Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace

AHMED ABBASI and HSINCHUN CHEN
The University of Arizona

One of the problems often associated with online anonymity is that it hinders social accountability, as substantiated by the high levels of cybercrime. Although identity cues are scarce in cyberspace, individuals often leave behind textual identity traces. In this study we proposed the use of stylometric analysis techniques to help identify individuals based on writing style. We incorporated a rich set of stylistic features, including lexical, syntactic, structural, content-specific, and idiosyncratic attributes. We also developed the Writeprints technique for identification and similarity detection of anonymous identities. Writeprints is a Karhunen-Loeve transforms-based technique that uses a sliding window and pattern disruption algorithm with individual author-level feature sets. The Writeprints technique and extended feature set were evaluated on a testbed encompassing four online datasets spanning different domains: email, instant messaging, feedback comments, and program code. Writeprints outperformed benchmark techniques, including SVM, Ensemble SVM, PCA, and standard Karhunen-Loeve transforms, on the identification and similarity detection tasks with accuracy as high as 94% when differentiating between 100 authors. The extended feature set also significantly outperformed a baseline set of features commonly used in previous research. Furthermore, individual-author-level feature sets generally outperformed use of a single group of attributes.

Categories and Subject Descriptors: I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Stylometry, online text, discourse, style classification, text mining

ACM Reference Format:

Abbasi, A. and Chen, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26, 2, Article 7 (March 2008), 29 pages. DOI = 10.1145/1344411.1344413 http://doi.acm.org/ 10.1145/1344411.1344413

This work was funded in part by a grant from NSF Information and Data Management "Multilingual Online Stylometric Authorship Identification: An Exploratory Study," August 2006–August 2007.

Authors' addresses: A. Abbasi, H. Chen, Department of Management Information Systems, University of Arizona, Room 430, McClelland Hall, 1130 E. Helen St., Tucson, AZ 85721; email: aabbasi@email.arizona.edu; hchen@eller.arizona.edu.

1. INTRODUCTION

The Internet's numerous benefits have been coupled with the realization of several vices attributable to the ubiquitous nature of computer-mediated communication and abuses of online anonymity. The Internet is often used for the illegal sale and distribution of software [Moores and Dhillon 2000; Zheng et al. 2006]. It also serves as an attractive medium for hackers indulging in online attacks [Oman and Cook 1989; Krsul and Spafford 1997] and cyber-wars [Garson 2006]. Furthermore, Internet-based communication is swarming with fraudulent schemes, including email scams. One well-known fraudulent scheme is the 4-1-9 scam [Airoldi and Malin 2004] where deceptive individuals convince users to provide bank account information, or to cash fake cashier checks through email and forum messages. The scam has been around for over a decade and has generated over 5 billion dollars in fraudulent revenues [Sullivan 2005]. Electronic marketplaces constitute another area susceptible to deception in the form of reputation rank inflation [Morzy 2005]. In this scheme, online sellers create fake sales transactions to themselves in order to improve reputation rank [Josang 2007]. While artificial accreditation can simply be a business ploy, it is also often done in order to defraud unsuspecting future buyers.

Tools providing greater informational transparency in cyberspace are necessary to counter anonymity abuses and garner increased accountability [Erickson and Kellogg 2000; Sack 2000]. The aforementioned forms of Internet misuse all involve text-based modes of computer-mediated communication. Hence, the culprits often leave behind potential textual traces of their identity [Li et al. 2006]. Peng et al. [2003] refer to an author's unique stylistic tendencies as an author profile. Ding et al. [2003] described such identifiers as text fingerprints that can discriminate authorship.

Stylometry is the statistical analysis of writing style [Zheng et al. 2006]. In lieu of these textual traces, researchers have begun to use online stylometric analysis techniques as a forensic identification tool, with recent application to email [De Vel et al. 2001], forums [Zheng et al. 2006], and program code [Gray et al. 1997]. Despite significant progress, online stylometry has several current limitations. The biggest shortcoming has been the lack of scalability in terms of number of authors and across application domains (e.g., email, forums, chat). This is partially attributable to use of feature sets that are insufficient in terms of the breadth of stylistic tendencies captured. Furthermore, previous work has also mostly focused on the identification task (where potential authorship entities are known in advance). There has been limited emphasis on similarity detection, where no entities are known a priori (which is more practical for cyberspace).

In this study we addressed some of the current limitations of online stylometric analysis. We incorporated a larger, more holistic feature set than those used in previous studies. We also developed the Writeprint technique, which is intended to improve stylometric analysis scalability across authors and domains for identification and similarity detection tasks. Experiments were conducted in order to evaluate the effectiveness of the proposed feature set and technique in comparison with benchmark techniques and a baseline feature set.

