection Method. Each online message consists of a sequence of characters. The characters include Chinese characters, English characters, digits, punctuations, blank lines, blank spaces, special symbols, emoticons etc.

Based on the following rules, the sequence of characters is combined into sequence elements.

Rule 1: if the element of a sequence is a Chinese character, then the Chinese character is a sequence element.

Rule 2: if the element of a sequence is an English character, then read the next element. All the English characters are treated as a single sequence element until the element is not English character.

Rule 3: if the element of a sequence is a punctuation, then read the next elements. All the punctuations are treated as a single sequence element until the element is not punctuation.

Likewise, the elements of digits, blank spaces, blank lines, tab spaces are suitable for the rules 2 and 3.

An emoticon is represented as three-digit numbers such as <001> or <002>. The representation of an emoticon is treated as a sequence element.

Thus, based on the above rules, the extensive linguistic elements for Chinese online messages is transformed to sequence elements.

Then, the sequence elements are combined to 1-gram sequences, 2-gram sequences, 3-gram sequences, 4-gram sequences, etc. In the experiment section, experimental results show that 4-gram sequences are more effective than other n-grams (details in Sect. 5).

Most n-gram sequences occur in a certain category only once. These features are rare features which are either noninformative for category prediction or not influential in global performance [34]. We define frequent features that occur in the unique category frequently. The features are treated as frequent features on condition that the frequency of features exceeds the given frequent support threshold defined as "Minimum Term Frequency, *TFmin*", which is the smallest number of times a feature can appear in a category. The sequences that meet the condition of *TFmin* are extracted as features to remove the rare features.

If most category profiles containing one certain frequent sequence, the frequent sequence does not have distinguished ability. Filtering common frequent sequences among profiles and selecting effective frequent sequences are essential. We define the distinguish ability of one frequent sequence as category frequency (*CF*), which is the number of category profiles in which a frequent sequence occurs. We computed the *CF* for each unique frequent sequence in the training set and removed from the search space those frequent sequence whose *CF*

exceed the given minimum category frequency threshold (*CFmin*). Let us suppose all the unique frequent sequences among category profiles are represented as $T = \{(t_1, e_1), (t_2, e_2), \cdots, (t_v, e_v)\}$, where t_i is the ith frequent sequence, e_i is the total number of category profiles that contain the frequent sequence t_i, v is the number of unique frequent sequences. Formula (1) is the condition of filtering common frequent sequences.

$$\frac{e_i}{n} >= CFmin \tag{1}$$

where n is the number of profiles. The frequent sequences that satisfy the Formula (1) are removed.

The usual *TF-IDF* scheme is widely used in the vector space model [30]. In this paper, we improve the *TF-IDF* method and propose a new *TFC-ICF* method to compute the feature value of frequent sequences in the category profiles. *TFC* represents the number of a frequent sequence that occurs in the online messages of a certain category. The inverse category frequency (*ICF*) [4] is similar to the inverse term frequency (*IDF*) [31]. *ICF* is given by $ICF = \log\left(\frac{N}{F_{t_k}}\right)$, where F_{t_k} is the number of category profiles in which a frequent sequence $t_k \in \{t_1, t_2, \cdots, t_u\}$ occurs, N is the total number of categories. Then, the category profile feature values ($w^i_{t_k}$) are computed by the formula (2).

$$w^i_{t_k} = TFC \times ICF = \sum\nolimits_{h \in C_i} G^h_{t_k} \times \log\left(\frac{N}{F_{t_k}}\right) \tag{2}$$

where $G^h_{t_k}$ is the number of frequent sequence t_k that occurs in the online message d_h. C_i is the ith category profile.

