A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques

Rong Zheng
Information, Operations, and Management Science, Stern School of Business, New York University, New York, NY 10012. E-mail: rzheng@stern.nyu.edu

Jiexun Li, Hsinchun Chen, and Zan Huang
Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona, Tucson, AZ 85721.

With the rapid proliferation of Internet technologies and applications, misuse of online messages for inappropriate or illegal purposes has become a major concern for society. The anonymous nature of online-message distribution makes identity tracing a critical problem. We developed a framework for authorship identification of online messages to address the identity-tracing problem. In this framework, four types of writing-style features (lexical, syntactic, structural, and content-specific features) are extracted and inductive learning algorithms are used to build feature-based classification models to identify authorship of online messages. To examine this framework, we conducted experiments on English and Chinese online-newsgroup messages. We compared the discriminating power of the four types of features and of three classification techniques: decision trees, backpropagation neural networks, and support vector machines. The experimental results showed that the proposed approach was able to identify authors of online messages with satisfactory accuracy of 70 to 95%. All four types of message features contributed to discriminating authors of online messages. Support vector machines outperformed the other two classification techniques in our experiments. The high performance we achieved for both the English and Chinese datasets showed the potential of this approach in a multiple-language context.

Introduction

The rapid development and proliferation of Internet technologies and applications have created a new way to share information across time and space. A wide range of activities have evolved over the Internet, ranging from simple information exchange and resource sharing to virtual communications and e-commerce activities. In particular, online messages are being extensively used to distribute information over Web-based channels such as e-mail, Web sites, Internet newsgroups, and Internet chat rooms.

Unfortunately, online messages also can be misused for the distribution of unsolicited or inappropriate information such as junk mail (commonly referred to "spamming") and offensive/threatening messages. Moreover, criminals have been using online messages to distribute illegal materials, including pirated software, child pornography materials, stolen properties, and so on. In addition, criminal or terrorist organizations also use online messages as one of their major communication channels. These activities have spawned the concept of "cybercrime." Cybercrime was defined by Thomas and Loader (2000) as illegal computer-mediated activities which can be conducted through global electronic networks.

A common characteristic of online messages is anonymity. People usually do not need to provide their real identity information such as name, age, gender, and address. In many misuse or crime cases of online messages, the sender will attempt to hide his/her true identity to avoid detection. For example, the sender's address can be forged or routed through an anonymous server, or the sender can use multiple usernames to distribute online messages via different anonymous channels. Therefore, the anonymity of online messages imposes unique challenges to identity tracing in cyberspace. As a result of the sheer growth of cyber users and activities, efficient automated methods for identity tracing are becoming imperative. Authorship identification based on analyzing stylistic features of online messages is suggested to be a possible solution.

In this article, we propose a framework of authorship identification on online messages. In this framework, four types of features that are identified in authorship-analysis

Received October 1, 2003; revised August 1, 2004; accepted February 10, 2005

 ● Published online 21 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20316

research are extracted from online messages, and several major inductive learning techniques are used to build feature-based classification models to perform automated authorship identification. We evaluated the effectiveness of this approach with different types of features and different classification techniques for online messages. Due to the international nature of the Internet, we evaluated the applicability of the proposed framework on different languages. Generally, our framework and experimental study aim to address the following research questions:

Q1: Given the special characteristics of online messages, can the authorship-identification techniques be applied to online messages?
Q2: Which types of writing-style features are effective for identifying the authorship of online messages?
Q3: Which classification techniques are effective for identifying the authorship of online messages?
Q4: What is the prediction power of authorship identification for online messages in a multiple-language context?
Q5: To what extent can authorship-identification techniques be applied to online messages with different numbers of authors and messages?"

Section 2 of this article surveys the literature on authorship analysis and summarizes major types of text features and techniques. Section 3 describes our proposed online-message authorship-identification framework in detail. Section 4 presents an experimental study that answers the research questions. We conclude the article in Section 5 by summarizing our research contributions and noting future directions.

Related Studies

In this section, we briefly review research in authorship analysis. Based on a review of writing-style features, analytical techniques, multiple language issue, and other related parameters, we propose a taxonomy for authorship-analysis research.

Authorship Analysis

Authorship analysis is a process of examining the characteristics of a piece of writing to draw conclusions on its authorship. Its roots are from a linguistic research area called stylometry, which refers to statistical analysis of literary style. As more sophisticated techniques, such as machine learning techniques, have been applied to this domain, this field of research has been generally recognized as authorship analysis. Gray, Sallis, and MacDonell (1997) identified four principle aspects of authorship analysis that can be applied to software forensics. Based on some definitions from Gray et al., we categorized authorship analysis studies into three major fields.

Authorship identification determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author. It also is called "authorship attribution" in some literature, especially by linguistic researchers. The origins of this field date back to the 18th century when English logician Augustus de Morgan suggested that authorship might be settled by determining if one text contained more longer words than another. His hypothesis was investigated by Mendenhall (1887), who subsequently published his results of authorship attribution among Bacon, Marlowe, and Shakespeare. The most thorough and convincing study in this field was conducted by Mosteller and Wallace (1964). In their study on the mystery of the authorship of the Federalist Papers, they attributed all 12 disputed papers to Madison. Their conclusion was generally accepted by historical scholars and became a milestone in this research field.

Authorship characterization summarizes the characteristics of an author and generates the author profile based on his or her writings. Some of these characteristics include gender, educational and cultural background, and language familiarity. This relatively new research direction grew out of the authorship-identification studies. Craig (1999) first made the nexus between authorship identification and characterization by analyzing the plays written by Middleton Thomas and others. He used salient common words which could best discriminate Middleton from others to define the descriptive writing habit of Middleton. More recently, other implicit characteristics of the authors have been investigated. Corney, Vel, Anderson, and Mohay (2002) explored differences of writing style with different education backgrounds of authors. Koppel, Argamon, and Shimoni (2002) and Argamon, Šarić, and Stein (2003) provided solid evidence showing that the writing style of males differs from that of females in the use of pronouns and certain types of noun modifiers. The evaluation is based on both fiction and nonfiction corpora.

Similarity detection compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author. Most studies in this category are related to plagiarism detection. Plagiarism involves the complete or partial replication of a piece of work without permission of the original author. Plagiarism detection attempts to detect the plagiarism activity through examining the similarity between two pieces of writings. Since similarity detection differs much from author identification in various aspects, it is beyond the scope of this article. We refer readers to Clough (2000), which provided a comprehensive review on current tools and technologies of plagiarism detection.

Authorship analysis has been applied to online messages in recent years. For example, de Vel (2000; del Vel, Anderson, Corney, & Mohay, 2001) conducted a series of experiments to identify the authors of e-mail messages. They explored some new characteristics of authorship identification on e-mail. Their results indicated a promising future for applying conventional authorship-identification methods to the online community. Different from the de Vel (2000; de Vel et al., 2001) studies, the current study aimed to construct a framework for authorship identification of online messages, with an emphasis on comparing different types of writing-style features and classification techniques.

DOI: 10.1002/asi

Authorship identification can be formulated as follows: Given a set of writings of a number of authors, assign a new piece of writing to one of them. The problem can be considered as a statistical hypothesis test or a classification problem. The essence of this classification is identifying a set of features that remain relatively constant for a large number of writings created by the same person. Once a feature set has been chosen, a given writing can be represented by an n-dimensional vector, where n is the total number of features. Given a set of precategorized vectors, we can apply many analytical techniques to determine the category of a new vector created based on a new piece of writing. Hence, the features set and the analytical techniques may significantly affect the performance of authorship identification. The multiple language issue is an important new research direction in this field. In the following sections, we review the literature from these perspectives.

