Information Sciences 231 (2013) 98–112

Contents lists available at ScienceDirect

Information Sciences

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

A unified data mining solution for authorship analysis in anonymous textual communications

Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung *, Mourad Debbabi

Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada H3G 1M8

ARTICLE INFO

Article history:
Available online 1 April 2011

Keywords:
Authorship identification
Authorship characterization
Stylometric features
Writeprint
Frequent patterns
Cyber forensics

ABSTRACT

The cyber world provides an anonymous environment for criminals to conduct malicious activities such as spamming, sending ransom e-mails, and spreading botnet malware. Often, these activities involve textual communication between a criminal and a victim, or between criminals themselves. The forensic analysis of online textual documents for addressing the anonymity problem called authorship analysis is the focus of most cybercrime investigations. Authorship analysis is the statistical study of linguistic and computational characteristics of the written documents of individuals. This paper is the first work that presents a unified data mining solution to address authorship analysis problems based on the concept of *frequent pattern-based writeprint*. Extensive experiments on real-life data suggest that our proposed solution can precisely capture the writing styles of individuals. Furthermore, the writeprint is effective to identify the author of an anonymous text from a group of suspects and to infer sociolinguistic characteristics of the author.

1. Introduction

Cyber criminals take advantage of the anonymous nature of the cyber world to conduct malicious activities such as phishing scams, identity theft, and harassment. In phishing scams, for instance, scammers send out e-mails and create websites to trick account holders into disclosing sensitive information such as account numbers and passwords. To solve these kinds of cybercrime cases, investigators usually have to backtrack to IP addresses based on information in the header of anonymous e-mail. However, identification based solely on the IP address is insufficient to identify the suspect (the author of the e-mail) if there are multiple users on the computer that sent out the e-mail, or if the e-mail was sent out from a proxy server.

The problem of anonymity in online communication is addressed by applying authorship analysis techniques. The study of authorship analysis has a long history in resolving authorial disputes over historic and poetic work [8,10]; however, the study of authorship analysis for online textual communication is very limited [12,47]. Traditional written works are voluminous and are usually well-structured following common syntactic and grammatical rules. In contrast, online documents such as e-mails and instant messages are short, poorly structured, and are usually written in *para language* containing several spelling and grammatical mistakes. These differences make some of the traditional works in authorship analysis not applicable to online textual data.

In this paper, we present a unified data mining approach to address the challenges of authorship analysis in anonymous online textual communication for the purpose of cybercrime investigation. Below, the term "text message" is broadly defined to include any textual communication, such as e-mails, blog postings, and instant messages. Specifically, this paper ad-

* Corresponding author. Tel.: +1 514 8482424x5919; fax: +1 514 8483171.
E-mail addresses: iqbal_f@ciise.concordia.ca (F. Iqbal), h_binsal@ciise.concordia.ca (H. Binsalleeh), fung@ciise.concordia.ca (B.C.M. Fung), debbabi@ciise.concordia.ca (M. Debbabi).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.03.006

dresses the following three authorship analysis problems, which were determined in a collaborative project with a Canadian law enforcement unit.

(1) *Authorship identification with large training samples:* A cybercrime investigator wants to identify the most plausible author of an anonymous text message from a group of suspects. We assume that the investigator has access to a large collection of messages that are previously written by suspects. In real-life investigation, the sample text messages can be obtained from the suspects' e-mail archives and chat logs on the seized personal computer(s), or from e-mail service providers with warrants. An investigator wants to precisely extract the writing styles of each suspect from the sample messages, use such patterns to identify the author of the anonymous message, and present such patterns as evidence to support the finding. Most of the previous works on authorship identification [1,14,15,23,47] assume that every suspect has only one writing style. We argue that a person's writing style may vary depending on the recipients or the topics. For example, when a student writes an e-mail, his writing style to a professor is probably different from his writing style to a friend. The challenge is how to precisely identify such stylistic variations and utilize the variations to further improve the accuracy of authorship identification.

(2) *Authorship identification with small training samples:* Given a collection of anonymous messages from a group of suspects, a cybercrime investigator wants to determine the author of each anonymous message in the collection. Unlike the previous problem, this problem assumes that the investigator has access to only a few training samples written by the suspects. In real-life investigation, the investigator can ask a suspect to produce a sample of his writing by listening to a story or watching a movie, then reproducing the played scene in his own writing. Clearly, the number of samples is very limited. The major challenge is how to identify the author of the anonymous messages when there are insufficient training data to build a classifier [1,47] or to extract any significant patterns.

(3) *Authorship characterization:* Given a collection of anonymous text messages, a cybercrime investigator sometimes has no clues about who the potential suspects are and, therefore, has no training samples of the suspects. Yet, the investigator would like to infer characteristics, such as gender, age group, and ethnic group, of the author(s) based on the writing styles in the anonymous messages. We assume the investigator has access to some external source of text messages such as blog postings and social network websites that disclose the authors' public profiles. The challenge is how to utilize such external sources to infer characteristics of the authors of the anonymous messages.

To address these authorship analysis problems, we propose a unified data mining approach that models the *writeprint* of a person [1]. The concept of writeprint, an analogy of a fingerprint in physical forensic analysis, is to capture the writing style of a person from his/her written text. Authorship studies [10,44] suggest that individual persons often leave traces of their personality in their written work. For instance, the selection of words, the composition of sentences and paragraphs, and the relative preference of one language artifact over another can help in identifying one individual from another. By capturing and analyzing the writeprint of anonymous text messages, an investigator may be able to identify the author of the messages from a group of suspects, or infer the characteristics of the author. The contributions of this paper are summarized as follows:

- *Frequent-pattern-based writeprint:* We precisely model the writeprint of a suspect by employing the concept of *frequent patterns* [5]. Intuitively, the writeprint of a suspect is the combination of stylistic features that are frequent in his/her text messages but not in other suspects' text messages. To ensure the uniqueness of the writeprint among the suspects, our approach ensures that any two writeprints among the suspects are disjoint, meaning that they do not share any frequent pattern. This is the first work that presents a unified data mining solution based on the frequent-pattern-based writeprint to address all three authorship analysis problems discussed above.
- *Capturing stylistic variation:* Our insight is that a person may have multiple writing styles depending on the recipients and the context of a text message. We present an algorithm to extract the writeprint and sub-writeprints of a suspect using the concept of frequent patterns. Experimental results suggest that the identification of sub-writeprints can improve the accuracy of authorship identification. Most importantly, the sub-writeprint reveals the fine-grained writing styles of an individual that can be valuable information for investigators or authorship analysis experts.
- *Analysis based on different training sample sizes:* Traditional authorship identification methods often require a reasonably large volume of training samples in order to build a classification model. Our proposed method is effective even if only a few training samples exist. If a training sample is not available, our approach can infer the characteristics of the authors based on stylometric features in the anonymous text messages.
- *Presentable evidence:* A writeprint is a combination of stylometric features that are frequently found in a suspect's text messages. Given that the concept is easy to understand, an investigator or an expert witness can present the writeprint and explain the finding in a court of law. Some traditional authorship identification methods, such as SVM and neural networks [41,47], do not have the same merit.
- *Removing burden from investigator:* One question frequently raised by a cybercrime investigator is how to determine the right set of stylometric features that should be used for the authorship analysis case in hand. Adding unrelated stylometric features can distort the accuracy of an analysis. Our notion of frequent-pattern-based writeprint resolves the problem because insignificant patterns are not frequent and, therefore, do not appear in the writeprint. Thus, an investigator can simply add all available stylometric features without worrying about the degradation of quality.

- *Clustering by stylometric features:* Our previous work [22] suggests that clustering by stylometric features is effective to identify anonymous e-mails written by the same authors; however, Ref. [22] does not show how to make use of the clustering results for further authorship analysis. To address the problem of authorship identification with few training samples, we propose to first cluster the anonymous e-mails by stylometric features and then match the writeprint between the training samples and the anonymous e-mail clusters. Our experimental results suggest that this approach is effective for authorship identification even with small training sample size.

The rest of the paper is organized as follows: Section 2 reviews the state-of-the-art in authorship analysis. Section 3 formally defines the authorship analysis problems and the notion of writeprint. Section 4 describes our unified data mining approach of authorship analysis in details. Section 5 evaluates our proposed method on real-life dataset. Section 6 concludes the paper.