Solving Class Imbalance Problems. In this paper, a frequent sequence standardization method to solve class imbalance problem is introduced. In category profile list $C = \{C_1, C_2, \cdots, C_n\}$, non-zero frequent sequences in each profile $\boldsymbol{C_i} = \left\{\left(t_1, w^i_{t_1}\right), \left(t_2, w^i_{t_2}\right), \cdots, \left(t_u, w^i_{t_u}\right)\right\}$ are sorted by the feature values in descending order. Then, every non-zero frequent sequence has its own ranking in the category profile. The total number of non-zero frequent sequences in each profile is counted. Let us suppose category profile C_i has the smallest number of non-zero frequent sequences, and the number is r. The number of non-zero frequent sequences in each profile is kept down to the same number r according to their ranking. The feature values of the rest of low ranking (lower than r) frequent sequences in profiles is set to 0. The feature values of non-zero frequent sequences in each profile are normalized to the range between 0 and 1.

The feature value is computed by the Formula (3).

$$w^i_{t_k} = \frac{r - s}{r - 1} \tag{3}$$

where $w^i_{t_k}$ is the feature value of non-zero frequent sequence t_k in the category profile C_i. s is the ranking of the non-zero frequent sequence t_k. r is the total number of the non-zero frequent sequences in profile C_i. $\frac{1}{r-1}$ is the interval between two non-zero frequent sequences.

3.3 Authorship Attribution

Let us suppose there is an illegal online message d. The message d is represented to character sequences, sequence elements and n-gram sequences. Ultimately, the message d is represented to $\boldsymbol{d} = \{(t_1, w_{t_1}), (t_2, w_{t_2}), \cdots, (t_u, w_{t_u})\}$, where $w_{t_i} = ICF \times G_{t_i}$, The computation formula of ICF is shown in Sect. 3.2, G_{t_i} is the number of feature t_i occurs in the message d. Formula (4) is the similarity function between unknown document d and category profile C_i.

$$sim(d, C_i) = \cos(\boldsymbol{d}, \boldsymbol{C_i}) = \sum_{k=1}^{u} \frac{w_{t_k} \cdot w_{t_k}^i}{|\boldsymbol{d}| \cdot |\boldsymbol{C_i}|} \tag{4}$$

where $sim(d, C_i)$ is the similarity degree between unknown online message d and category profile C_i, w_{t_k} is the feature value of sequence t_k in the message d, $w_{t_k}^i$ is the feature value of frequent sequence t_k in category profile C_i, u is the number of frequent sequences.

The most likely category of an unknown online message d is computed by Formula (5).

$$author\,(d) = \arg\max sim(d, C_i) \tag{5}$$

The unknown online message d is attributed to the maximum similarity between the unknown online message d and profile $C_i \in C$.

4 Experiment Design

In this section, experiments are designed to evaluate the performance of the proposed profile-based authorship attribution approach. The overall experimental objective is to verify whether our method can effectively identify the author from a list of suspects. Three experiments are performed to test the overall experimental objectives. (1) The first experiment is to test the influence of different parameters *TFmin* and *CFmin* on experimental results. (2) The second experiment is to test whether the accuracy is influenced by the class imbalance problem obviously, and show whether our frequent sequence standardization method is effective. (3) The third experiment is to compare the experimental results of the proposed profile-based approach and two benchmark methods, and test whether the proposed profile-based approach can achieve better performance than the two benchmarks.

4.1 Datasets

Three real-life datasets including BBS, Blog, and E-mail were collected. Table 1 shows the detailed information of the three datasets. There is no public Chinese datasets for online messages. So, we collected 10 most popular Bloggers on the website http://blog.sina.com.cn/ as the blog dataset. BBS dataset were gained from 6 active moderators on the Zhihu web forum http://www.zhihu.com/. Involving personal privacy, E-mails were collected from 5 staff members' mailboxes that were used to announce notifications on our university's mail server.