Writing-Style Features for Authorship Identification

In early studies, researchers analyzed word usage of different authors to identify authors; however, the effectiveness of this approach is limited since word usage is highly dependent on the topic of the article. To achieve generic authorship identification in various applications, we need "content-free" features. In early work, features such as sentence length (Yule, 1938) and vocabulary richness (Yule, 1944) were proposed. Later, Burrows (1987) developed a set of more than 50 high-frequency words, which were tested on the Federalist Papers. Holmes (1998) analyzed the use of "shorter" words (i.e., two- or three-letter words) and "vowel words" (i.e., words beginning with a vowel). Such word-based and character-based features require intensive efforts in selecting the most appropriate set of words that best distinguish a given set of authors (Holmes & Forsyth, 1995), and sometimes those features are not reliable discriminators when applied to a wide range of applications.

In the seminal work conducted by Mosteller and Wallace (1964), they first used the frequency of occurrence of function words (e.g., "while" and "upon") to clarify the disputed work, the Federalist Papers. For the first time, their analysis demonstrated the discriminating capability of function words. Subsequently, many researchers have confirmed the good discriminating capability of function words (Baayen et al., 1996; Burrows, 1989; Holmes & Forsyth, 1995; Tweedie & Baayen, 1998). Since function-word usage determines how to syntactically form a sentence, we classify function-word usage into syntactic features. Rooted from linguistic research, part of speech (POS) and punctuation usage are other important syntactic features which have been applied to authorship research. Stamatatos, Fakotakis, and Kokkinakis (2001) introduced a fully automatic method to extract POS features, and a better performance was achieved compared to pure lexical-feature-based approaches. Binongo and Smith (1999) used the frequency of occurrence of 25 prepositions to discriminate between Oscar Wilde's plays and essays. Baayen et al. (2002) concluded that incorporating punctuation frequency as a feature can improve the performance of authorship identification. These studies demonstrated that syntactic features might be more reliable in authorship-identification problems than are lexical features.

As a recently explored feature type, structural features attracted more attention. People have different habits when organizing an article. These habits, such as paragraph length, use of indentation, and use of signature, can be strong authorial evidence of personal writing style. This is more prominent in online documents, which have less content information but more flexible structures or richer stylistic information. De Vel et al. (2001) proposed to use structural layout traits and other features for e-mail author identification and achieved high identification performance.

As discussed earlier, key-word-based features are widely accepted to be ineffective in author identification in multiple-topic corpora, hence this type of feature has long been ignored in this field; however, in some circumstances, the proper use of such features can improve the performance of authorship identification. We refer to such consistently used and content-related words or phrases as content-specific features. Usually, these features can express personal interest in a specific domain. One successful application of content-specific features was conducted by Martindale and McKenzie (1995). They correctly attributed 12 disputed Federalist Papers and concluded that content features outperformed lexical features, but not function-word features. In Zheng et al. (2003), approximately 10 content-specific features were introduced in a cybercrime context, and the results showed that they were helpful in improving the author-identification accuracy.

Rudman (1998) summarized that almost 1,000 writing-style features had been used in authorship-analysis applications. There is no agreement on a best set of features for a wide range of application domains. It also is generally accepted that the performance of authorship analysis depends on the combination of the selected features and analytical techniques.

Techniques for Authorship Identification

In early studies, most analytical tools used in authorship analysis were statistical univariate methods. The pioneering study by Mendenhall (1887) was based on histograms of word-length distribution of various authors. Another popular attribution tool of characterizing the stationary distribution of words or letters is a Naïve Bayes (NB) classifier of Mosteller and Wallace (1964), developed during their long work over disputed Federalist Papers. Their systematic work not only provided solid evidence to clarify the disputation but also grounded this field. The CUSUM statistics procedure is another tool applied to authorship analysis by Farringdon (1996). The essence of this procedure is to create the cumulative sum of the deviations of the measured variable and plot that in a graph to compare among authors. This technique showed some success and even became a forensic tool to assist experts conducting authorship analysis. Nevertheless, Holmes (1998) found that the CUSUM analysis was unreliable because the

DOI: 10.1002/asi

stability of that test over multiple topics was not warranted. Univariate methods have another constraint in that they can only deal with one or two features. These constraints called for the application of the multivariate approaches. Burrows (1987) first employed principle component analysis (PCA) on the frequency of function words. PCA is capable of combining many measures and project them into a graph. The geographic distance represents the similarity between different authors' style. The good results encouraged many follow-up studies based on the multivariate method. Biber (1995) applied factor analysis to model the differences between four languages. His finding showed that factor analysis can identify the features dimensions and model the variation. Cluster analysis and discriminant analysis were introduced to this field later by Holmes (1992) and Ledger and Merriam (1994). Mutually supportive results obtained by a variety of multivariate methods have further validated the effectiveness of multivariate approaches.

The advent of powerful computers instigated the extensive use of machine learning techniques in authorship analysis. Tweedie et al. (1996) used a feedforward neural network, also called multilayer perceptron, to attribute authorship to the disputed Federalist Papers. They used three hidden layers and two output layers, and trained with a conjugate gradient. The result was consistent with the results of the previous work over the disputed Federalist Papers. Radial basis function (RBF) networks were applied by Lowe and Matthews (1995) to investigate the extent of Shakespeare's collaboration with his contemporary, John Fletcher, on various plays. More recently, Khmelev and Tweedie (2001) presented a technique for authorship attribution based on a simple Markov Chain, the key idea of which is using the probabilities of the subsequent letters as features. Diederich et al. (2000) introduced Support Vector Machines (SVM) to this field. Experiments were carried out to identify the writings of seven target authors from a set of 2,652 newspaper articles written by several authors covering three topic areas. This method detected the target authors in 60 to 80% of the cases. A new area of study is the identification of e-mail authors based on message content. De Vel et al. (2001) used SVM to classify 150 e-mail documents from three authors. In this experiment, an average accuracy of 80% was achieved. Argamon et al. (2003) furthered the investigation on the 500 e-mail messages written by 20 authors. A variant of Exponentiated Gradient algorithm was examined, and showed that this algorithm outperforms other popular classifiers such as NB and Ripper.

In general, machine learning methods achieved higher accuracy than did statistical methods. The machine learning methods can deal with a larger set of features with fewer requirements on mathematical models or assumptions. Mealand (1995) also noted that machine learning methods were tolerant to noise and nonlinear interactions among features. Besides techniques, parameters such as the number of authors to be identified and the number of messages used to train the classification model also can impact the performance of authorship identification.

Parameters for Authorship Identification

In essence, authorship identification is a classification problem. The complexity level of this problem can be determined by several parameters. For example, the number of authors and the number of available sample documents in the training set may affect the prediction accuracy. Hoorn, Frank, Kowalczyk, and Ham (1999) conducted an experiment to identify poets using neural networks and letter-sequence features. They achieved 80 to 90% accuracy when differentiating two poets. When the choice was between three poets, the accuracy was below 70%. They suggested that the reliability of the classification would decrease when the number of poets increased to more than three poets. Stamatatos et al. (2001) examined the influence of the training data size on authorship-identification performance. They observed that the classification performance was improved as the number of writings by each author in the training dataset increased from 10 to 20.

Most previous experimental studies of authorship identification worked on a relatively small-scale classification problem (i.e., two or three authors). Training size (i.e., the total number of writings) varied widely in different applications. In previous literature, we also noticed decreased identification performance with more than four authors. Although these parameters are considered critical to the complexity of the problem and therefore the prediction accuracy, there are no studies examining their impact on the authorship-identification performance in a systematic way.

Multiple Languages in Authorship Identification

Due to the international nature of the Internet, it is of critical importance to study authorship identification in a multilingual context. Writing-style features are largely language dependent. For instance, Chinese has no explicit word boundaries. Consequently, the features and feature-extraction techniques for Chinese are very different from those for English. Stamatatos et al. (2001) conducted authorship identification on Greek newspaper articles. He proposed a computer-based feature-extraction approach which may be applied to multiple languages. But Greek is to some extent similar to English because they share similar linguistic characteristics, such as the existence of word boundaries. Peng, Schuurmans, Keselj, & Wang (2003) conducted experiments on Greek, English, and Chinese data to examine the performance of authorship attribution across different languages. They identified unique linguistic characteristics of Chinese and concluded that some character-based features such as *n*-gram should be used to avoid word-segmentation problems. They also noted that the Chinese vocabulary is much larger than the English vocabulary, which may give rise to sparse data problems. They examined the *n*-gram language model on Greek newspaper articles, English documents, and Chinese novels using the Bayesian classification technique. In all three languages, the best accuracy achieved was 90%. But the performance for Chinese writings was not as good as that for English writings. Multiple-language support of authorship technology is an

DOI: 10.1002/asi

TABLE 1. A taxonomy for authorship analysis.