2. Literature review

Authorship analysis is the study of linguistic and computational characteristics of the written documents of individuals [8,10]. Writing styles or specific writing traits extracted from an individual's previously written documents can be used to differentiate one person from another [31]. The writing styles can be broadly categorized into five different types of stylometric features, namely lexical, syntactic, structural, content-specific, and idiosyncratic features [1,23].

Authorship analysis has been very successful for resolving authorship identification disputes over literary and conventional writings [29]. However, analysis of online documents is more challenging [13]. Online documents are relatively short in size, resulting in insufficient training data to learn about the writing patterns of an author. Furthermore, the writing style of online documents is informal, and people are not conscientious about spelling and grammatical mistakes in informal chat and instant messages. Consequently, techniques that are very successful in literary and traditional works are not applicable to online documents. Therefore, it is imperative to develop analytical techniques that are suitable for online documents. Authorship is applied to e-mails [13,41], web forums [33], chat logs [28], and Web postings [3].

The detailed survey on authorship studies by Stamatatos [39] and Koppel et al. [27] show that the authorship problem is studied from three main perspectives: authorship identification, authorship similarity detection, and authorship characterization.

Authorship identification is applied to an anonymous document to identify the most plausible author from a group of suspects. In most studies, a classification model is developed by using the stylometric features extracted from a set of sample documents written by the suspects, and then is applied to the anonymous document to determine the most plausible author. These traditional classification approaches assume that the writing style of an author is consistent regardless of the subject-matter of a document and the type of target recipient. We argue that this assumption may not hold because the writing style of an author may be different depending on the target recipient [12]. Thus, we study this problem in Section 3.1 and propose a data mining approach to capture different writing styles of an author in Section 4.2.

Authorship similarity detection is used to determine whether or not the given two objects are produced by the same entity, without knowing who the entity is [1]. There are several applications of similarity detection including plagiarism detection and online marketplace. The digital revolution has greatly simplified the ability to copy and distribute creative works, which has led to increased copyright violation worldwide [42]. Similarly, the reputation system of online marketplaces, built by using customers' feedback, is most often manipulated by entering the system with multiple names (aliases) [16]. Abbasi and Chen [3] have developed techniques for detecting aliases in an online systems (e.g., eBay) by analyzing the users' feedback. Abbasi et al. [2,4] have developed similarity detection techniques for identifying malicious and fraudulent websites.

Some studies address authorship similarity as a *authorship verification* problem in which the task is to confirm whether or not a suspect is the author of a document in question. In traditional verification studies, a classification model is developed on a suspect's sample documents, which is then employed to verify if the given anonymous document was written by the suspect. Our previous work [24] borrows the NIST'speaker recognition evaluation framework to address the problem. Our recent work [22] has verified that clustering by stylometric features can effectively identify the messages written by the same author and, therefore, is also applicable to similarity detection.

Authorship characterization [12,27] is used to collect sociolinguistic attributes, such as gender, age, occupation, and educational level, of the potential author of an anonymous document. Corney et al. [12], Koppel et al. [26,27], and Argamon et al. [7] study the effects of gender-preferential attributes on authorship analysis. Other profiling studies discuss educational level [12], age, language background [27], and neuroticism [6]. Neuroticism is the tendency of a person to experience negative emotional states such as anxiety, anger, or guilt.

Machine learning techniques employed in most authorship analysis studies fall into three main categories: (1) probabilistic classifiers (e.g., Bayesian classifiers [36] and its variants); (2) decision trees [34]; and (3) support vector machine (SVM) [25] and its variants. Each of these techniques has its own limitations in terms of classification accuracy, scalability, and interpretability. An extensive survey on text categorization [37] suggests that SVM outperforms other classifiers, such as decision tree methods [35], the probabilistic naive Bayes classifier, and batch linear classifiers (Rocchio). However, SVM is a black box method and it is very difficult to interpret the reasons for reaching a conclusion; therefore, SVM is not suitable for evidence collection and presentation, which are important steps in cyber forensic analysis.

Feature selection is viewed as an important preprocessing step in the area of machine learning and data mining [21]. In authorship studies too, one may apply different feature selection techniques [20,21,30] to determine a subset of stylometric features that can discriminate the authors. Feature selection has two general approaches [38]: *Forward selection* starts with no features and, at each step, adds the feature that decreases the error the most until any further addition does not decrease the error significantly. *Backward selection* starts with all the features and, at each step, removes the one that decreases the error the most until any further removal increases the error significantly. These approaches consider only one attribute at a time. In contrast, our proposed approach employs the notion of frequent stylometric patterns that capture the combined effect of features. Irrelevant features will not be frequent in our approach. Thus, there is no need to apply feature selection. More importantly, feature selection does not guarantee the property of uniqueness among the writeprints of suspects.

3. Three authorship analysis problems

In this section, we formally define three authorship analysis problems described in the introduction section. The problems are carefully chosen to cover the most commonly encountered scenarios of authorship analysis in cybercrime investigation.

3.1. Authorship identification with large training samples

The first problem of authorship analysis is to identify the most plausible author S_a of a given anonymous text message ω from a group of suspects $\{S_1,\ldots,S_n\}$, with large sets of sample text messages M_i from each suspect S_i. In real-life investigation, the sample text messages can be obtained from the suspects' e-mail archives and chat logs on the seized personal computer(s), or from the e-mail service providers with warrants. An investigator wants to precisely extract the writing patterns of each suspect from the sample messages, use such patterns to identify the author of the anonymous message, and present such patterns as evidence to support the findings. Intuitively, a collection of text messages M_i *matches* an anonymous message ω if M_i and ω share similar patterns of stylometric features called *writeprint* [1]. The writeprint of a suspect *uniquely* represents the stylometric patterns of a suspect S_i, but does not represent the stylometric patterns of *any* other suspect S_j, where $i \neq j$. Below, we formally define the notions of stylometric patterns and writeprint [23].

3.1.1. Stylometric patterns

The stylometric patterns in a set of text messages M_i written by suspect S_i are a combination of stylometric feature items that *frequently* occurs in M_i. We concisely model and capture such frequently occurring patterns by the concept of *frequent itemset* [5], described as follows.

Let $U = \{f_1,\ldots,f_z\}$ denote all *stylometric feature items*. Let M_i be a set of text messages where each message $m \in M_i$ is represented as a set of stylometric feature items such that $m \subseteq U$. A text message m contains a stylometric feature item f_j if the numerical feature value of the text message m falls within the interval of f_j. The writing style features of some sample messages can be represented as vectors of feature items in Table 1.

Let $F \subseteq U$ be a set of stylometric feature items called a *stylometric pattern*. A text message m *contains* a stylometric pattern F if $F \subseteq m$. A stylometric pattern that contains k stylometric feature items is a *k-pattern*. For example, the stylometric pattern $F = \{f_1,f_4,f_6\}$ is a 3-pattern. The *support* of a stylometric pattern F is the percentage of text messages in M_i that contains F. A stylometric pattern F is *frequent* in a set of messages M_i if the support of F is greater than or equal to a user-specified minimum support threshold.

Definition 3.1 (*Frequent stylometric pattern*). Let M_i be the set of text messages written by suspect S_i. Let $support(F|M_i)$ be the percentage of text messages in M_i that contain the pattern F, where $F \subseteq U$. A pattern F is a *frequent stylometric pattern* in M_i if $support(F|M_i) \geqslant min\_sup$, where the minimum support threshold min_sup is a real number in an interval of $[0,1]$. □

Table 1
Stylometric feature vectors (each row representing one message).

Message	Feature X			Feature Y		Feature Z	
	X_1	X_2	X_3	Y_1	Y_2	Z_1	Z_2
m_1	0	1	0	1	0	1	0
m_2	0	1	0	1	0	1	0
m_3	0	1	0	1	0	1	0
m_4	1	0	0	1	0	1	0
m_5	0	0	0	1	0	1	0
m_6	0	0	1	0	1	0	1
m_7	0	0	0	1	0	0	1
m_8	0	0	1	0	1	0	1
m_9	0	1	0	1	0	0	1
m_{10}	1	0	0	1	0	0	1

The writing style of a suspect S_i is represented as a set of frequent stylometric patterns, denoted by $FP(M_i) = \{F_1, \ldots, F_k\}$, extracted from his/her set of text messages M_i.