Table 1. The information of three datasets

Dataset	Authors	No. of instances	Average no. of words per instance
Blog	Author 1	200	273
	Author 2	200	926
	Author 3	200	354
	Author 4	107	336
	Author 5	200	413
	Author 6	200	461
	Author 7	200	184
	Author 8	183	289
	Author 9	197	610
	Author 10	200	227
BBS	Author 1	68	78
	Author 2	117	40
	Author 3	92	34
	Author 4	79	25
	Author 5	90	16
	Author 6	90	65
E-mail	Author 1	22	479
	Author 2	28	194
	Author 3	19	171
	Author 4	10	320
	Author 5	19	168

From the information of the three datasets in Table 1, we can see that the Blog dataset has more instances and the average number of words per instance is longer than other datasets. The average number of words per instance in the BBS dataset is fewer than that of other datasets. The class imbalanced problem in BBS and E-mail datasets are more obvious than Blog dataset.

4.2 Benchmarks for Comparison

To test the effectiveness of the proposed profile-based authorship attribution approach, two instance-based methods were selected as benchmarks for comparison.

There were little related studies aiming at authorship attribution method for Chinese online messages on forensic purpose except for our artificial intelligence and data mining research group in agricultural university of Hebei. So, we selected our previous instance-based authorship attribution method [25] as the first benchmark. We term the first benchmark as *Instance-lexi-stru* method. In the *Instance-lexi-stru* method, word segmentation softwares were used for word segmentation and part of speech tagging due to Chinese language's special

characteristics. The information gain method was used to select effective lexical features. The feature value of lexical features was calculated by the traditional *TF-IDF* Formula [29].

$$w(t, \boldsymbol{d}) = tf(t, \boldsymbol{d}) \times log(N/n_t + 0.01) \tag{6}$$

where $w(t, \boldsymbol{d})$ is the feature value of feature t in document d, $tf(t, \boldsymbol{d})$ is the frequency of feature t in document d, N is the total number of documents, and n_t is the number of documents that contain feature t.

Structural features include structural characteristics (shown in Table 2) [25], punctuations features (30 categories including Chinese and English punctuations), and part of speech features (12 categories part of speech features). An SVM are used as learning algorithm.

Table 2. Structural characteristics

Features
Number of distinct punctuations/total number of punctuations
Number of distinct words/total number of words
Mean sentence length
Mean paragraph length
Number of digital characters/total number of words
Number of lowercase letters/total number of words
Number of uppercase letters/total number of words
Number of space/total number of words
Number of blank lines/total number of lines
Number of indents/total number of words

In the second benchmark, we employed n-gram techniques to extract n-gram frequent elements (detail shown in Sect. 3.2). The information gain method was used to select effective n-gram frequent elements. The *TF-IDF* in Formula (6) is used to compute the feature value of n-gram frequent elements. An SVMs are used as learning algorithm. We term the second benchmark as *instance-n-gram* method.

In the experiments, the samples in each dataset were randomly divided into two sets, namely, 70 % as the training set, and 30 % as the test set. The accuracy was used to evaluate the experimental results.

5 Experimental Results and Discussions

5.1 The Influence of *TFmin* and *CFmin* on Experimental Results

To test the influence of different parameters *TFmin* and *CFmin* on experimental results, the first experiment was conducted. Different parameter combinations were experimented on the three datasets to find effective parameters combinations. The experimental results were shown in Table 3.