Problems		
Category	Description	Label
Authorship Identification	Determines the likelihood of a particular author having produced a piece of writing by examining other writings by that author.	P1
Authorship Characterization	Summarizes the characteristics of an author and determines the author profile based on his/her works.	P2
Similarity Detection	Compares multiple pieces of work and determines whether they are produced by a single author without actually identifying the author.	P3

Features			
Category	Examples		Label
Lexical	Average word/sentence length	Vocabulary richness	F1
Syntactic	Frequency of function words	Use of punctuation	F2
Structural	Paragraph length	Indentation	
	Use of a greeting statement	Use of a farewell statement	F3
Content-specific	Frequency of keywords		F4

Techniques		
Category	Description	Label
Statistical Analysis	Uses statistical methods for calculating document statistics based on metrics to analyze the characteristics of the author or to examine the similarity between various pieces of work.	T1
Machine Learning	Uses classification methods to predict the author of a piece of work based on a set of metrics.	T2

Parameters		
Category	Description	Label
Problem Scope	The number of authors	A
Training Size	Total number of documents in training set	D

important new research direction in this field in light of the continuous globalization of Internet applications.

A Taxonomy of Authorship-Analysis Research

Based on the previous review, we present a taxonomy for authorship-analysis research in Table 1.

Most previous studies addressed the authorship-identification problem, which actually initiated this research domain. Table 2 summarizes major studies in authorship identification since the 1960s using the proposed framework.

Some general conclusions can be drawn from Table 2. First, lexical and syntactic features were most commonly used. Second, statistical approaches were extensively used in the past, but more applications of machine learning techniques have been observed recently. Third, few researchers have addressed multiple-language issues. Finally, most of the previous studies used a small parameter set whereas larger sets of parameters have been investigated in recent years.

A Framework for Authorship Identification of Online Messages

Online messages, as the major channel of Web communication, are important sources for identity tracing in cyberspace. Although authorship-identification methods have achieved successes in many literary and forensic applications, very limited studies have been undertaken specifically on online messages. Due to the special characteristics of online messages, the conventional authorship-identification approach needs to be reexamined.

Characteristics of Online Messages

Compared with conventional objects of authorship identification such as literary works or published articles, one challenge of author identification of online messages is the limited length of online messages. As Ledger and Merriam (1994) claimed, authorship characteristics would not be strongly apparent below 500 words. Based on their review of the size of writings in related studies, Forsyth and Holmes (1996) found that it was very difficult to attribute a text of less than 250 words to an author. The short length of online messages may cause some identifying features in normal texts to be ineffective. For example, since the vocabulary used in short documents is usually limited and relatively unstable, measures such as vocabulary richness may be not as effective as in previous studies on literary works. Therefore, how to correctly identify the authors of

DOI: 10.1002/asi

TABLE 2. Previous studies in authorship identification.

Previous studies	Features				Techniques		Multilanguage		Parameters	
	F1	F2	F3	F4	T1	T2	N	Y	A	D
(Mosteller & Wallace, 1964)		√			√		√		3	85
(Ledger & Merriam, 1994)	√				√		√		2	N/A
(Merriam & Matthews, 1994)	√					√	√		2	50
(Kjell, 1994)	√					√	√		3	85
(Martindale & McKenzie, 1995)	√	√		√	√	√	√		3	85
(Mealand, 1995)	√	√			√		√		1	N/A
(Holmes & Forsyth, 1995)	√	√			√	√	√		3	85
(Farringdon, 1996)	√	√			√		√		N/A	N/A
(Baayen et al., 1996)	√	√			√		√		2	2
(Tweedie et al., 1996)		√				√	√		3	85
(Tweedie & Baayen, 1998)	√				√		√		8	16
(Craig, 1999)	√				√		√		2	97
(Hoorn et al., 1999)	√					√	√		3	90
(Binongo & Smith, 1999)	√				√		√		2	5
(Diederich et al., 2000)	√					√	√		7	700
(De Vel et al., 2001)	√	√	√	√		√	√		4	1259
(Stamatatos et al., 2001)		√			√			√	10	300
(Khmelev & Tweedie, 2001)	√					√	√		45	380
(Corney et al., 2002)	√	√	√	√		√	√		N/A	N/A
(Baayen et al., 2002)	√	√		√	√		√		8	72
(Peng et al., 2003)	√	√				√		√	8–10	200
(Zheng et al., 2003)	√	√	√	√	√	√		√	3–9	150
(Argamon et al., 2003)	√	√	√	√		√	√		20	500

Note. N/A: not available in the paper.

these relatively short documents with appropriate features becomes a challenge.

On the other hand, online messages also have some characteristics which may help reveal the writing style of the author. Since Web-based channels such as e-mail, newsgroup, and chat rooms are relatively casual compared with formal publications, authors are more likely to leave their own "writeprints" in their articles. For example, the structure or composition style used in online messages is often different from normal text documents, possibly because of the different purposes of these two kinds of writings. Some special features, such as structural layout traits, unusual language usage, unusual content markers, and substylistic features, may be useful in forming a suitable feature collection.

Another important characteristic of online messages is the multilingual nature. The Internet is a global network. Cyber users can distribute online messages in any language over cyberspace. For example, most large international criminal and terrorist organizations, such as Osama bin Laden and Al Qaeda, are using the Internet to formulate plans, raise funds, spread propaganda, and communicate (see http://www.usatoday.com/tech/news/2001-02-05-binladen.htm). Online messages are distributed on the Internet across multiple countries in a variety of languages. Therefore, the prediction power of authorship identification for different languages is an immediate concern.

Furthermore, the problem scope and training-set size of authorship identification also present a challenge for online messages. As mentioned earlier, most previous studies, such as authorship identification of Shakespeare's works and the Federalist Papers, dealt with a relatively small number of authors, typically no more than 10; and the average number of messages per author ranged from less than 10 to 300. Under these levels of parameter settings, satisfactory classification performance could be achieved. But in the context of identity tracing on the Internet, the number of potential authors for an online message could be large. Since cyber users often use different usernames on different Web channels, the number of available messages for each author may be limited. Therefore, in this research, we are interested in examining the prediction power of the authorship-identification method for online messages under different parameter settings.

A Framework for Authorship Identification of Online Messages

Based on the taxonomy of authorship analysis we summarized earlier, we propose a framework for authorship identification of online messages in the next subsection (see Figure 1).

Feature set. A feature set is composed of writing-style features predefined by researchers. As an important component of our framework, the feature set may significantly affect the performance of authorship identification. For the special characteristics of online messages discussed earlier, new feature-selection heuristics are necessary. Based on the review of previous studies and analysis, we integrated four