Example 3.1. Consider the messages in Table 1. Digit '1' indicates the presence of a feature item within a message. Suppose the user-specified threshold $min\_sup = 0.3$, which means that a stylometric pattern $F = \{f_1, \ldots, f_k\}$ is frequent if at least 3 out of the 10 e-mails contain all feature items in F. $\{X_1\}$ is not a frequent stylometric pattern because it has support 2/10 = 0.2. $\{X_2\}$ is a frequent stylometric 1-pattern because it has support 0.4. $\{X_2, Y_1\}$ is a frequent stylometric 2-pattern because it has support 0.4. $\{X_2, Y_1, Z_1\}$ is a frequent stylometric 3-pattern because it has support 0.3. Example 4.1 shows how to efficiently compute all frequent patterns. □

3.1.2. Writeprint

The notion of frequent stylometric patterns in Definition 3.1 captures the writing style of a suspect. However, two suspects, S_i and S_j, may share some similar writing styles. Therefore, it is important to filter out the common frequent stylometric patterns and retain the patterns that are unique to each suspect. This leads us to the notion of "writeprint". Intuitively, a writeprint is a set of frequent stylometric patterns that can uniquely represent the writing style of a suspect S_i if every pattern in the writeprint is found *only* in the text messages written by S_i, but not in other suspects' text messages. In other words, the writeprint of a suspect S_i is a set of stylometric patterns that are frequent in the text messages M_i written by S_i but not frequent in the text messages M_j written by any other suspect S_j where $i \neq j$.

Definition 3.2 (*Writeprint*). A *writeprint*, denoted by $WP(M_i)$, is a set of stylometric patterns where each stylometric pattern F has $support(F|M_i) \geqslant min\_sup$ and $support(F|M_j) < min\_sup$ for any M_j where $i \neq j$, min_sup is a user-specified minimum threshold. In other words, $WP(M_i) \subseteq FP(M_i)$, and $WP(M_i) \cap WP(M_j) = \emptyset$ for any $1 \leqslant i, j \leqslant n$ and $i \neq j$. □

Unlike the physical fingerprint, we do not claim that the writeprint can uniquely distinguish every individual in the world, but our experimental results strongly suggest that the writeprint defined above is accurate enough to uniquely identify the writing style of an individual among a limited number of suspects. Our notion of writeprint has two special properties that make it different from the traditional notion of writeprint in the literature [1].

First, the *combination* of feature items that composes the writeprint of a suspect is dynamically generated based on the embedded patterns in the text messages. This flexibility allows us to succinctly model the writeprint of different suspects by using different combinations of feature items. In contrast, the traditional notion of writeprint considers one feature at a time without considering *all* combinations.

Second, every frequent stylometric pattern F in our notion of writeprint captures a piece of writing pattern that can be found *only* in one suspect's text messages, but not in other suspects' text messages. A cybercrime investigator could precisely point out such matched patterns in the anonymous message to support the conclusion of authorship identification. In contrast, the traditional classification method, e.g., decision tree, attempts to use the *same* set of features to capture the writeprint of different suspects. It is quite possible that the classifier would capture some common writing patterns and the investigator could unintentionally use the common patterns to draw a wrong conclusion of authorship. Our notion of writeprint avoids such problem and, therefore, provides more convincing and reliable evidence.

Definition 3.3 (*Authorship identification with large training samples*). Let $\{S_1, \ldots, S_n\}$ be a set of suspected authors of an anonymous text message ω. Let $\{M_1, \ldots, M_n\}$ be the sets of text messages previously written by suspects $\{S_1, \ldots, S_n\}$, respectively. Assume the number of messages of each set of M_i, denoted by $|M_i|$, is reasonably large (say >30). The *problem of authorship identification with large training samples* is to identify the most plausible author S_a of ω from $\{S_1, \ldots, S_n\}$ with presentable and intuitive evidence. The most plausible author S_a is the suspect whose writeprint of his text messages M_a has the "best match" with stylometric features in ω. □

3.2. Authorship identification with small training samples

The second problem of authorship analysis is to identify the most plausible author S_a of a set of anonymous text messages Ω from a group of suspects $\{S_1, \ldots, S_n\}$, with only *few* sample text messages M_i for each suspect S_i. Note, this problem is different from the first problem in Definition 3.3 in two ways: (1) The number of training samples $|M_i|$ is small (say less than 30 sample e-mails). Therefore, it is infeasible to build a classifier as in the traditional classification method [23,47] or to extract the *frequent* stylometric patterns based on low support counts. (2) The first problem focuses on how to identify the author of one anonymous message. In contrast, this problem focuses on how to cluster the anonymous messages by stylometric features such that the messages written by the same author are clustered together, and how to identify the author of each cluster of anonymous messages.

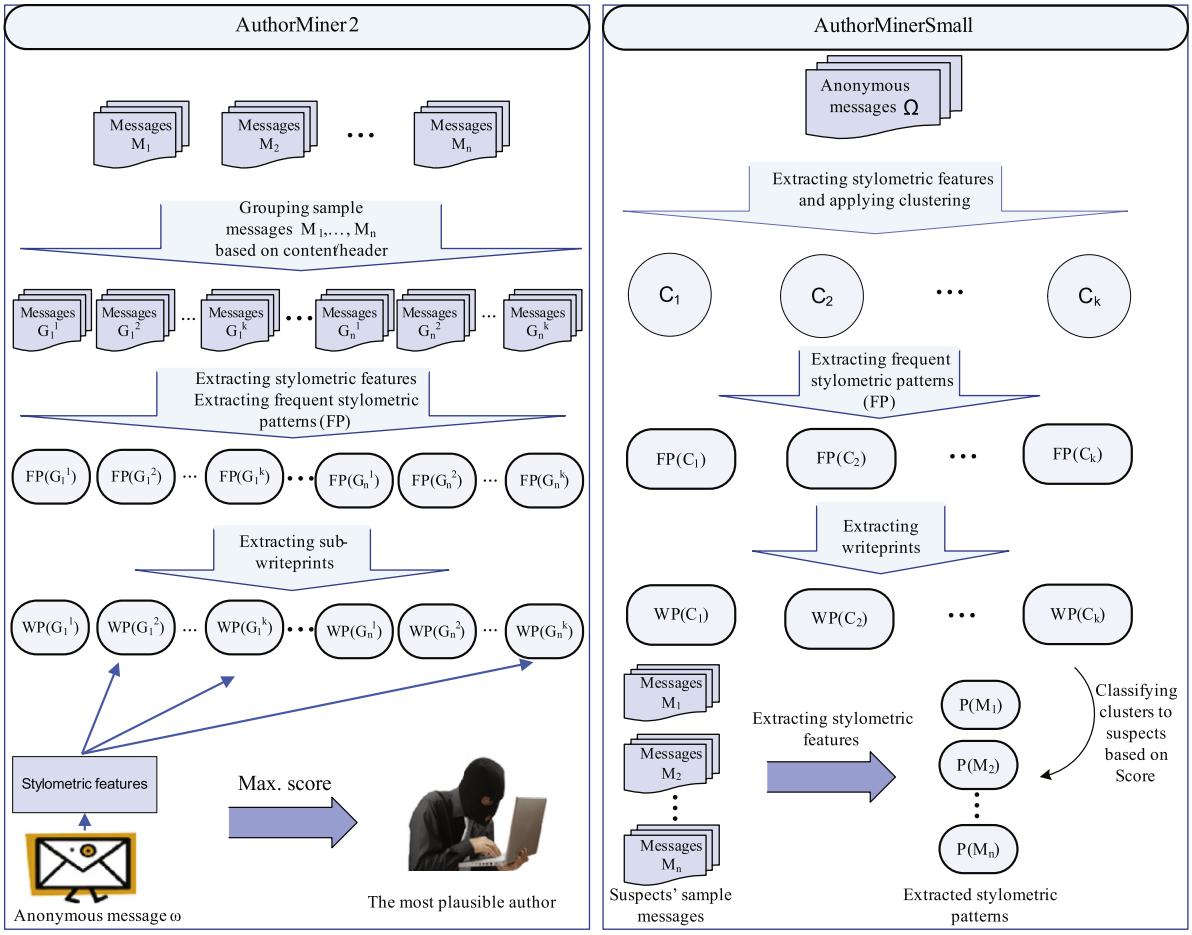
Definition 3.4 (*Authorship identification with small training samples*). Let Ω be a set of anonymous text messages. Let $\{S_1, \ldots, S_n\}$ be a set of suspected authors of Ω. Let $\{M_1, \ldots, M_n\}$ be the sets of text messages previously written by suspects $\{S_1, \ldots, S_n\}$, respectively. Assume $|M_i|$ is small. The *problem of authorship identification with small training samples* is to first

group the messages in Ω into clusters $\{C_1,\ldots,C_k\}$ by stylometric features, and then to identify the plausible author S_a from $\{S_1,\ldots,S_n\}$ for each cluster of anonymous messages $C_j \in \{C_1,\ldots,C_k\}$, with presentable evidence. The most plausible author S_a of C_j is the suspect whose stylometric patterns of his/her text messages M_a have the "best match" with writeprint in C_j. □