Table 3. The experimental results of different parameter combinations

Dataset	Parameters combinations		Accuracy (%)			
	TFmin	*CFmin*	2-gram	3-gram	4-gram	5-gram
Blog	2	0.5	84.18	87.03	87.50	86.87
		0.6	83.86	87.50	87.82	87.18
		0.7	85.76	87.82	88.13	87.34
		0.8	85.76	88.13	**88.29**	8.66
		0.9	85.92	87.82	88.13	88.13
	3	0.5	81.65	84.34	85.29	85.13
		0.6	82.60	85.44	85.44	85.13
		0.7	82.44	86.23	87.76	86.08
		0.8	82.75	84.97	86.23	86.08
		0.9	83.70	84.97	87.76	85.60
	4	0.5	80.22	82.91	82.91	82.75
		0.6	80.06	84.02	83.86	83.86
		0.7	80.54	83.70	84.81	84.34
		0.8	81.96	84.45	85.60	86.08
		0.9	81.65	84.97	85.13	85.13
BBS	2	0.5	69.60	70.64	69.80	69.05
		0.7	73.68	74.44	**75.39**	75.19
		0.9	69.63	71.85	70.37	71.11
	3	0.5	67.48	67.74	69.05	65.08
		0.7	59.56	59.26	60.45	61.19
		0.9	55.97	55.97	58.96	58.21
	4	0.5	0	0	0	0
		0.7	52.73	51.15	55.00	55.00
		0.9	41.84	42.17	43.02	47.71
E-mail	2	0.4	77.42	77.42	77.42	77.42
		0.6	80.65	80.65	**83.87**	80.65
		0.8	80.66	80.66	77.42	77.42
	3	0.4	55.00	61.62	77.78	73.42
		0.6	61.91	55.00	70.22	66.67
		0.8	56.57	64.29	64.29	70.74
	4	0.4	0	57.94	65.07	65.52
		0.6	53.15	53.15	65.00	61.67
		0.8	56.57	56.57	62.26	53.00

From Table 3, we can see that the experimental results of the 4-gram column are better than those of the 2-gram, 3-gram and the 5-gram columns in most cases, which suggests that the 4-gram frequent sequences is more effective than

other n-grams. With regard to parameter *TFmin*, the experimental results on the condition that the parameter value is 2 are better than those of other parameter values, which suggests that the sequences are frequent sequences when at least 2 documents in a certain category contain the sequences. As for parameter *CFmin*, the highest accuracy on the Blog dataset is 88.29 % on the condition that the parameter value of *CFmin* is 0.8, and the parameter value of *TFmin* is 2. There is little difference among the experimental results on blog dataset when the parameter value of *CFmin* is 0.7, 0.8 and 0.9. Likewise, the highest experimental result on the BBS dataset is 75.39 % on the condition that the parameter value of *CFmin* is 0.7, and the parameter value of *TFmin* is 2. The highest experimental result on the E-mail dataset is 83.87 % on the condition that the parameter value of *CFmin* is 0.6, and the parameter value of *TFmin* is 2. If the parameter value of *CFmin* is too high, the purpose to filter common frequent sequence would not be achieved. If the parameter value of *CFmin* is too low, some frequent sequences that have distinguished ability would be filtered away. So, the appropriate parameter value of *CFmin* should be set at a range of 0.6 to 0.8.

From Table 3, we can see that highest accuracy of Blog, BBS, and E-mail are 88.29 %, 75.39 %, and 83.87 % respectively. The accuracy exceeds 80 % by experimenting on the Blog and E-mail datasets. The accuracy of the E-mail dataset is high, because E-mail documents have obvious structural features such as greetings, farewells, and signatures, and the n-gram based frequent sequences feature extraction and selection method can represent the E-mail author's writing-styles well. The accuracy of the BBS dataset is relatively low, which might be caused by too few words in the BBS test set. Some BBS test samples even can not find the matching frequent sequences in any category profiles. The writing-styles in too short online messages are not obvious. Experimental results show that the authors can be identified from a list of suspects effectively, and promising performance is achieved.

5.2 The Experiments on Class Imbalance Problem

In this paper, a frequent sequence standardization method is used to deal with the class imbalance problem. To test the effectiveness of the frequent sequence standardization method, experiments were done on the Blog dataset. We changed the number of instances for authors randomly to produce the class imbalance problem. Firstly, the two authors' instances as shown in Table 1 were reduced by half. Then, the four authors' instances were reduced by half, and so on. The experimental results are shown in Table 4. We assume the parameter η denote the proportion of the number of authors whose instances are reduced by half in all the authors.