The remainder of this article is organized as follows. Section 2 presents a general review of stylometric analysis and a taxonomy of online stylometric analysis studies. Section 3 describes research gaps, questions, and our proposed research design. Section 4 describes the system design, which includes the stylometric features and techniques utilized in our analysis. Section 5 presents two experiments used to evaluate the effectiveness of the proposed approach and discussion of the results. Section 6 concludes with a summary of our research contributions, closing remarks, and future directions.

2. RELATED WORK

In this section we present a summary of stylometry, followed by a taxonomy and review of online stylometric analysis research.

2.1 Stylometry

Stylometric analysis techniques have been used for analyzing and attributing authorship of literary texts for numerous years (e.g., Mosteller and Wallace [1964]). Three important characteristics of stylometry are the analysis tasks, writing-style features used, and the techniques incorporated to analyze these features [Zheng et al. 2006]. These characteristics are discussed next.

2.1.1 *Tasks.* Two major stylometric analysis tasks are identification and similarity detection [Gray et al. 1997; De Vel et al. 2001]. The objective in the identification task is to compare anonymous texts against those belonging to identified entities, where each anonymous text is known to be written by one of those entities. The Federalist papers [Mosteller and Wallace 1964] are a good example of a stylometric identification problem. Twelve anonymous/disputed essays were compared against writings belonging to Madison and Hamilton. Since all possible author classes are known a priori, identification problems can use supervised or unsupervised classification techniques.

The objective in the similarity detection task is to compare anonymous texts against other anonymous texts and assess the degree of similarity. Examples include online forums, where there are numerous anonymous identities (i.e., screen names, handles, email addresses). Similarity detection tasks can only use unsupervised techniques, since no class definitions are available beforehand.

2.1.2 *Features.* Stylistic features are the attributes or writing-style markers that are the most effective discriminators of authorship. The vast array of stylistic features includes lexical, syntactic, structural, content-specific, and idiosyncratic style markers.

Lexical features are word, or character-based statistical measures of lexical variation. These include style markers such as sentence/line length [Yule 1938; Argamon et al. 2003], vocabulary richness [Yule 1944], and word-length distributions [De Vel et al. 2001; Zheng et al. 2006].

Syntactic features include function words [Mosteller and Wallace 1964], punctuation [Baayen et al. 2002], and part-of-speech tag *n*-grams [Baayen et al. 1996; Argamon et al. 1998]. Function words have been shown to be highly effective discriminators of authorship, since the usage variations of such words are a strong reflection of stylistic choices [Koppel et al. 2006].

Structural features, which are especially useful for online text, include attributes relating to text organization and layout [De Vel et al. 2001; Zheng et al. 2006]. Other structural attributes include technical features such as the use of various file extensions, fonts, sizes, and colors [Abbasi and Chen 2005]. When analyzing computer programs, different structural features (e.g., the use of braces and comments) are utilized [Oman and Cook 1989].

Content-specific features are comprised of important keywords and phrases on certain topics [Martindale and McKenzie 1995] such as word *n*-grams [Diederich et al. 2003]. For example, content-specific features on a discussion of computers may include "laptop" and "notebook."

Idiosyncratic features include misspellings, grammatical mistakes, and other usage anomalies. Such features are extracted using spelling and grammar checking tools and dictionaries [Chaski 2001; Koppel and Schler 2003]. Idiosyncrasies may also reflect deliberate author choices or cultural differences, such as use of the word "center" versus "center" [Koppel and Schler 2003].

Over 1,000 different features have been used in previous authorship analysis research, with no consensus on a best set of style markers [Rudman 1997]. However, this could be attributable to certain feature categories being more effective at capturing style variations in different contexts. This necessitates the use of larger feature sets comprised of several categories of features (e.g., punctuation, word-length distributions, etc.) spanning various feature groups (i.e., lexical, syntactic, etc.). For instance, the use of feature sets containing lexical, syntactic, structural, and syntactic features has been shown more effective for online identification than feature sets containing only a subset of these feature groups [Abbasi and Chen 2005; Zheng et al. 2006].

2.1.3 *Techniques.* Stylometric analysis techniques can be broadly categorized into supervised and unsupervised methods. Supervised techniques are those that require author-class labels for categorization, while unsupervised techniques make categorizations with no prior knowledge of author classes.