3.3. Authorship characterization

The third problem of authorship analysis is to determine the characteristics of the authors of a given set of anonymous text messages. Unlike the previous two problems that have training samples, there are no suspects and no training samples available for investigation in this problem. Instead, the goal of the cybercrime investigator is to infer as much sociolinguistic information as possible about the potential authors based on the stylometric features of the anonymous messages. The sociolinguistic information may reveal the characteristics of the authors, such as ethnicity, age, gender, level of education, and religion. The investigator is assumed to have access to some online text documents with known authors who come from the same population as the suspects. Such online text documents can be collected from blog postings and social networks that disclose the authors' public profiles. The problem of authorship characterization is how to infer the characteristics of authors of the anonymous text messages by matching writeprint in the online text documents.

Definition 3.5 (*Authorship characterization*). Let Ω be a set of anonymous text messages. Let M be a set of online text documents with known authors' characteristics. The *problem of authorship characterization* is to first group the messages in Ω into clusters $\{C_1,\ldots,C_k\}$ by stylometric features, then identify the characteristics of the author of each cluster C_j by matching the writeprint extracted from the online text documents M. □



(a) *AuthorMiner2* (b) *AuthorMinerSmall*

Fig. 1. Overview of *AuthorMiner2* and *AuthorMinerSmall*.

4. A unified data mining approach

In this section, we present a unified data mining approach that utilizes the concept of frequent stylometric patterns to address the three authorship analysis problems described in Definitions 3.3, 3.4, 3.5.

4.1. AuthorMiner2: identify author based on large training samples

To address the authorship problem in Definition 3.3, we propose the algorithm, called *AuthorMiner2*, to identify the author of an anonymous message ω from the suspects $\{S_1,\ldots,S_n\}$ based on the writeprints extracted from their previously written messages $\{M_1,\ldots,M_n\}$. The number of messages in each M_i is assumed to be reasonably large (say >30) for patterns extraction. AuthorMiner2 is an improved version of *AuthorMiner* presented in our previous work [23]. AuthorMiner2 is developed based on the hypothesis that a person may have different writing styles and, therefore, different writeprints, when writing to different recipients. The extension of AuthorMiner2 is to identify such stylistic variations, capture the sub-writeprints from the variations, and utilize the sub-writeprints to further improve the accuracy of authorship identification. Our experimental results support the hypothesis and suggest that the author identification accuracy of AuthorMiner2 is higher than AuthorMiner. Most importantly, AuthorMiner2 can capture and concisely present the fine-grained writing styles of an individual.

Fig. 1(a) shows an overview of AuthorMiner2 in four steps. Step 1 is grouping the training sample messages in M_i (of suspect S_i) by the types of message recipients, for example, by the e-mail address domains. Each set of training sample messages M_i is divided into groups $\{G_i^1,\ldots,G_i^k\}$. Step 2 is extracting the frequent stylometric patterns $FP(G_i^g)$ from each group G_i^g. Step 3 is filtering out the common frequent stylometric patterns shared between any two of the groups across all suspects. The remaining frequent stylometric patterns form the sub-writeprint of each group G_i^g, denoted by $WP(G_i^g)$. Step 4 is identifying the most plausible author S_a of ω by comparing every extracted writeprint $WP(G_i^g)$ with ω. Algorithm 1 illustrates each step in detail.

Algorithm 1. *AuthorMiner2*

Input: An anonymous message ω
Input: Sets of messages $\{M_1,\ldots,M_n\}$ by $\{S_1,\ldots,S_n\}$
for all $M_i \in \{M_1,\ldots,M_n\}$ **do**
  Divide M_i into groups $\left\{G_i^1,\ldots,G_i^k\right\}$;
  for all $G_i^g \in \left\{G_i^1,\ldots,G_i^k\right\}$ **do**
    extract frequent stylometric patterns $FP(G_i^g)$ from G_i^g;
  end for
end for
for all $M_i \in \{M_1,\ldots,M_n\}$ **do**
  for all $G_i^g \in \left\{G_i^1,\ldots,G_i^k\right\}$ **do**
    for all $M_j \in \{M_{i+1},\ldots,M_n\}$ **do**
      for all $G_j^h \in \left\{G_j^1,\ldots,G_j^k\right\}$ **do**
        if $G_i^g \neq G_j^h$ **then**
          for all frequent stylometric pattern $F_x \in FP(G_i^g)$ **do**
            for all frequent stylometric pattern $F_y \in FP\left(G_j^h\right)$ **do**
              if $FP_x == FP_y$ **then**
                $FP(G_i^g) = FP(G_i^g) - F_x$;
                $FP\left(G_j^h\right) = FP\left(G_j^h\right) - F_y$;
              end if
            end for
          end for
        end if
      end for
    end for
    $WP(G_i^g) = FP(G_i^g)$;
  end for
end for
highest_score = −1;
for all $M_i \in \{M_1,\ldots,M_n\}$ **do**

Algorithm 1 (*continued*)

Algorithm 1. *AuthorMiner2*

```
27:   for all G_i^g ∈ {G_i^1, ..., G_i^k} do
28:     if Score(ω ≈ WP(G_i^g)) > highest_score then
29:       highest_score = Score(ω ≈ WP(G_i^g));
30:       author = S_i;
31:     end if
32:   end for
33: end for
34: return author;
```

4.1.1. Grouping messages

The intuition of Step 1 (Lines 1–2 in Algorithm 1) is to divide sample messages M_i of each suspect S_i into different groups $\{G_i^1, \ldots, G_i^k\}$ so that each group of messages G_i^g captures one type of stylistic variation of S_i and, therefore, Step 2 can extract the frequent stylometric patterns of such variation from G_i^g. In real-life application of the method, there is no single grouping method that can guarantee each group captures only one stylistic variation nor can yield optimal authorship identification accuracy because the division of stylistic variations is different depending on the suspect in question. Fortunately, an investigator can usually determine an appropriate grouping based on some basic background information about the suspect, such as his occupation and working hours. Suppose the suspect is a sales manager of a company. His writing to his customers and colleagues will be more formal than his writing to his friends. Thus, the grouping can be based on the domain name of the recipients' e-mail addresses, message timestamps (e.g., during/after office hours), or message contents.

4.1.2. Extracting frequent stylometric patterns

Step 2 (Lines 3–6 in Algorithm 1) extracts the frequent stylometric patterns from each group G_i^g for every M_i of suspect S_i. Frequent patterns mining is a well-known subject in the field of data mining. Any frequent patterns mining algorithm, for example, *Apriori* [5], *FP-growth* [18], *CP-tree* [40], and *MaxClique* [45], can be used to extract the frequent stylometric patterns from the messages. For completeness, we present an overview of the Apriori algorithm in the context of mining frequent stylometric patterns.

Apriori is a level-wise iterative search algorithm that uses frequent stylometric κ-patterns to explore the frequent stylometric $(\kappa + 1)$-patterns. First, the set of messages for writeprint mining is organized as sets of stylometric feature items as shown in Table 1. The set of frequent stylometric patterns are found by scanning the messages in G_i^g, accumulating the support count of each stylometric pattern, and collecting the stylometric pattern F that has $support(F|G_i^g) \geqslant min\_sup$. The features in the resulting frequent 1-patterns are then used to find frequent stylometric 2-patterns, which are used to find frequent stylometric 3-patterns, and so on, until no more frequent stylometric $(\kappa + 1)$-patterns can be found. The generation of frequent stylometric $(\kappa + 1)$-patterns from frequent stylometric κ-patterns is based on the following Apriori property.