DOI: 10.1002/asi

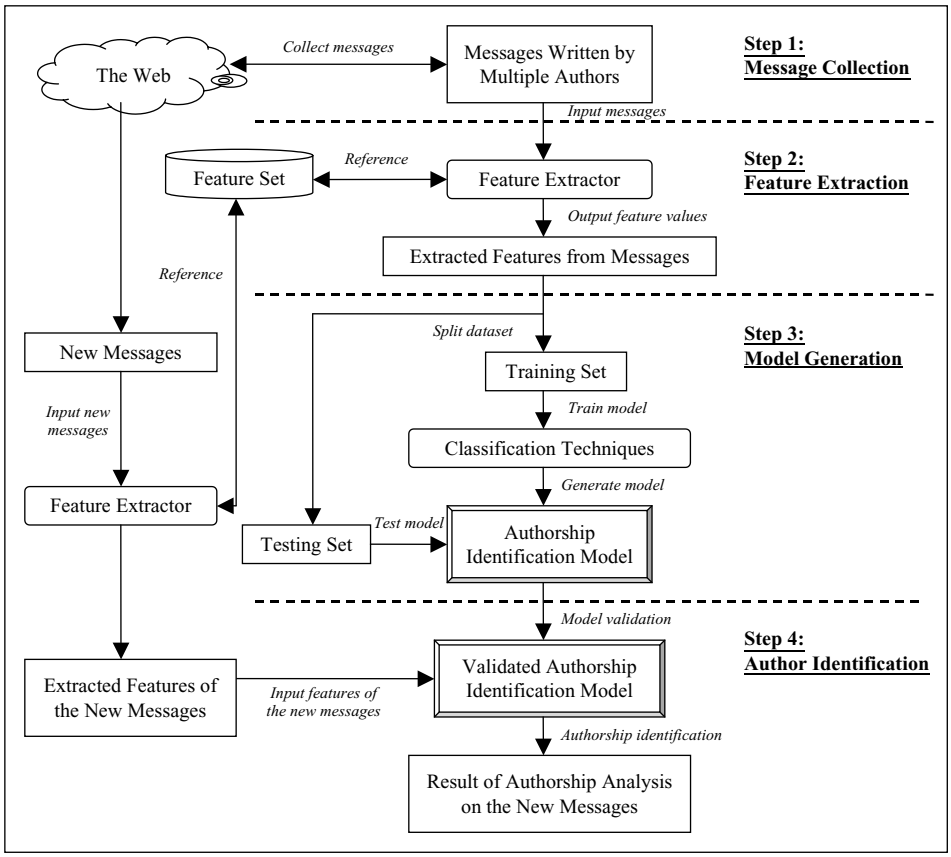

FIG. 1. A framework of authorship identification of online messages.

types of features into the feature set: lexical, syntactic, content-specific, and structural features (see Table 3).

Lexical features can be further divided into character-based and word-based features. In our research, we included character-based lexical features used in de Vel (2000), Forsyth and Holmes (1996), and Ledger and Merriam (1994), vocabulary richness features in Tweedie and Baayen (1998), and word-length frequency features used in Mendenhall (1887) and de Vel et al. (2000). In total, we adopted 87 lexical features for English messages (see Table 3).

Syntactic features, including function words, punctuation, and part of speech, can capture an author's writing style at the sentence level. The discriminating power of syntactic features is derived from people's different habits of organizing sentences. We do not use POS tags as features in this study because POS tagging is still immature for some languages such as Chinese.

One disputed issue in feature selection is how many function words should be considered. Different sets of function words, ranging from 12 to 122, have been tested in various studies (Baayen et al., 1996; Burrows, 1989; de Vel et al., 2001; Holmes & Forsyth, 1995; Tweedie & Baayen, 1998). There is no generally accepted good set of function words for authorship identification because of the varying discriminating power of function words in different applications. In this research, we adopted a large set of 150 function words, which was selected based on previous research, as listed in Appendix A.

We also adopted the punctuation features suggested by Baayen et al. (1996). Combing the function words and punctuation features, we considered 158 syntactic features for English online messages in our framework.

In general, **structural features** represent the way an author organizes the layout of a piece of writing. De Vel (2000) introduced several structural features specifically for e-mail. Since e-mail contains many general structural features of online messages, we adopted those features applicable for online messages. In addition, we added features such as paragraph indentation and signature-related features. In total, we adopted 14 structured features, including 10 features from de Vel (2000) and four newly proposed features (see Table 3).

In addition to "content-free" features, **content-specific features** are important discriminating features for online messages. The selection of such features is dependent on specific application domains. On the Web, one user may often post online messages involving a relatively small range of topics whereas different users may distribute messages on different topics. For this reason, special words or characters closely related to specific topics may provide some clue about the identity of the author. For example, a criminal selling pirate software may use such words as "obo" (or best offer), "for sale," and so on; a criminal

DOI: 10.1002/asi

TABLE 3. Adopted features in the framework.

Features	Description
Lexical features	
Character-based features	
1. Total number of characters(C)	
2. Total number of alphabetic characters/C	
3 Total number of upper-case characters/C	
4. Total number of digit characters/C	
5. Total number of white-space characters/C	
6. Total number of tab spaces/C	
7–32. Frequency of letters (26 features)	A–Z
33–53. Frequency of special characters (21 features)	~ , @, #, $, %, ^, &, *, -, _, = ,+, >, <, [,], {, }, /, \, |
Word-based features	
54. Total number of words (M)	
55. Total number of short words (less than four characters)/M	e.g., and, or
56. Total number of characters in words/C	
57. Average word length	
58. Average sentence length in terms of character	
59. Average sentence length in terms of word	
60. Total different words/M	
61. Hapax legomena*	Frequency of once-occurring words
62. Hapax dislegomena*	Frequency of twice-occurring words
63. Yule's K measure*	A vocabulary richness measure defined by Yule
64. Simpson's D measure*	A vocabulary richness measure defined by Simpson
65. Sichel's S measure*	A vocabulary richness measure defined by Sichele
66. Brunet's W measure*	A vocabulary richness measure defined by Brune
67. Honore's R measure*	A vocabulary richness measure defined by Honore
68–87. Word length frequency distribution /M (20 features)	Frequency of words in different length
Syntactic Features	
88–95 Frequency of punctuations (8 features)	",", ".", "?", "!", ":", ";", " ' ", " " "
96–245 Frequency of function words (303 features)	The whole list of function words is in the appendix.
Structural Features	
246. Total number of lines	
247. Total number of sentences	
248. Total number of paragraphs	
249. Number of sentences per paragraph	
250. Number of characters per paragraph	
251. Number of words per paragraph	
252. Has a greeting	
253. Has separators between paragraphs	
254. Has quoted content	Cite original message as part of replying message
255. Position of quoted content	Quoted content is below or above the replying body
256. Indentation of paragraph	Has indentation before each paragraph
257. Use e-mail as signature	
258. Use telephone as signature	
259. Use url as signature	
Content-specific Features	
260–270. Frequency of content specific keywords (11 features)	"deal", "obo", "sale", "wtb", "thx", "paypal", "check", "windows", "software", "offer", "Microsoft"

Note. The definitions of measures with "*" can be found in Tweedie and Baayen (1998).

distributing pornography materials may frequently use words such as "sexy." In this study, by manually observing and analyzing historical messages, we identify 11 key words as content-specific features particularly for English "for-sale" online messages. Such key words either represent characteristics of the materials for sale or the writing habits of the author when he or she was selling something online. When applied to a different application domain, different content-specific features need to be developed based on specific application characteristics.

Although some languages share similar style features, many languages exhibit unique characteristics, such as function words. Most Western languages use a blank space as a boundary to segment two words. By contrast, in many Oriental languages such as Chinese, such word boundaries often do not exist and words are adjacent to each other in a sentence. These phenomena require different writing-style features for different languages.

We investigate the writing-style features in Chinese in an attempt to examine the capability of our framework in a

DOI: 10.1002/asi

multiple-language context. We choose Chinese because it is the second most popular language on the Web (14.1%, English 35.8%, http://www.glreach.com/globstats/). In addition, Chinese is a typical Oriental language that differs from English on some important aspects (e.g., word boundary). Due to the language differences, some English features (e.g., word length or frequency of the 26 different English letters) do not exist in Chinese—we remove such features. Based on analysis of comparable English features, 117 features from Table 3 were selected for Chinese online messages, including 16 lexical features (Features 1–6, 54, and 59–67), 77 syntactic features (Features 88–95 and 69 Chinese function words), 14 structural features (Features 246–259), and 10 content-specific features. Function words and content-specific features for the Chinese dataset are listed in Appendix A.

Feature extractor. Authors' writing-style features need be extracted from the unstructured text for further analytical purposes. Given the predefined feature set, human beings are able to extract the features with very high accuracy; however, due to the large amount of online messages and the large number of writing-style features, manual feature extraction is too labor intensive and time consuming.

In addition, due to language differences, different feature-extraction procedures are required. For example, since there are no word boundaries in Chinese, automatic word extraction is more difficult. We can solve this problem by extracting characters from a text and reconstructing words from the extracted characters.