Supervised techniques used for authorship analysis include support vector machines (SVMs) [Diederich 2000; De Vel 2001; Li et al. 2006], neural networks [Merriam 1995; Tweedie et al. 1996; Zheng et al. 2006], decision trees [Apte 1998; Abbasi and Chen 2005], and linear discriminant analysis [Baayen 2002; Chaski 2005]. SVM is a highly robust technique that has provided powerful categorization capabilities for online authorship analysis. In head-to-head comparisons, SVM significantly outperformed other supervised learning methods such as neural networks and decision trees [Abbasi and Chen 2005; Zheng et al. 2006].

Unsupervised stylometric categorization techniques include principal component analysis (PCA) and cluster analysis [Holmes 1992]. PCA's ability to capture essential variance across large numbers of features in a reduced dimensionality makes it attractive for text analysis problems, which typically involve large feature sets. PCA has been used in numerous previous authorship studies

Table I. A Taxonomy for Online Stylometric Analysis

Tasks		
Category	Description	Label
Identification	Comparing text from anonymous identities against known classes.	T1
Similarity Detection	Texts from anonymous identities are compared against each other in order to assess degree of similarity with no prior class definitions.	T2
Domains		
Category	Examples	Label
Asynchronous CMC	Asynchronous conversation including email, web forums, and blogs.	D1
Synchronous CMC	Persistent text including chat rooms and instant messaging.	D2
Documents	Electronic documents including nonliterary texts and news articles.	D3
Program Code	Text containing code snippets and examples.	D4
Features		
Category	Examples	Label
No. of Categories	Maximum number of stylistic feature categories used in experiments.	Cat.
Number of Features	Maximum number of style marking attributes incorporated.	No.
Feature-Set Type	Whether a single author-group-level feature set or multiple individual author-level subsets were used.	Type
Classes		
Category	Description	Label
No. of Classes	Maximum number of classes used in experiments.	No.

(e.g., Burrows 1987, Baayen et al. 1996]), and has also been shown effective for online stylometric analysis [Abbasi and Chen 2006].

2.2 Online Stylometric Analysis

Online stylometric analysis is concerned with categorization of authorship style in online texts. Here, we define "online texts" as any textual documents that may be found in an online setting. This includes computer-mediated communication (CMC), nonliterary electronic documents (e.g., student essays, news articles, etc.), and program code. Previous online studies have several important characteristics pertaining to the tasks, domains, features, and number of author classes utilized. These are summarized in the taxonomy presented in Table I.

Based on the proposed taxonomy, Table II shows previous studies dealing with online stylometric classification. For some previous studies, the number of features and categories used are marked with a dash ("-") or a not available ("n/a"). The dashes are for studies where authorship was evaluated manually, without the use of any defined set of features. For studies marked "n/a" the authors were unable to determine the number of features and categories used in the study. We discuss the taxonomy and related studies in detail in the following.

2.2.1 *Tasks.* As described in the previous section, two important stylometric analysis tasks are identification and similarity detection. For online

Table II. Previous Studies in Online Stylometric Analysis

	Tasks		Domains				Features			Classes
Previous Studies	T1	T2	D1	D2	D3	D4	Cat.	#	Type	#
Oman and Cook 1989		√				√	2	16	Group	6
Hayne and Rice 1997	√		√				—	—	Group	26
Krsul and Spafford 1997	√					√	3	49	Group	29
Stamatatos et al. 2000	√				√		3	22	Group	10
De Vel et al. 2001	√		√				7	191	Group	3
Chaski 2001	√	√			√		1	33	Both	4
Baayen et al. 2002	√				√		1	60	Group	8
Corney et al. 2002	√		√				10	221	Group	2
Argamon et al. 2003	√		√				5	506	Group	20
Diederich et al. 2003	√				√		3	120,000	Group	7
Hayne et al. 2003	√		√				—	—	Group	5
Koppel and Schler 2003	√		√				3	4,060	Group	11
Ding and Samadzadeh 2004	√					√	3	56	Group	46
Whitelaw and Argamon 2004	√		√		√		2	109	Group	3
Abbasi and Chen 2005	√		√				13	418	Group	5
Chaski 2005	√				√		3	6	Group	2
Juola and Baayen 2005	√	√			√		1	n/a	Group	2
Abbasi and Chen 2006	√		√				6	106	Group	10
Li et al. 2006	√		√				11	270	Group	5
Pan et al. 2006	√		√				4	56	Group	2
Zheng et al. 2006	√		√				11	270	Group	20

texts, these two tasks can be performed at the message/document or identity level [Pan 2006]. Message-level analysis attempts to categorize individual texts (e.g., emails), whereas identity-level analysis is concerned with classifying identities belonging to a particular entity. For example, let's assume that the entity John Smith has various email accounts (identities) in cyberspace (e.g., js@hotmail.com, john@yahoo.com, etc.). The message-level identification task may attempt to determine if an anonymous email was written by js@hotmail.com, while the identity-level identification task would attempt to determine whether js@hotmail.com and john@yahoo.com are identities belonging to the same entity.