From the experimental results in Table 4, we can see that the accuracy does not change obviously. The accuracy decreases slightly when the two authors' instances are reduced by half. Then, the accuracy keeps stable when more authors' instances are reduced by half, which suggests that class imbalance problem has little effect on experimental performance and shows that the frequent sequence standardization method to solve class imbalance problem is effective.

Table 4. The experimental results of class imbalance situation for the Blog dataset

η	Accuracy (%)
0/10	88.29
2/10	85.76
4/10	84.65
6/10	84.86
8/10	85.13
10/10	84.49

5.3 Comparison of the Proposed Profile-Based Method with Two Benchmark Methods

In Sect. 4.2, two instance-based benchmarks (*Instance-lexi-stru* and *instance-n-gram*) are described. Each sample of the training set is treated as an instance and is extracted to a separate writing-style. The method is different from the proposed profile-based approach in this paper. We term the proposed profile-based approach as *Profile-n-gram*. To compare the performance of *Profile-n-gram*, *Instance-lexi-stru* and *Instance-n-gram*, experiments were made on three datasets. In instance-based methods, the kernel function was set to the linear kernel function in SVMs. 1000 lexical features were selected by the information gain features selection method. In *Profile-n-gram*, the parameters of *TFmin* and *CFmin* were set to optimal parameter combination based on the Sect. 5.1. The experimental results on different datasets are shown in Table 5.

From Table 5, we can see that the accuracy of the proposed *Profile-n-gram* method on each dataset is higher than that of the *Instance-lexi-stru* and *Instance-n-gram* methods. The accuracy of the *Instance-lexi-stru* on Blog and BBS datasets is higher than that of the *Instance-n-gram*. The accuracy of the

Table 5. The experimental results of *Profile-n-gram*, *Instance-lexi-stru* and *Instance-n-gram*

Dataset	Method type	Accuracy (%)
Blog	Profile-n-gram	88.29
	Instance-lexi-stru	82.06
	Instance-n-gram	80.15
BBS	Profile-n-gram	75.39
	Instance-lexi-stru	62.94
	Instance-n-gram	61.24
E-mail	Profile-n-gram	83.87
	Instance-lexi-stru	72.74
	Instance-n-gram	80.66

Instance-n-gram on the E-mail dataset is higher than that of the *Instance-lexi-stru*, because the E-mail dataset has obvious structural features, and n-gram feature extraction and selection method can mine the writing-style features effectively. The accuracy gap between the profile-based and the instance-based methods on the BBS dataset is wider than that of the Blog dataset, which might be caused by the characteristic of the two datasets. From Table 1, we can see that the length of samples in the Blog dataset is longer than that of the BBS dataset. The short samples would decrease the performance of the instance-based methods greatly [7]. The proposed profile-based methods can take full advantages of short online messages, and extract frequent sequences that almost can not be mined by traditional feature extraction methods [3,20,35].

Examples of frequent sequence extraction results for profiles on the BBS dataset is shown in Table 6. The left side of "=" is the frequent sequence. The right side of "=" is the feature value of the frequent sequence. Some fixed collocations in author's profiles include the combinations of Chinese words, English words, punctuations, modal words, digits, and slang words. The experimental results show that the n-gram based frequent sequence feature extraction methods are suitable for Chinese online message characteristics, and the proposed profile-based method is effective for solving authorship attribution problems with short online message samples.