In this study, we developed an automated feature-extraction program for English and Chinese, respectively. We randomly selected 30 messages to validate the accuracy of feature extraction for each language. All extracted features were manually examined in the original document. We used accuracy (i.e., the percentage of correctly extracted features of all extracted features) as the effectiveness measure for feature-extraction process. The results showed that the feature extractor achieved an accuracy of over 95 and 90% for English and Chinese, respectively; however, some features, especially structural features, were found difficult to extract accurately. For example, if an author uses an uncommon word for greeting, the program may not be able to extract this feature; if a character "@" is found in the last few lines of a message, the program may incorrectly identify it as an e-mail address. Manual checking of these features is recommended after automated feature extraction.

Classification techniques. To accomplish the authorship-identification task in this framework, many classification techniques can be adopted. In this research, we adopted three popular, yet powerful classifiers: C4.5 decision tree (Quinlan, 1986), backpropagation neural networks (Lippmann, 1987), and SVM (Cristianini & Shawe-Taylor, 2000) (refer to Sebastiani's 2002 survey on machine learning applied to text learning problem). In this study, we chose WEKA data mining tools (publicly available on the Internet at http://www. cs.waikato.ac.nz/ml/weka).

Among the various symbolic learning algorithms developed over the past decades, ID3, the older version of C4.5 dealing with integers, and its variants have been tested extensively and shown to rival other machine learning techniques in predictive power (Chen, Shankaranarayanan, Iyer, & She, 1998; Dietterich, Hild, & Bakiri, 1990; Murthy, Kasif, & Salzberg, 1994). C4.5, an extension of the ID3 algorithm, is a decision-tree building algorithm developed by Quinlan (1986) that adopts a divide-and-conquer strategy and the entropy measure for object classification. Its goal is to classify mixed objects into their associated classes based on the objects' attribute values. In WEKA data mining tools, a standard C4.5 algorithm is implemented.

Backpropagation neural networks have been popular for their unique learning capability (Widrow, Rumelhart, & Lehr, 1994) and have achieved good performance in different applications (Giles, Sun, & Zurada, 1998; Kim & Lewis, 2000; Tolle, Chen, & Chow, 2000). They also were introduced to authorship analysis by Kjell (1994) and Tweedie et al. (1996). WEKA data mining tools provided a standard three-layer, fully connected backpropagation neural network. We chose the heuristic number (i.e., number of nodes in input layer + number of nodes in output layer)/2 as the number of nodes in the hidden level. The input layer nodes are style features, and output nodes are author identities.

An SVM is a novel learning machine first introduced by Vapnik (1995) and is based on the Structural Risk Minimization principle from computational learning theory. Due to the fact that an SVM is capable of handling millions of inputs and good performance (Cristianini & Shawe-Taylor, 2000; Joachims, 2002), it was introduced to authorship-analysis research in previous work (Argamon, 2003; de Vel et al., 2001; Diederich et al., 2000). In this study, we chose WEKA data mining tools, which implemented Platt's (1999) sequential minimal optimization algorithm for training a support vector classifier (see Keerthi et al., 2001, and Platt, 1999, for more information on the sequential minimal optimization algorithm). This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Multiclass problems are solved using the one-against-all method. We used polynomial kernel when we built the classification model.

Procedure of authorship identification. The authorship-identification process, as shown in Figure 1, can be divided into four steps.

Step 1. Message Collection

Based on previous investigations on the Web, investigators need to collect a set of online messages written by potential authors to profile the writing styles of each author.

Step 2. Feature Extraction

Online messages on the Web are in unstructured text format. Based on the predefined writing-style features, the

DOI: 10.1002/asi

From: "The Collectaholic" <mkusz@comcast.net>
Subject: Software Titles - Only $3.00
Newsgroups: misc.forsale.computers.other.software
Date: 2002-10-04 12:07:22 PST
All CDs are the original CDs in working condition and come with all the original documentation.
Shipping is $3.00 for first title and $0.50 for each additional title.
$1.00 Titles
PC World The Best of MediaClips: sounds and graphics that can be used on media projects
…
$3.00 Titles
Boggle: classic word game
Canon Publishing Suite: layout, drawing & photo editing tools

FIG. 2. An example of English online messages.

feature extractor can analyze the messages and extract the features in textual online messages. After feature extraction, each unstructured text is represented as a vector of writing-style features.

Step 3. Model Generation

As in a typical classifier learning process, the online message collection is divided into two subsets. One subset, called the training set, is used to train the classification model. The classification techniques applied in this process may lead to models with different predictive powers. The other subset is called the testing set, which is used to validate the prediction power of the authorship-identification model generated by the classification model. If the performance of the classifier is verified by the testing set, it can be used to identify the authorship of newly found online messages. An iterative training and testing process may be needed to develop a good authorship-prediction model.

Step 4. Author Identification

After the authorship-identification model is developed, it can be used to predict the authorship of unknown online messages. The result of authorship identification will help the investigator focus his or her effort on a small set of messages and authors.

Experimental Evaluation

We conducted an experimental study to evaluate the proposed framework—in particular, the effects of feature types and classification techniques on authorship identification of online messages. Data collection, feature selection, implementation details, and experimental results are presented next.

Data Collection

Among various types of online messages, personal e-mail and chat messages often involve privacy issues and are difficult to collect. Moreover, the format of e-mail is often similar to newsgroup messages. Therefore, publicly available newsgroup messages are selected and collected as the test bed in this study.

English Internet newsgroup messages. In this research, we are particularly interested in illegal online messages such as those concerning the distribution of pirate software. For the experiment, we collected online for-sale messages as potentially illegal messages since they have similar format and content. Over a time period of 2 weeks, we observed the activities of several USENET newsgroups involving computer software and music CD trading. Based on the average number of reads, posts, and unique user IDs per day, we identified *misc.forsale.computers.** (including 27 subgroups) in Google newsgroups to be the most relevant and active for-sale newsgroup in the United States. We list one message in Figure 2 as an example.

Most previous studies addressed the identification problem for two to four authors. But in the cybercrime context, messages to be identified usually have more potential authors, and normally classification algorithms are not adapted to large number of target classes. To examine the capability of the induction algorithm, we identified 20 of the most active users (represented by a unique ID and e-mail address) who frequently posted messages in these newsgroups. The number of messages we collected for each author varies from 30 to 92. The average length of the articles written by each author was 84 to 346 words. The main characteristics of the English dataset are listed in Table 4.

Chinese Bulletin Board System (BBS) messages. To examine the impact of different languages on the performance of authorship identification, we created a Chinese dataset. We downloaded 532 messages from *bbs.mit.edu* and *smth.org*, the most popular Chinese BBS. Although we attempted to create a Chinese dataset comparable to the English dataset, the number of Chinese for-sale messages we could access was quite limited. Instead, we collected online messages on seven topics—movies, music, sports, travel, beauty, love,

TABLE 4. Descriptions of English and Chinese test-beds.

	No. of authors	Average no. of messages per author	Average length of messages per author
English	20	48	169 words
Chinese	20	37	807 words

DOI: 10.1002/asi

and novels—to create the Chinese dataset. Similar to the English dataset, these Chinese messages were written by 20 authors, and each author has 30 to 40 messages. The main characteristics of the dataset are listed in Table 4.

Experimental Design

To examine different features and techniques, we designed several author-identification tasks. First, four feature sets were created. Here, we use F1, F2, F3, and F4 to denote lexical, syntactic, structural, and content-specific features, respectively. The first feature set contained lexical features (F1) only. Syntactic features were added to F1 to form the second feature set (F1+F2). Structural features were added to form the third feature set (F1+F2+F3). The fourth feature set contains all four types of features (F1+F2+F3+F4). We choose this incremental method in such order because it represents the evolutionary sequence of style features, and we intend to examine the effect of adding relatively new features to existing ones. Second, we adopted C4.5, NN, and SVM classifiers, respectively, as the classifiers. We randomly chose five authors and all messages of these five authors from the test bed for all the experiments. A 30-fold cross-validation was used to estimate the accuracy of the classification model. The same procedure was repeated for both the English and Chinese test beds.