Property 4.1 (*Apriori property*). *All nonempty subsets of a frequent stylometric pattern must also be frequent.* □

By definition, a stylometric pattern F is not frequent if $support(F|G_i^g) < min\_sup$. The above property implies that adding a stylometric feature x to a non-frequent stylometric pattern F will never make it more frequent. Thus, if an κ-pattern F is not frequent, then there is no need to generate $(\kappa + 1)$-pattern $F \cup \{x\}$ because the superset of F must not be frequent. Each suspect S_i may have multiple writing styles, and each writing style is represented as a set of frequent stylometric patterns, denoted by $FP(G_i^g) = \{F_1, \ldots, F_z\}$.

Example 4.1. Consider the messages in Table 1. Suppose $min\_sup = 0.3$. First, we identify all frequent 1-patterns by scanning the table once to obtain the support of every stylometric feature item. The items having support $\geqslant 0.3$ are frequent stylometric 1-patterns, denoted by $L_1 = \{\{X_2\},\{Y_1\},\{Z_1\},\{Z_2\}\}$. Then, we join L_1 with itself, i.e., $L_1 \bowtie L_1$, to generate the candidate list $\ell_2 = \{\{X_2,Y_1\},\{X_2,Z_1\},\{X_2,Z_2\},\{Y_1,Z_1\},\{Y_1,Z_2\},\{Z_1,Z_2\}\}$ and scan the table once to identify the patterns in ℓ_2 that have support $\geqslant 0.3$, called frequent stylometric 2-patterns $L_2 = \{\{X_2,Y_1\},\{X_2,Z_1\},\{Y_1,Z_1\},\{Y_1,Z_2\}\}$. Similarly, we perform $L_2 \bowtie L_2$ to generate ℓ_3 and scan the table once to identify $L_3 = \{X_2,Y_1,Z_1\}$. The finding of each set of frequent κ-patterns requires one full scan of the table. □

4.1.3. Filtering common stylometric patterns

Step 3 (Lines 7–25 in Algorithm 1) filters out the common stylometric frequent patterns between any two $FP(G_i^g)$ and $FP\left(G_j^h\right)$ where $i \neq j$. The general idea is to compare every frequent pattern F_x in $FP(G_i^g)$ with every frequent pattern F_y in

all other $FP\left(G_j^h\right)$, and to remove them from $FP(G_i^g)$ and $FP\left(G_j^h\right)$ if F_x and F_y are the same. The computational complexity of this step is $O\left(\left|\cup FP(G_i^g)\right|^2\right)$, where $\left|\cup FP(G_i^g)\right|$ is the total number of stylometric frequent patterns. The remaining stylometic frequent patterns in $FP(G_i^g)$ represents a sub-writeprint $WP(G_i^g)$ of suspect S_i. A suspect S_j may have multiple sub-writeprints $\left\{WP\left(G_j^1\right),\ldots,WP\left(G_j^k\right)\right\}$ depending on how the messages are grouped in Step 1.

Example 4.2. Suppose there are two suspects S_1 and S_2 having two sets of text messages M_1 and M_2, respectively, where M_1 is divided into G_1^1 and G_1^2, and M_2 is divided into G_2^1 and G_2^2. Suppose $FP\left(G_1^1\right)=\{\{X_1\},\{Y_1\},\{X_1,Y_1\}\}$, $FP\left(G_1^2\right)=\{\{X_1\},\{Y_2\},\{X_1,Y_2\}\}, FP\left(G_2^1\right)=\{\{X_1\},\{Z_1\},\{X_1,Z_1\}\}, FP\left(G_2^2\right)=\{\{Y_2\},\{Z_2\},\{X_2,Z_2\}\}$. After filtering, $WP\left(G_1^1\right)=\{\{Y_1\},\{X_1,Y_1\}\}, WP\left(G_1^2\right)=\{\{X_1,Y_2\}\}, WP\left(G_2^1\right)=\{\{Z_1\},\{X_1,Z_1\}\}, WP\left(G_2^2\right)=\{\{Z_2\},\{X_2,Z_2\}\}$. □

4.1.4. Identifying author

Step 4 (Lines 26–35 in Algorithm 1) determines the author of the anonymous message ω by comparing ω with each writeprint $WP(G_i^g)$ of every suspect S_i and identifying the writeprint that is the most similar to ω. Intuitively, a writeprint $WP(G_i^g)$ is similar to ω if many frequent stylometric patterns in $WP(G_i^g)$ match the stylometric feature items found in ω. Formally, a frequent stylometric pattern $F \in WP(G_i^g)$ *matches* ω if ω contains every stylometric feature item in F.

Eq. (1) shows the score function that quantifies the similarity between the anonymous message ω and a writeprint $WP(G_i^g)$. The frequent stylometric patterns in a writeprint are not all equally important, and their importance is reflected by their support count in G_i^g, i.e., the percentage of text messages in G_i^g having such combination of stylometric feature items. Thus, the proposed score function accumulates the support count of a frequent stylometric pattern

$$Score(m \approx WP(G_i^g)) = \sum_{x=1}^{p} support(MP_x|G_i^g) \tag{1}$$

where $MP=\{MP_1,\ldots,MP_p\}$ is a set of matched patterns between $WP(G_i^g)$ and the anonymous message ω. The score is a real number within the range of $[0,1]$. The higher the score means the higher similarity between the writeprint and the malicious message ω. The message ω is assigned to the writeprint of a message group G_a^g with the highest score. The suspect S_a of such group G_a^g is the most plausible author of ω among the suspects.

In the unlikely case that multiple suspects have the same highest score, AuthorMiner2 returns the suspect whose the number of matched patterns $|MP|$ is the largest. In case multiple suspects have the same highest score and the same number of matched patterns, AuthorMiner2 returns the suspect whose the size of matched k-pattern is the largest because having a match on large sized frequent stylometric k-pattern is more significant than a small sized pattern. To facilitate the evaluation procedure in our experiment, the method presented here is designed to return only one suspect. In the actual deployment of the method, a more preferable solution is to return a list of suspects ranked by their scores, followed by the number of matched patterns and the size of the largest matched pattern.

4.2. AuthorMinerSmall: identify author based on small training samples

To address the authorship problem in Definition 3.4, we propose the algorithm, called *AuthorMinerSmall*, to identify the author of a set of anonymous messages Ω from the suspects $\{S_1,\ldots,S_n\}$ based on the writeprints in their previously written messages $\{M_1,\ldots,M_n\}$. The number of messages in M_i is small (say <30) and, therefore, insufficient for extracting frequent stylometric patterns as shown in Section 4.1.

Fig. 1(b) depicts an overview of AuthorMinerSmall, which can be summarized into three steps. Step 1 is grouping the anonymous messages Ω into clusters $\{C_1,\ldots,C_k\}$ by the stylometric features such that each cluster contains the anonymous messages written by the same suspect. Step 2 is extracting the writeprint from each cluster of messages. Step 3 is identifying the most plausible author S_a for each cluster C_j by comparing the extracted writeprint $WP(C_j)$ with every set of training samples M_i.

Step 1 groups the anonymous messages Ω into clusters $\{C_1,\ldots,C_k\}$ by the stylometric features. This step is based on the hypothesis that the writing style of every suspect is different, so clustering by stylometric features could group the messages written by the same author into one cluster. Our previous work [22] has already verified the hypothesis. This clustering step is very different from Step 1 in AuthorMiner2 in Section 4.1.1, which groups *training sample messages* with the goal of identifying the sub-writeprints of a suspect. In contrast, the reason of clustering *anonymous messages* in AuthorMinerSmall is to facilitate more precise writeprint extraction, which is not available from the training samples due to the limited size. The subsequent steps then utilize the extracted writeprint to identify the author of every anonymous e-mail cluster.

One may employ any clustering methods, such as k-means, to group the anonymous messages into clusters $\{C_1,\ldots,C_k\}$ such that messages in the same cluster have similar stylometric features and messages in different clusters have different stylometric features. Often, k is an input parameter to a clustering algorithm. In this case, k can be the number of suspects.

Step 2 extracts the frequent stylometric patterns and writeprints from each cluster C_j. The procedure is the same as the one described in AuthorMiner2 in Sections 4.1.2 and 4.1.3. The frequent stylometric patterns and writeprints are denoted by $\{FP(C_1),\ldots,FP(C_k)\}$ and $\{WP(C_1),\ldots,WP(C_k)\}$, respectively, in Fig. 1(b).

Step 3 identifies the most plausible author for each cluster of anonymous messages C_j by comparing C_j with the training samples $\{M_1,\ldots,M_n\}$. For each message in M_i, we extract the stylometric feature items and take the average of the feature items over all the messages in M_i. The similarity between C_i and M_i is computed by using Eq. (1). The most plausible author is the suspect having the highest score. In the unlikely case that multiple suspects have the same highest score for a given cluster, the strategy discussed in Section 4.1.4 is applicable.