Table 6. Examples of frequent patterns for profiles on BBS dataset

Authors	Examples of frequent patterns for different profiles
Author 1	大叔=0.91, 吗?=0.88, 呵呵..=0.75, 嘿嘿..=0.63,了?=0.63 了 =0.55, 呵呵 =0.34, 阿邦=0.34, 拜拜=0.34, 嫣然=0.34
Author 2	???=0.94, ...=0.92, 苹果=0.81, 了。=0.72, 的。=0.72 我们=0.64, 但=0.56, 球员=0.42, 这=0.42, iphone=0.19
Author 3	。=0.97, ?=0.91, ...=0.92,如果=0.67, 中国=0.56 然后=0.27,时候=0.34, 国民=0.27, 可能=0.27, 微博=0.15
Author 4	呵呵=0.92, 了....=0.86, 啊 =0.8, ～～～=0.72, 哈哈=0.72 这=0.72, 就=0.55, 挺好的=0.47, 嗯的=0.47, 空看=0.47
Author 5	～～～=0.97, 我=0.94, 呵呵=0.92, 帖子=0.89, 大哥=0.84 吧主=0.73, 哈哈=0.64, 不过=0.58, 真的=0.47, 谢谢!=0.19
Author 6	.....=0.98, bye-bye=0.86, hi=0.76, 但是=0.67, 游戏=0.59 不能=0.45, 还是=0.45, 科技=0.45, 开始=0.34, 那个=0.25

6 Conclusions

In this paper, a profile-based approach to authorship attribution for Chinese online messages was firstly proposed. N-gram techniques were employed to extract frequent sequence from extensive linguistic elements in Chinese online

messages, and category frequency was used to filter common frequent sequences. A novel frequent sequence standardization method was used to deal with class imbalance problem. The profile method was used to represent the authors as category profiles. A similarity computing method was employed to attribute illegal online messages to the most likely authorship.

To test the effectiveness of the proposed profile-based authorship attribution approach, Blog, BBS and E-mail datasets were collected. Experimental results showed that highest accuracy of Blog, BBS, and E-mail were 88.29 %, 75.39 %, and 83.87 %, respectively. Comparing with the two benchmark methods, the proposed profile-based authorship attribution approach achieved better performance than the two instance-based benchmark methods. The frequent sequence standardization method to solve class imbalanced problem was effective. This study showed that the proposed profile-based authorship attribution approach for Chinese online messages could identify the author from a list of suspects effectively, and present convictive evidence for cybercrime investigation.

Our future research will focus on the following directions. First, we intend to extend the writing-style features by semantical analysis on the training samples based on Chinese linguistic characteristics. Second, no standard benchmark datasets are currently available. To achieve more meaningful experimental results, we will try to collect more real-world datasets and experiment on those datasets. Third, we will investigate cybercrime forensic methods that combine our authorship attribution methods with other specific forensic technologies such as data recovery, tracing IP address, tracing log file, and social relation network investigation.

Acknowledgments. This work was supported by grants from Department of Education of Hebei Province(No.QN20131150), Program of Study Abroad for Young Teachers by Agricultural University of Hebei. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. 12321: 12321 statistics figures (2015). http://12321.cn/report.php
2. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intell. Syst. **20**(5), 67–75 (2006)
3. Abbasi, A., Chen, H.: Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. (TOIS) **26**(2), 1–29 (2008)
4. Basili, R., Moschitti, A., Pazienza, M.T.: A text classifier based on linguistic processing. In: Proceedings of IJCAI99, Machine Learning for Information Filtering. Citeseer, Stockholm, Sweden (1999)
5. Basili, R., Moschitti, A., Pazienza, M.T.: Robust inference method for profile-based text classification. In: Proceedings of JADT 2000, 5th International Conference on Statistical Analysis of Textual Data. Lausanne, Switzerland (2000)
6. Casey, E.: Digital Evidence and Computer Crime: Forensic science, Computers, and the Internet. Academic press, Cambridge (2011)