To examine the impact of different combinations of parameters (i.e., the number of authors and the number of messages per author) on the performance of authorship identification, we set up four tasks to identify 5, 10, 15, and 20 authors, respectively. For each task, we trained our classification model using different subsets of the initial dataset (10, 15, 20, 25, 30 messages per author). We repeated this procedure for C4.5, NN, and SVM. The same procedure was conducted for both the English and Chinese datasets.

To evaluate the prediction, we used the accuracy measure, which has been commonly adopted in data mining and authorship analysis. Accuracy indicates the overall prediction of a particular classifier, which is defined as in Equation 1 for our experiments:

TABLE 5. Accuracy for different feature sets and different techniques.

	English dataset			Chinese dataset		
	C4.5	NN	SVM	C4.5	NN	SVM
F1	78.06%	86.09%	**89.36%**	52.22%	55.56%	**57.78%**
F1+F2	79.09%	88.72%	**90.03%**	66.22%	68.05%	**69.16%**
F1+F2+F3	88.75%	93.06%	**94.66%**	72.50%	80.27%	**82.77%**
F1+F2+F3+F4	93.36%	96.66%	**97.69%**	74.16%	83.05%	**88.33%**

$$\text{Accuracy} = \frac{\text{Accuracy of messages whose author was correctly identified}}{\text{Total number of messages}} \quad (1)$$

Experimental Results and Discussion

The results for the comparison of different feature types and techniques are summarized in Table 5 and Figure 3. We observed that the accuracy kept increasing as more types of features were used. SVM outperformed NN, which in turn outperformed the C4.5 classifier. The best accuracy was achieved when using SVM and all feature types. The results for the Chinese dataset also were consistent. We discuss the results based on two aspects: feature types and techniques.

Comparison of features types. To examine the effect of adding one type of feature on the accuracy of authorship identification for a certain classification technique and a certain language, we conducted 18 individual pairwise *t* tests. Table 6 shows the *p* values of the *t* tests for feature comparison for the English and Chinese datasets. We chose $\alpha = 5\%$ for each individual *t* test.

Lexical features (F1). When using lexical features alone, C4.5, NN, and SVM achieved 78.06, 86.09, and 89.36% average accuracy, respectively, for the English dataset. The results are comparable to most previous studies. Although the vocabulary richness measure, a lexical feature, may not

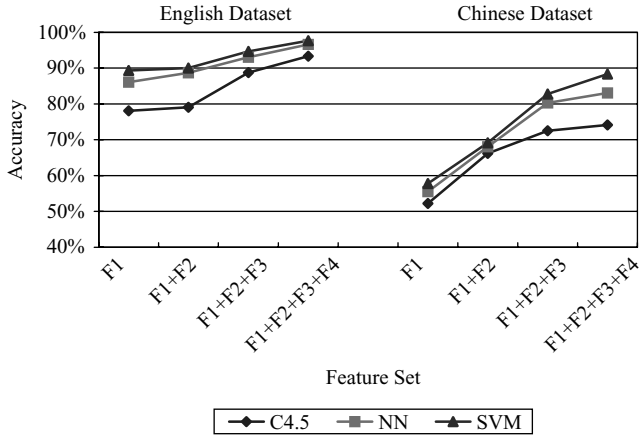

FIG. 3. Authorship identification accuracies for different types of features.

DOI: 10.1002/asi

TABLE 6. Pairwise *t* tests on accuracy for different feature types.

	C4.5		NN		SVM	
	t	*p*	*t*	*p*	*t*	*p*
***t*-test results for the English dataset**						
Feature Types						
F1 < F1+F2	0.2725	.3936	1.3007	.1018	0.2544	.4005
F1+F2 < F1+F2+F3	**3.2142**	**.0016**	**1.8987**	**.0338**	**1.9874**	**.0282**
F1+F2+F3 < F1+F2+F3+F4	**1.8004**	**.0411**	**1.7172**	**.0483**	**1.9960**	**.0277**
***t*-test results for the Chinese dataset**						
Feature Types						
F1 < F1+F2	**2.0706**	**.0237**	**2.6153**	**.0070**	**2.2094**	**.0176**
F1+F2 < F1+F2+F3	1.3090	.1004	**2.3568**	**.0127**	**2.6276**	**.0068**
F1+F2+F3 < F1+F2+F3+F4	0.3188	.3761	0.5671	.2875	1.1618	.1274

be useful when the text length is short (Tweedie & Baayen, 1998), we believe lexical features, in general, are applicable to English online messages. For the Chinese dataset, C4.5, NN, and SVM achieved 52.22, 55.56, and 57.78% average accuracy when using lexical features alone. This indicates that lexical features themselves are not effective for authorship identification for Chinese online messages. More features need to be used to improve the performance.

Syntactic features (F2). For the English dataset, when the syntactic features were added, the accuracy of all three classifiers did not improve significantly (C4.5: $p = .3936$, NN: $p = .1018$, SVM: $p = .4005$). This result is not consistent with previous studies that showed the good discriminating capability of function words. One possible reason is that online messages in our datasets were too short to represent people's usage habits of function words. Another possible reason is that compared with the small number of words used in one message, the number of function words we used as features may be too large. Stamatatos et al. (2001) examined the impact of the number of selected function words, ranging from 10 to 100, on the performance. Their results showed that the best accuracy was achieved by using 60 function words. Similarly, de Vel et al. (2001) observed a decrease in performance when the number of the function words was increased from 122 to 320. Further research is needed to identify a suitable set of function words for online messages.

For the Chinese dataset, adding syntactic features improved the performance significantly (C4.5: $p = .0237$, NN: $p = .0070$, SVM: $p = .0176$). Since the average length (807 words) of the Chinese dataset is longer than that of the English dataset (74 words), it seems that syntactic features are more effective to distinguish longer messages.

Structural features (F3). For the English dataset, adding structural features improved the performance significantly for all three techniques (C4.5: $p = .0016$, NN: $p = .0338$, SVM: $p = .0282$). For the Chinese dataset, the improvement was significant for NN ($p = .0127$) and SVM ($p = .0068$). It appears that an author's consistent writing patterns were reflected in the structural features. For example, some authors always write long sentences and paragraphs and some always use a greeting in every message. Structural features appeared to be a good discriminator among authors for online messages.

Content-specific features (F4). Content-specific features improved the performance of the three classifiers significantly for the English datasets (C4.5: $p = .0411$, NN: $p = .0483$, SVM: $p = .0277$). Authors seem to have certain content-specific keywords (e.g., some people preferred check as a payment method; some people mostly sell Microsoft products). For the Chinese dataset, the content-specific features for the seven topics also helped to improve the classification performance slightly, but not at a significant level.

Comparison of classification techniques. To compare the performances of classification techniques (C4.5, NN, and SVM) on the accuracy of authorship identification given a certain feature set and a certain language, we conducted 24 individual pairwise *t* tests. Table 7 shows the *p* value of the *t* tests for classifier comparison for both datasets. Alpha = 5% is again chosen for each individual *t* test.

We found that for each individual *t* test, SVM achieved significantly higher accuracy than C4.5 decision-tree algorithms in six configurations (of eight). NN achieved significantly higher accuracy than C4.5 decision-tree algorithms in five configurations (of eight). The performance differences between SVM and neural networks were statistically insignificant. In general, our results were consistent with previous studies in that NN and SVM typically had better performance than decision-tree algorithms (Diederich et al., 2000). The good performance of SVM also conformed to its success in many other applications (Joachims, 1998; Osuna, Freund, & Girosi, 1997), although it did not outperform neural networks significantly in our research. We also observed that the performance differences among the three techniques in the Chinese datasets were much smaller as compared to the English dataset. Significant differences between the C4.5 technique and NN and SVM techniques were found only when using most of the feature types. Further research is needed to fully explain the language difference in this aspect.

DOI: 10.1002/asi

TABLE 7. Pairwise *t* tests on accuracy for different classification techniques.