4.3. AuthorCharacterizer: characterize an unknown author

To address the authorship problem in Definition 3.5, we propose the algorithm, called *AuthorCharacterizer*, to characterize the properties of an unknown author of some anonymous messages. Fig. 2 shows an overview of AuthorCharacterizer in three steps. Step 1 is identifying the major groups of stylometric features from a given set of anonymous messages Ω. Step 2 is extracting the writeprints for different categories of online users from the public forums, such as blogs and chat rooms. Step 3 is characterizing the unknown authors of Ω by comparing the writeprints with Ω.

Step 1 groups the anonymous messages Ω into clusters $\{C_1,\ldots,C_k\}$ by the stylometric features. This step is similar to the Step 1 described in AuthorMinerSmall in Section 4.2. The only difference is that the number of clusters k is the number of categories identified for a characteristic. For instance, $k = 2$ (male/female) for gender, $k = 3$ (Australia/Canada/United Kingdom) for region or location.

Step 2 extracts the writeprints from the text messages U in other online sources, in which characteristics such as gender and location are known. In our experiment, we use the blog postings from blogger.com because this website allows bloggers to disclose their personal information. Each collected blog is converted into a set of stylometric feature items. Then we group them by the characteristics that we want to make inferences on the anonymous messages C_j. For example, if we want to infer

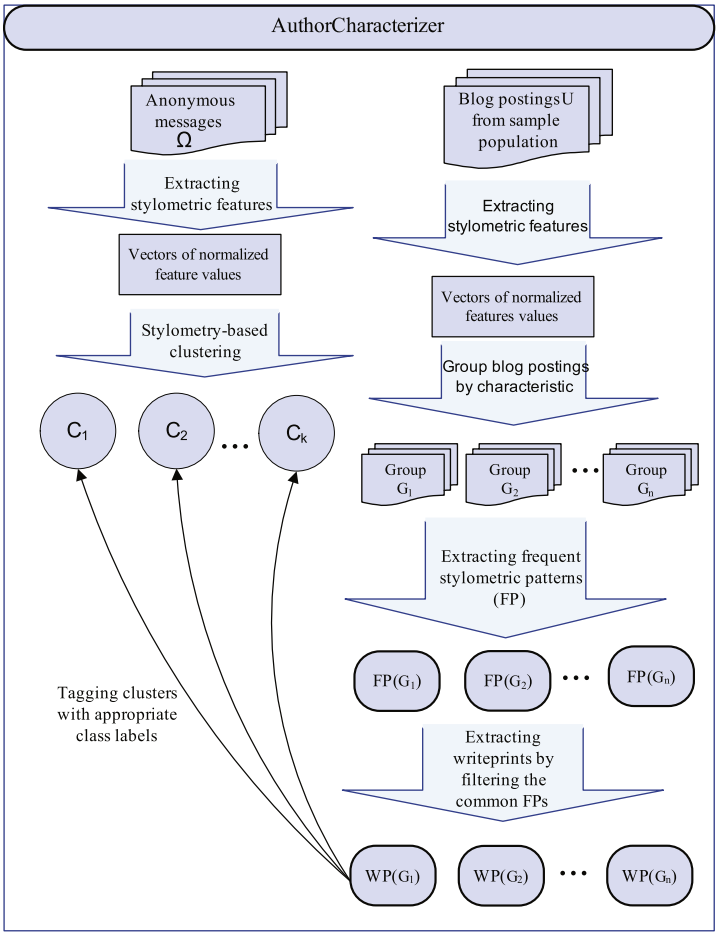
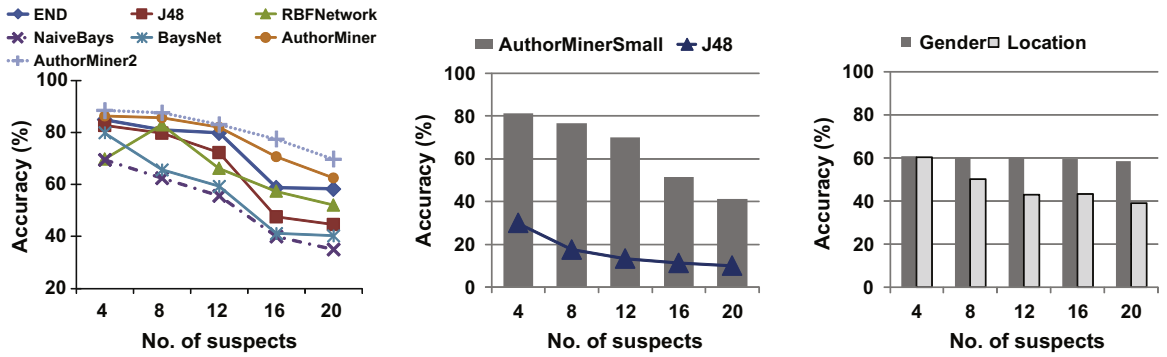


Fig. 2. Overview of *AuthorCharacterizer*.



(a) AuthorMiner2: identification with large samples (b) AuthorMinerSmall: identification with small samples (c) AuthorCharacterizer: characterization of gender and location

Fig. 3. Authorship identification accuracy and characterization accuracy.

the gender of the author of C_j, then we divide the blog postings into groups $G_1,\ldots,G_k$ by gender. Next, we extract the writeprints, denoted by $WP(G_x)$, from each G_x as described in Step 2 in Section 4.2.

Step 3 infers the characteristic of the unknown author of anonymous messages C_j by comparing the stylometric feature items of each message ω in C_j with the writeprint $WP(G_x)$ of every group G_x. The similarity between ω and $WP(G_x)$ is computed using Eq. (1). Message ω is labeled with class x if $WP(G_x)$ has the highest $Score(\omega \approx WP(G_x))$. All anonymous messages C_j are characterized to label x that has the major class.

5. Experimental evaluation

The objectives of the experiments are (1) to evaluate the authorship identification accuracy of our proposed method, AuthorMiner2, and to compare the results with some previously developed classification methods; (2) to evaluate whether or not clustering text messages by writing styles can identify a set of anonymous messages written by the same author; (3) to evaluate the authorship identification accuracy of our proposed method, AuthorMinerSmall, in case the training samples are small; and (4) to evaluate the accuracy of authorship characterization method, AuthorCharacterizer, based on the training data collected from blog postings. All experiments are conducted on a PC with Intel Core2 Quad Q6600 4 GB RAM.

In the experiments, we use 302 stylometric features including 105 lexical features, 159 syntactic features (150 function words and 9 punctuation marks), 15 structural features, 13 domain-specific features, and 10 gender-preferential features.[1] The function words used in our study are listed in [47] while the gender-specific attributes are discussed in [12]. 13 content-specific terms that are common across the Enron dataset are used. The content-specific features are commonly used in the literature of authorship analysis for online messages [1,13,47].

5.1. AuthorMiner2

We perform our experiments on the publicly available e-mail corpus, the Enron e-mail dataset[2], written by former Enron employees. After preprocessing, the corpus contains 200,399 real-life e-mails from 158 individuals [11]. We randomly select 40 messages from each suspect and employ 10-fold cross-validation to measure the authorship identification accuracy. An identification is correct if the AuthorMiner2 or the traditional classification method can correctly identify the true author of an anonymous text message among the group of suspects.

Fig. 3(a) depicts the average identification accuracy over the 10-fold of AuthorMiner2, AuthorMiner [23], which is predecessor of AuthorMiner2, and other classification methods implemented in WEKA [43], namely Ensemble of Nested Dichotomies (END) [17], J48 (a.k.a. C4.5) [34], Radial Basis Function Network (RBFNetwork) [9], Naive Bayes Classifier [36], and BayesNet [32]. These methods are chosen because they are either popular in the field or the state-of-the-arts in their category. For example, RBFNetwork is an artificial neural network, J48 is a commonly employed decision tree classification method, and Naive Bayes is often used as a benchmark classifier for comparison [46].

Fig. 3(a) suggests that AuthorMiner2 and AuthorMiner consistently outperform other traditional classification methods in terms of identification accuracy. This indicates the robustness of our frequent-pattern-based writeprint mining for authorship identification. According to our discussion with a cybercrime team of a law enforcement unit in Canada, the number of suspects is usually less than 10 in most real-life investigations. Compared with other methods, the accuracy of AuthorMiner2

[1] <http://www.ciise.concordia.ca/fung/pub/Stylometric.pdf>.
[2] <http://www.cs.cmu.edu/~enron/>.