7. Chen, M., Jin, X., Shen, D.: Short text classification improved by learning multi-granularity topics. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 1776–1781. Citeseer, Barcelona, Spain (2011)
8. De Vel, O.: Mining e-mail authorship. In: Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (KDD 2000). Boston, USA (2000)
9. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. ACM SIGMOD Rec. **30**(4), 55–64 (2001)
10. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Multi-topic e-mail authorship attribution forensics. In: Proceedings of ACM Conference on Computer Security - Workshop on Data Mining for Security Applications. ACM, Philadelphia, PA, USA (2001)
11. Ding, S.H.H., Fung, B.C.M., Debbabi, M.: A visualizable evidence-driven approach for authorship attribution. ACM Trans. Inf. Syst. Secur. (TISSEC) **17**(3), 12 (2015)
12. Elliot, W., Valenza, R.: Was the earl of oxford the true shakespeare. Notes Queries **38**(4), 501–506 (1991)
13. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Tat: an author profiling tool with application to arabic emails. In: Proceedings of the Australasian Language Technology Workshop, Melbourne, Australia, pp. 21–30 (2007)
14. Fisher, B.A., Fisher, D.R.: Techniques of Crime Scene Investigation. CRC Press, Boca Raton (2012)
15. Forsyth, R.S., Holmes, D.I.: Feature-finding for text classification. Literary Linguist. Comput. **11**(4), 163–174 (1996)
16. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary Linguist. Comput. **13**(3), 111–117 (1998)
17. Holmes, D.I., Forsyth, R.S.: The federalist revisited: new directions in authorship attribution. Literary Linguist. Comput. **10**(2), 111–127 (1995)
18. Hoorn, J.F., Frank, S.L., Kowalczyk, W., van Der Ham, F.: Neural network identification of poets using letter sequences. Literary Linguist. Comput. **14**(3), 311–338 (1999)
19. ICT: Ict facts and figures (2015). http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx
20. Iqbal, F., Binsalleeh, H., Fung, B.C., Debbabi, M.: Mining writeprints from anonymous e-mails for forensic investigation. Digit. Invest. **7**(1), 56–64 (2010)
21. Iqbal, F., Hadjidj, R., Fung, B.C.M., Debbabi, M.: A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digit. Invest. **5**, S42–S51 (2008)
22. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
23. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING, vol. 3, pp. 255–264. Halifax Canada, (2003)
24. Kjell, B.: Authorship attribution of text samples using neural networks and Bayesian classifiers. In: Proceedings of IEEE International Conference on Systems. Man, and Cybernetics, vol. 2, pp. 1660–1664. IEEE, San Antonio, USA (1994)
25. Ma, J.B., Li, Y., Teng, G.F.: CWAAP: an authorship attribution forensic platform for chinese web information. J. Softw. **9**(1), 11–19 (2014)
26. Merriam, T.V., Matthews, R.A.: Neural computation in stylometry II: an application to the works of Shakespeare and Marlowe. Literary Linguist. Comput. **9**(1), 1–6 (1994)

27. Mosteller, F., Wallace, D.: Inference and Disputed Authorship: The Federalist. Addison-Wesley, Boston (1964)
28. Peng, F., Schuurmans, D., Wang, S., Keselj, V.: Language independent authorship attribution using character level language models. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics. vol. 1, pp. 267–274. Association for Computational Linguistics, Stroudsburg, USA (2003)
29. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets, vol. 77. Cambridge University Press, Cambridge (2011)
30. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
31. Sichel, H.S.: On a distribution law for word frequencies. J. Am. Stat. Assoc. **70**(351a), 542–547 (1975)
32. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. **60**(3), 538–556 (2009)
33. Sun, J., Yang, Z., Liu, S., Wang, P.: Applying stylometric analysis techniques to counter anonymity in cyberspace. J. Netw. **7**(2), 259–266 (2012)
34. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of Fourteenth International Conference on Machine Learning, vol. 97, pp. 412–420, Nashville, TN, USA (1997)
35. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. **57**(3), 378–393 (2006)
36. Zheng, R., Qin, Y., Huang, Z., Chen, H.: Authorship analysis in cybercrime investigation. In: Chen, H., Miranda, R., Zeng, D.D., Demchak, C.C., Schroeder, J., Madhusudan, T. (eds.) ISI 2003. LNCS, vol. 2665, pp. 59–73. Springer, Heidelberg (2003)