	F1		F1+F2		F1+F2+F3		F1+F2+F3+F4	
	t	*p*	*t*	*p*	*t*	*p*	*t*	*p*
***t*-test results for the English dataset**								
Classification Techniques								
C4.5 < NN	**2.0283**	**.0259**	**3.5907**	**.0006**	**1.6296**	**.0570**	**1.7076**	**.0492**
C4.5 < SVM	**2.8092**	**.0044**	**4.0001**	**.0002**	**2.2146**	**.0174**	**3.0712**	**.0023**
NN < SVM	1.1533	.1291	0.5842	.2818	0.6359	.2649	0.6712	.2537
***t*-test results for the Chinese dataset**								
Classification Techniques								
C4.5 < NN	0.5035	.3092	0.4534	.3268	1.5694	.0637	**1.7597**	**.0445**
C4.5 < SVM	0.8067	.2132	0.5626	.2890	**2.2703**	**.0154**	**3.5327**	**.0007**
NN < SVM	0.4470	.3291	0.2169	.4149	0.5963	.2778	1.0769	.1452

Parameters of authorship identification. Several experiments also were conducted to examine the impact of parameters of authorship identification, including number of authors and number of messages per author. The results are summarized in Figures 4 and 5.

Generally, the results showed that the accuracy of classification increases as the number of authors decreases or the number of messages per author increases. The results are consistent among different classifiers. This result confirmed the general rationale of the classification model and previous studies (Hoorn et al., 1999; Peng et al., 2003; Zheng et al., 2003). Additionally we observed that given a training set of 10 messages per author, SVM achieved 69 to 83% accuracy when the number of authors varied from 5 to 20. Given 20 authors, SVM achieved 69 to 83% accuracy when the number of messages varied from 10 to 30. It appears that the writing-style features are generally effective for a large number of authors.

Performance of authorship identification for different languages. Our experiments showed that the proposed approaches can identify authors of online messages with satisfactory accuracy for both the English and Chinese datasets. In particular, when all features were used, the three classifiers achieved accuracy of 90 to 97% and 72 to 88% for the English and Chinese datasets, respectively. Several factors may account for the discrepancy of performance between the two languages. Different writing-style features are used for English and Chinese. Moreover, the automated feature extraction for Chinese is not as accurate as for English. Better predictions would be achieved if more Chinese writing-style features were developed. Despite the significant language differences, our proposed approach appears promising in a multilingual context.

Conclusion and Future Directions

In this research, we proposed a framework for authorship identification of online messages. To evaluate the effectiveness of this framework, we conducted experiments on English and Chinese online newsgroup messages. The experimental results showed that the proposed approach is able to identify the authors of online messages. Structural features and content-specific features showed particular discriminating capabilities for authorship identification on online messages. SVM and neural networks outperformed C4.5 and neural networks significantly for the authorship-identification task. Different parameter settings of authorship identification had

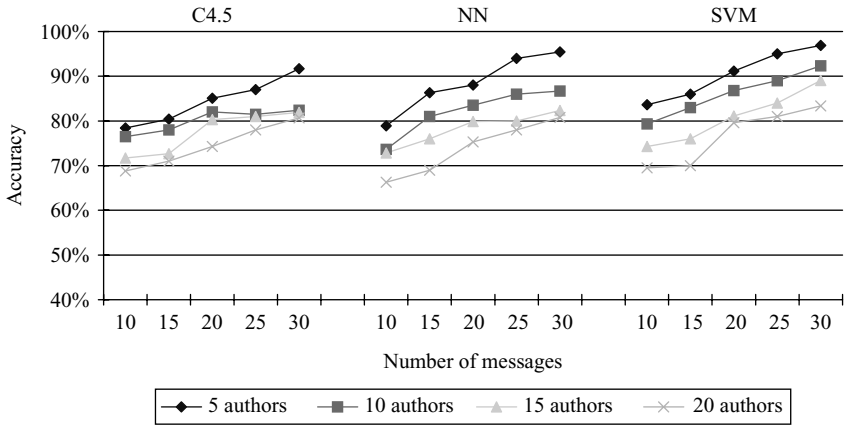

FIG. 4. Accuracies for different numbers of authors and different numbers of messages in the English dataset.

DOI: 10.1002/asi

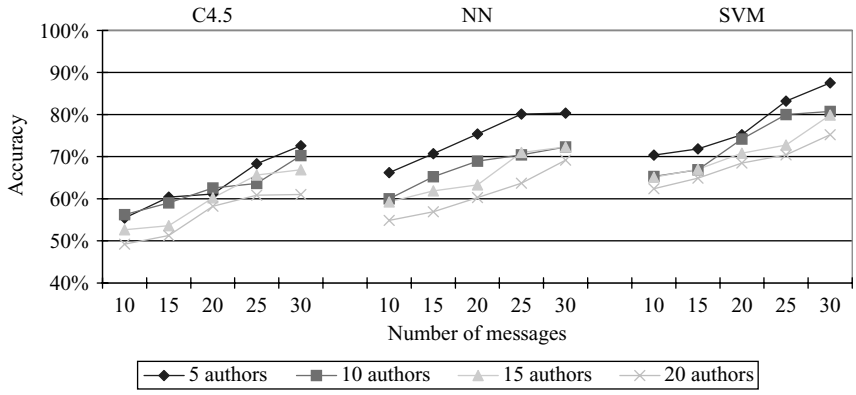

FIG. 5. Accuracies for different numbers of authors and different numbers of messages in the Chinese dataset.

an impact on the performance. The satisfactory performance we achieved for both English and Chinese datasets indicates a promising application of this approach in a multiple-language context. We believe that the proposed framework has the potential to assist in tracing identities in cyberspace.

We have identified several future research directions based on the current study. First, we will explore issues about how to identify the optimal set of features for online messages. How much does each feature contribute to the performance? How can we find a minimal feature set which still allows good performance? Is there an optimal function-word list? Second, this study will be expanded to include more languages to further study language differences. We also aim to examine the feasibility of authorship identification of online messages with mixed languages. Third, content-specific feature selection would benefit from an automatic topic or content classifier which can generate the most distinguishing content features. Another challenging research direction is to validate the proposed technique in the field so that it can be truly useful to assist in tracing identities or even cybercrime investigation in the future.

Acknowledgments

This project was funded primarily by National Science Foundation Grant 9983304, Digital Government Program, "COPLINK Center: Information and Knowledge Management for Law Enforcement," July 2000 to June 2003. We thank Robert Chang from the Taiwan Crime Investigation Bureau for his domain expertise. We also thank Detective Tim Petersen, Sergeant Jennifer Schroeder, and Daniel Casey from the Tucson Police Department for their assistance on the project. Other members of Artificial Intelligence Laboratory who have directly contributed to this article are Yi Qin, Michael Chau, Jie Xu, and Wingyan Chung.

References

Argamon, S., Šarić, M., & Stein, S.S. (2003). Style mining of electronic messages for multiple authorship discrimination: First results. Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 475–480). ACM Press.

Baayen, R.H., Van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. In Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002).

Baayen, R.H., Van Halteren, H., & Tweedie, F.J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 2, 110–120.

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge, UK: Cambridge University Press.

Binongo, J.N.G., & Smith, M.W.A. (1999). The application of principal component analysis to stylometry. Literary and Linguistic Computing, 14(4), 445–466.

Burrows, J.F. (1987). Word patterns and story shapes: The statistical analysis of narrative style. Literary and Linguistic Computing, 2, 61–67.

Burrows, J.F. (1989). "An ocean where each kind…": Statistical analysis and some major determinants of literary style. Computers and the Humanities, 23, 309–321.

Chen, H., Shankaranarayanan, G., Iyer, A., & She, L. (1998). A machine learning approach to inductive query by examples: An experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing. Journal of the American Society for Information Science, 49(8), 693–705.

Clough, P. (2000). Plagiarism in natural and programming languages: An overview of current tools and technologies (Research Memoranda: CS-00–05). Department of Computer Science, University of Sheffield, United Kingdom.