Table 2
Paired t test ($\alpha = 0.05$, $df = 4$, critical value $t_{0.05,4} = 2.132$).

AuthorMiner2 vs.	END	J48	RBFNetwork	NaiveBays	BaysNet	AuthorMiner
Test statistic value	2.966	3.242	5.555	8.552	5.207	2.848
Reject H_0?	Yes	Yes	Yes	Yes	Yes	Yes
p-Value	0.0206	0.0158	0.00257	0.000513	0.00324	0.0232

is relatively flat when the number of suspects is less than 10, implying that it is more robust to the number of suspects. Yet, the accuracy of AuthorMiner2 decreases quickly from 88.37% to 69.75% as the number of suspects increases from 4 to 20. This trend generally holds in all classification methods shown in the figure, and is typical in any classification problem. Previous studies [47] also report a similar trend.

The accuracy gap between AuthorMiner and AuthorMiner2 widens as the number of suspects increases. The improvement of AuthorMiner2 over AuthorMiner is contributed by the precise modeling of sub-writeprints. To illustrate the statistical significance of the performance difference between AuthorMiner2 and other methods, we perform a paired t-test on the data in Fig. 3(a) with the null hypothesis H_0: $\mu_D = 0$ and the alternative hypothesis H_a: $\mu_D > 0$ where $\mu_D = \mu_{AuthorMiner2} - \mu_{other\_method}$. H_0 will be rejected if the test statistic value is greater than or equal to the critical value $t_{0.05,4} = 2.132$ at significance level 0.05. H_0 is rejected in all cases as shown in Table 2. The experimental result strongly suggests that the performance of AuthorMiner2 is better than the other six compared methods.

In addition to identification accuracy, AuthorMiner2 can precisely model the writeprint of a suspect in a presentable format. For example, the writeprint of an author called *fossum-d* consists of 86 frequent stylometric patterns. We show two of them below:

$$\{f61 : \text{low}, f62 : \text{low}\} \text{ with } support = 23$$
$$\{f243 : \text{high}, f244 : \text{high}\} \text{ with } support = 18$$

where $f61$ measures the ratio of the number of distinct words and total words, $f62$ measures the vocabulary richness using hapax legomena, $f243$ measures the frequency of the function word "where", and $f244$ measures the frequency of the function word "whether". These two patterns imply that *fossum-d*'s vocabulary richness is low and *fossum-d* often uses the words "where" and "whether" in his/her e-mails.

We also measure the identification accuracy of AuthorMiner2 with respect to the training sample size. With 4 suspects, the average accuracies of sample sizes 10, 20, 30, and 40 are 33.33%, 50%, 71.4%, and 88.37%, respectively. This result suggests that increasing the number of training samples can improve the accuracy. The choice of minimum support threshold (*min_sup*) also affects the performance of our proposed algorithms. Setting it too high will result in having very few or even no frequent stylometric patters. Setting it too low will result in too many insignificant patterns. For both AuthorMiner and AuthorMiner2, we set *min_sup* = 0.1. The efficiency is inversely proportional to *min_sup* due to the increased number of frequent stylometric patterns. An investigator can determine an appropriate *min_sup* by measuring the identification accuracy on the testing data with *min_sup* between 0.05 and 0.5 before matching the writeprint with the actual anonymous messages. If the accuracy does not meet the desired accuracy on the testing data, the investigator may consider decreasing the *min_sup* or increasing the sample size.

Is the accuracy demonstrated in our experiments good enough for investigation? According to our discussion with the law enforcement unit, having 70–90% of identification accuracy is acceptable, especially in the early phase of an investigation when a crime investigator often has little clue to begin with. Yet, we emphasize that our proposed methods cannot (and should not) substitute the role of an expert witness in the court of law. The methods can speed up the analysis process and can identify some less obvious combination of stylometric patterns, but the expert witness still has to apply her expert knowledge to verify the consistency of the extracted results with other available evidences.

The authorship identification process includes reading files, identifying writeprints, and classifying an anonymous e-mail. The total runtime is dominated by the Apriori-based process of the frequent stylometric patterns extraction in the writeprint identification process. Thus, the complexity of AuthorMiner2 is the same as the complexity of Apriori, which is $O(|U|^l \times |M_i|)$, where $|U|$ is the number of distinct stylometric feature items, l is the maximum number of stylometric features of any e-mail, and $|M_i|$ is the number of training samples from suspect S_i. The number of candidate frequent patterns usually peaks at level 2 [19], making the algorithm feasible to run in practice. For any test case of AuthorMiner2 shown in Fig. 3(a), the total runtime is less than 7 min.

5.2. AuthorMinerSmall

AuthorMinerSmall has two steps. The first step clusters the anonymous messages by stylometric features. The second step identifies the author of each anonymous e-mail cluster. The second step is meaningful only if clustering by stylometric features in the first step can successfully group the messages written by the same suspect. The experimental result presented in our previous work [22] suggests that clustering by stylometric features is effective to identify e-mails written by the same

Table 3
Experimental results of AuthorCharacterizer for location identification of anonymous messages.

No. of Authors	Location	Actual accuracy (%)	Weighted accuracy (%)	Sum (%)
4	AU	62.31	20.26	60.44
	CA	51.28	17.95	
	UK	65.39	22.23	
8	AU	46.99	15.04	50.18
	CA	62.00	21.7	
	UK	39.52	13.44	
12	AU	40.81	13.05	43.06
	CA	50.9	17.82	
	UK	35.95	12.22	
16	AU	39.98	12.79	43.21
	CA	49.16	17.21	
	UK	38.6	13.21	
20	AU	40.39	12.92	39.13
	CA	38.13	13.35	
	UK	37.83	12.86	

author. Thus, in this section, we focus on the experimental results of the second step, i.e., to verify whether or not the clustered anonymous e-mails can help authorship identification with few training samples.

Fig. 3(b) depicts the authorship identification accuracy for AuthorMinerSmall and J48 with the number of suspects ranging from 4 to 20. When the number of suspects increases from 4 to 20, the accuracy of AuthorMinerSmall drops from 81.18% to 41.26%. Given that the training dataset is so small, the accuracy above 70% is in fact very encouraging when the number of suspects is not too large. In contrast, the accuracy of J48 ranges from 30% to 10% due to the small training dataset.

AuthorMinerSmall has two phases. The computational complexity of clustering phase depends on the clustering algorithm. For instance, it is $O(k \times |\Omega|)$ for k-means, where k is the number of clusters and $|\Omega|$ is the number of anonymous messages. The computational complexity of writeprint extraction phase is $O(|U|^l \times |C_i| \times k)$, where $|U|$ is the number of distinct stylometric feature items, l is the maximum number of stylometric features of any e-mail, and $|C_i|$ is the number of anonymous messages in cluster C_i. As discussed in Section 5.2, the number of candidate frequent patterns usually peaks at 2. For any test case of AuthorMinerSmall shown in Fig. 3(b), the total runtime is less than a minute.

5.3. AuthorCharacterizer

The evaluation of AuthorCharacterizer has three steps. In the first step, we develop a small robot program to collect blog postings with authors' profiles from a blogger website, group them by gender and location, and extract the writeprint of each group. For characterizing the gender information, we collect 50 postings/messages for each gender type. Thus, if the total number of suspects is n, we collect $50 \times n \times 2$ blog postings in total. The average size of each posting is about 300 words. For characterizing the location information, we collect 737 postings from Australia, 800 postings from Canada, and 775 postings from the United Kingdom. In the second step, we cluster the collected postings by stylometric features as discussed in Section 5.2, and use 2/3 of the messages for training and 1/3 for testing. In the third step, we extract the writeprints from the training messages and characterize the testing messages. A characterization of an anonymous message is correct if AuthorCharacterizer can correctly identify the characteristic of the message. The computational complexity of AuthorCharacterizer is same to the complexity of AuthorMiner2.

Fig. 3(c) depicts the characterization accuracy of gender and location. The accuracy stays almost flat at around 60% for gender, and decreases from 60.44% to 39.13% as the number of authors increases for location. Table 3 shows the detailed experimental results for location. The *actual accuracy* is the percentage of records that are correctly characterized into a class. The *weighted accuracy* is the accuracy weighted by the actual number of records having the class over the total number of records. The *sum* (%) is the sum of the weighted accuracy.

6. Conclusion

In this paper, we study three typical authorship analysis problems encountered by cybercrime investigators: authorship identification with large training samples, authorship identification with small training samples, and authorship characterization. Furthermore, we present a unified data mining approach based on the novel concept of frequent-pattern-based writeprint to address the three problems. Our notion of writeprint, presented in the form of frequent patterns, is suitable for forensic purposes. Due to its intuitiveness, non-technical personnel including the judge and jury in a law court can understand it. Experimental results on real-life data suggest that our proposed approach, together with the concept of frequent-pattern-based writeprint, is effective for identifying the author of online text messages and for characterizing an unknown author.

Future research can focus on the following directions. (1) Our current version of AuthorMiner2 relies on the investigator to divide the messages into groups such that sub-writeprints can be derived. As a result, the identification result varies depending on the subjective grouping. One possible improvement is to devise a clustering method to group the training messages by sub-writeprints. (2) Our current study utilizes blog postings to infer characteristics of an e-mail author. Though our approach demonstrates some initial success, some stylometric features of e-mails are not applicable to blog postings. To further improve the characterization accuracy on e-mails, one research direction is to collect large volume of sample e-mails from authors with different backgrounds, extract the writeprints, and use the writeprints to infer the characteristics of future suspects based on their e-mails.

Acknowledgment

The research is supported in part by the National Cyber-Forensics and Training Alliance Canada (NCFTA Canada) and Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

References

[1] A. Abbasi, H. Chen, Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace, ACM Transactions on Information Systems 26 (2) (2008) 1–29.
[2] A. Abbasi, H. Chen, A comparison of tools for detecting fake websites, IEEE Computer 42 (10) (2009) 78–86.
[3] A. Abbasi, H. Chen, J. Nunamaker, Stylometric identification in electronic markets: scalability and robustness, Journal of Management Information Systems 5 (1) (2008) 49–78.
[4] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, J.F. Nunamaker Jr., Detecting fake websites: the contribution of statistical learning theory, MIS Quarterly 34 (3) (2010) 435–461.
[5] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.
[6] S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2) (2009) 119–123.
[7] S. Argamon, M. Šarić, S.S. Stein, Style mining of electronic messages for multiple authorship discrimination: first results, in: Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2003, pp. 475–480.
[8] R.H. Baayen, H. van Halteren, F.J. Tweedie, Outside the cave of shadows: using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing 2 (1996) 110–120.
[9] M.D. Buhmann, Radial Basis Functions: Theory and Implementations, second ed., Cambridge University Press, 2003.
[10] J.F. Burrows, Word patterns and story shapes: the statistical analysis of narrative style, Literary and Linguistic Computing 2 (1987) 61–67.
[11] V.R. Carvalho, W.W. Cohen, Learning to extract signature and reply lines from email, in: Proceedings of the Conference on Email and Anti-Spam, 2004.
[12] M. Corney, O. de Vel, A. Anderson, G. Mohay, Gender-preferential text mining of e-mail discourse, in: Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC), 2002, p. 282.
[13] O. de Vel, A. Anderson, M. Corney, G. Mohay, Mining e-mail content for author identification forensics, SIGMOD Record 30 (4) (2001) 55–64.
[14] O. de Vel, A. Anderson, M. Corney, G. Mohay, Multi-topic e-mail authorship attribution forensics, in: Proceedings of ACM Conference on Computer Security – Workshop on Data Mining for Security Applications, 2001.
[15] J. Diederich, J. Kindermann, E. Leopold, G. Paass, Authorship attribution with support vector machines, Applied Intelligence 19 (2000) 109–123.
[16] F. Fouss, Y. Achbany, M. Saerens, A probabilistic reputation model based on transaction ratings, Information Sciences 180 (11) (2010) 2095–2123.
[17] E. Frank, S. Kramer, Ensembles of nested dichotomies for multi-class problems, in: Proceedings of the 21st International Conference of Machine Learning (ICML), 2004, pp. 305–312.
[18] J. Han, J. Pei, Mining frequent patterns by pattern-growth: methodology and implications, SIGKDD Exploration Newsletter 2 (2) (2000) 14–20.
[19] M. Hegland, The apriori algorithm – a tutorial, WSPC/Lecture Notes Series 9 (7) (2005). <http://www2.ims.nus.edu.sg/preprints/2005-29.pdf>.
[20] Q. Hu, S. An, D. Yu, Soft fuzzy rough sets for robust feature evaluation and selection, Information Sciences 180 (22) (2010) 4384–4400.
[21] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.
[22] F. Iqbal, H. Binsalleeh, B.C.M. Fung, M. Debbabi, Mining writeprints from anonymous e-mails for forensic investigation, Digital Investigation (2010) 1–9.
[23] F. Iqbal, R. Hadjidj, B.C.M. Fung, M. Debbabi, A novel approach of mining write-prints for authorship attribution in e-mail forensics, Digital Investigation 5 (1) (2008) 42–51.
[24] F. Iqbal, L.A. Khan, B.C.M. Fung, M. Debbabi, E-mail authorship verification for forensic investigation, in: Proceedings of the 25th ACM SIGAPP Symposium on Applied Computing (SAC), Sierre, Switzerland, March 2010, pp. 1591–1598.
[25] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: Proceedings of European Conference on Machine Learning (ECML'98), 1998, pp. 137–142.
[26] M. Koppel, S. Argamon, A.R. Shimoni, Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17 (4) (2002) 401–412.
[27] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, Journal of the American Society for Information Science and Technology 60 (1) (2009) 9–26.
[28] T. Kucukyilmaz, B.B. Cambazoglu, F. Can, C. Aykanat, Chat mining: predicting user and message attributes in computer-mediated communication, Information Processing and Management 44 (4) (2008) 1448–1466.
[29] T.C. Mendenhall, The characteristic curves of composition, Science 11 (11) (1887) 237–249.
[30] A. Miller, Subset Selection in Regression, Chapman & Hall/CRC, 2002.
[31] F. Mosteller, D.L. Wallace, Applied Bayesian and Classical Inference: The Case of the Federalist Papers, second ed., Springer-Verlag, New York, 1964.
[32] J. Pearl, Bayesian networks: a model of self-activated memory for evidential reasoning, in: Proceedings of the 7th Conference of the Cognitive Science Society, 1985, pp. 329–334.
[33] S.R. Pillay, T. Solorio, Authorship attribution of web forum posts, in: eCrime Researchers Summit (eCrime), Dept. of Comput. & Inf. Sci., Univ. of Alabama at Birmingham, Birmingham, AL, USA, 2010, pp. 1–7.
[34] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[35] J.R. Quinlan, C4.5: Programs for machine learning, in: Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993, pp. 343–348.
[36] S.E. Robertson, Sparck K. Jones, Relevance weighting of search terms, Journal of the American Society for Information Science 27 (3) (1976) 129–146.
[37] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys 34 (1) (2002) 1–47.
[38] M. Sewell, Feature selection, 2007. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.6348&rep=rep1&type=pdf>.
[39] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556.

[40] S.K. Tanbeer, C.F. Ahmed, B. Jeong, Y. Lee, Efficient single-pass frequent pattern mining using a prefix-tree, Information Sciences 179 (5) (2009) 559–583.
[41] G. Teng, M. Lai, J. Ma, Y. Li, E-mail authorship mining based on svm for computer forensic, in: Proceedings of the 3rd International Conference on Machine Learning and Cyhemetics, August 2004.
[42] K. Wimmer, The First Amendment and the Media, 2002. <http://www.mediainstitute.org/ONLINE/FAM2002/toc.html>.
[43] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Elsevier, 2005.
[44] G.U. Yule, On sentence length as a statistical characteristic of style in prose, Biometrika 30 (1938) 363–390.
[45] M.J. Zaki, Scalable algorithms for association mining, IEEE Transactions on Knowledge and Data Engineering (TKDE) 12 (2000) 372–390.
[46] Y. Zhao, J. Zobel, Effective and scalable authorship attribution using function words, in: Proceedings of the 2nd AIRS Asian Information Retrieval Symposium, 2005, pp. 174–189.
[47] R. Zheng, J. Li, H. Chen, Z. Huang, A framework for authorship identification of online messages: writing-style features and classification techniques, Journal of the American Society for Information Science and Technology 57 (3) (2006) 1532–2882.