Corney, M., de Vel, O., Anderson, A., & Mohay, G. (2002, December). Gender-preferential text mining of E-mail discourse. Paper presented at the 18th annual Computer Security Applications Conference (ACSAC 2002), Las Vegas, NV.

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? Literary and Linguistic Computing, 14(1), 103–113.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to Support Vector Machines. Cambridge, UK: Cambridge University Press.

de Vel, O. (2000). Mining E-mail authorship. Paper presented at the Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD 2000).

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. SIGMOD Record, 30(4), 55–64.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2000). Authorship attribution with Support Vector Machines. Applied Intelligence, 19, 109–123.

Dietterich, T.G., Hild, H., & Bakiri, G. (1990). A comparative study of ID3 and backpropagation for English text-to-speech mapping. In Proceedings of the Seventh International Conference on Machine Learning (pp. 24–31). Morgan Kaufmann.

Farringdon, J.M. (1996). Analyzing for authorship: A guide to the Cusum technique. Cardiff: University of Wales Press.

DOI: 10.1002/asi

Forsyth, R.S., & Holmes, D.I. (1996). Feature finding for text classification. Literary and Linguistic Computing, 11(4), 163–174.
Giles, C.L., Sun, R., & Zurada, J.M. (1998). Neural networks and hybrid intelligent models—Foundations, theory, and applications. IEEE Transactions on Neural Networks, 9(5), 721–723.
Gray, A., Sallis, P., & MacDonell, S. (1997). Software forensics: Extending authorship analysis techniques to computer programs. Paper presented at the 3rd biannual conference of the International Association of Forensic Linguists (IAFL '97).
Holmes, D.I. (1992). A stylometric analysis of Mormon scripture and related texts. Journal of Royal Statistical Society, 155, 91–120.
Holmes, D.I. (1998). The evolution of stylometry in humanities. Literary and Linguistic Computing, 13(3), 111–117.
Holmes, D.I., & Forsyth, R.S. (1995). The Federalist revisited: New directions in authorship attribution. Literary and Linguistic Computing, 10, 111–127.
Hoorn, J.F., Frank, S.L., Kowalczyk, W., & Ham, F.V.D. (1999). Neural network identification of poets using letter sequences. Literary and Linguistic Computing, 14(3), 311–338.
Hsu, C.W., & Lin, C.J. (2002). A comparison on methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13, 415–425.
Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine learning (ECML) (pp. 137–142). Springer-Verlag.
Joachims, T (2002). Learning to classify text using support vector machines: Methods, theory, and algorithms. Dordrecht, The Netherlands: Kluwer.
Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., & Murthy, K.R.K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation, 13(3), 637–649.
Khmelev, D.V., & Tweedie, F.J. (2001). Using Markov chains for identification of writers. Literary and Linguistic Computing, 16(4), 299–307.
Kim, Y.H., & Lewis, F.L. (2000). Optimal design of CMAC Neural-Network controller for robot manipulators. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 30(1), 22–31.
Kjell, B. (1994). Authorship determination using letter-pair frequency features with Neural Network classifiers. Literary and Linguistic Computing, 9, 119–124.
Koppel, M., Argamon, S., & Shimoni, A.R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412.
Ledger, G.R., & Merriam, T.V.N. (1994). Shakespeare, Fletcher, and the two Noble Kinsmen. Literary and Linguistic Computing, 9, 235–248.
Lippmann, R.P. (1987). An introduction to computing with Neural Networks. IEEE Acoustics Speech and Signal Processing Magazine, 4(2), 4–22.
Lowe, D., & Matthews, R. (1995). Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions. Computers and the Humanities, 29, 449–461.
Martindale, C., & McKenzie, D. (1995). On the utility of content analysis in author attribution: The Federalist. Computer and the Humanities, 29, 259–270.
Mealand, D.L. (1995). Correspondence analysis of Luke. Literary and Linguistic Computing, 10, 171–182.
Mendenhall, T.C. (1887). The characteristic curves of composition. Science, 11(11), 237–249.
Merriam, T.V.N., & Matthews, R.A.J. (1994). Neural computation in stylometry: II. An application to the works of Shakespeare and Marlowe. Literary and Linguistic Computing, 9, 1–6.
Mitton, R. (1987). Spelling checkers, spelling correctors and the misspellings of poor spellers. Information Processing and Management, 23(5), 495–505.
Mosteller, F., & Wallace, D.L. (1964). Applied Bayesian and classical inference: The case of the Federalist Papers (2nd ed.). New York: Springer-Verlag.
Mosteller, F., & Wallace, D.L. (1964). Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley.
Murthy, S.K., Kasif, S., & Salzberg, S. (1994). A system for induction of oblique decision trees. Journal of Artificial Intelligence Research, 2, 1–32.
Osuna, E., Freund, R., & Girosi, F. (1997). Training Support Vector Machines: An application to face detection. Proceedings of Computer Vision and Pattern Recognition (pp. 130–136). IEEE Computer Society.
Peng, F., Schuurmans, D., Keselj, V., & Wang, S. (2003). Automated authorship attribution with character level language models. Paper presented at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) (pp. 267–274). Association for Computational Linguistics.
Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. Cambridge, MA: MIT Press.
Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities, 31, 351–365.
Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34, 1–47.
Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35(2), 193–214.
Thomas, D., & Loader, B.D. (2000). Cybercrime: Law enforcement, security and surveillance in the information age. Routledge.
Tolle, K.M., Chen, H., & Chow, H. (2000). Estimating drug/plasma concentration levels by applying Neural Networks to pharmacokinetic data sets. Decision Support Systems, 30(2), 139–152.
Tweedie, F.J., & Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32, 323–352.
Tweedie, F.J., Singh, S., & Holmes, D.I. (1996). Neural Network applications in stylometry: The Federalist Papers. Computers and the Humanities, 30(1), 1–10.
Vapnik, V. (1995). The nature of statistical learning theory. New York: Springer-Verlag.
Widrow, B., Rumelhart, D.E., & Lehr, M.A. (1994). Neural Networks: Applications in industry, business, and science. Communications of the ACM, 37, 93–105.
Yule, G.U. (1938). On sentence length as a statistical characteristic of style in prose. Biometrika, 30, 363–390.
Yule, G.U. (1944). The statistical study of literary vocabulary. Cambridge, UK: Cambridge University Press.
Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2003). Authorship analysis in cybercrime Investigation. Proceedings of the 1st NSF/NIJ Symposium, ISI2003 (pp. 59–73). Springer-Verlag.

DOI: 10.1002/asi

Appendix

English function words in our feature set:

a	between	in	nor	some	upon
about	both	including	nothing	somebody	us
above	but	inside	of	someone	used
after	by	into	off	something	via
all	can	is	on	such	we
although	cos	it	once	than	what
am	do	its	one	that	whatever
among	down	latter	onto	the	when
an	each	less	opposite	their	where
and	either	like	or	them	whether
another	enough	little	our	these	which
any	every	lots	outside	they	while
anybody	everybody	many	over	this	who
anyone	everyone	me	own	those	whoever
anything	everything	more	past	though	whom
are	few	most	per	through	whose
around	following	much	plenty	till	will
as	for	must	plus	to	with
at	from	my	regarding	toward	within
be	have	near	same	towards	without
because	he	need	several	under	worth
before	her	neither	she	unless	would
behind	him	no	should	unlike	yes
below	i	nobody	since	until	you
beside	if	none	so	up	your

Chinese function words in our feature set:

我 偶 你 他 我们 偶们 俺们 咱们 你们 他们 的 地 得 着 了 过 啊 呀 哎 呢 吧 哦 喔 噢 呵 呵呵 也 也许 都 又 是 就 一个 以后 然后 然而 虽然 但是 到底 随着 不然 后来 之后 总之 直到 往往 其实 反正 觉得 我想 认为 为什么 什么 怎么 怎样 难道 特别 却是 确实 的确 要不

Content-specific words in Chinese feature set:

美容 旅游 电影 结尾 女人 主角 景色 歌坛 皮肤 润肤

DOI: 10.1002/asi