The majority of previous studies focused on message-level analysis (e.g., De Vel et al. [2001], Abbasi and Chen [2005], Zheng et al. [2006]), which is useful for forensic applications with a small number of potential authors (e.g., Chaski [2001]). However, message-level analysis is not highly scalable to larger numbers of authors in cyberspace due to difficulties in consistently identifying texts shorter than 250 words [Forsyth and Holmes 1996]. Consequently, Zheng et al. [2006] noted a 14% drop in accuracy when increasing the number of author classes from 5 to 20 in their classification of forum postings. Argamon et al. [2003] also observed as much as a 23% drop in message classification accuracy when increasing the number of authors from 5 to 20.

Identity-level analysis attempts to categorize identities based on all texts written by that identity. It is somewhat less challenging than message-level categorization due to the presence of larger text samples, making identity-level analysis more suitable for cyber-content [Pan 2006]. Figure 1 presents

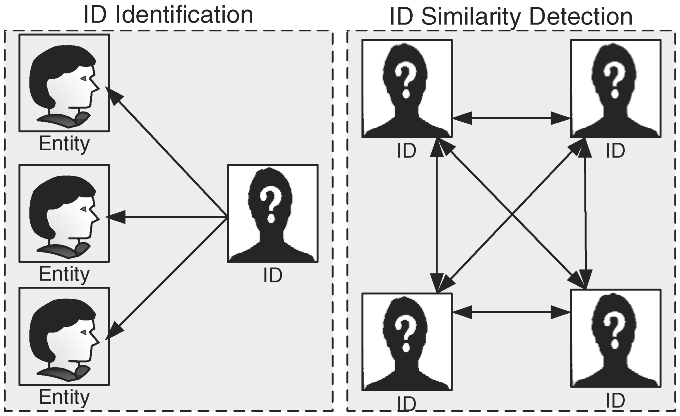

Fig. 1. Identity-level tasks.

illustrations of identity-level identification and similarity detection tasks. For ID identification, each anonymous identity is compared against all known entities. The identity is assigned to that entity with the highest similarity score (classification task). For ID similarity detection, each anonymous identity is compared to all other identities. Identities with a similarity score above a certain threshold are grouped together and considered to belong to the same entity (clustering task).

2.2.2 *Domains.* Online text includes various modes of computer-mediated communication (CMC), such as asynchronous and synchronous mediums [Herring 2002]. Relevant asynchronous modes for stylometric analysis are email, web forums, blogs, feedback/comments, etc. Many previous authorship studies focused on email (e.g., De Vel et al. [2001], Argamon et al. [2003]), forums (e.g., Abbasi and Chen [2005], Li et al. [2006]), and feedback comments [Hayne and Rice 1997; Hayne et al. 2003]. Synchronous forms of textual communication include instant messaging and chatrooms. We are unaware of any stylometric analysis relating to persistent conversation, despite its prevalence as a communication medium. The continuous nature of synchronous mediums makes them especially interesting since authors have less time to craft their responses [Hayne et al. 2003]. It is difficult to surmise without investigation what impact this may have on the ability to categorize authorship of persistent conversation.

Online documents encompass nonliterary texts, essays, and news articles. Electronic documents tend to be lengthier, more structured, and better written as compared to CMC text. Many previous studies focused on electronic documents with high levels of accuracy (e.g., Stamatatos et al. [2000], Chaski [2001], Whitelaw and Argamon [2004]).

The domain of program code is important for identifying hackers and attackers [Garson 2006], as well as detecting software plagiarism. Program-style analysis studies have developed programming-specific features, often tailored towards specific programming languages (e.g., Oman and Cook [1989], Krsul and Spafford [1997]).

2.2.3 *Features and Classes.* Feature sets used in previous online studies typically consist of a handful of categories and less than 500 features. Here, we define a "category" as a set of similar features (e.g., word-length distribution, punctuation, part-of-speech tag bigrams, etc.). Studies that did utilize larger feature sets typically incorporated only a couple of syntactic- or content-specific feature categories such as bag-of-words and part-of-speech bigrams [Koppel and Schler 2003; Diederich et al. 2003]. Consequently, online stylometric analysis has typically been applied to less than 20 authors, with only a few studies exceeding 25 authors (e.g., Krsul and Spafford [1997], Ding and Samadzadeh 2004]).

2.3 Feature-Set Types for Stylometry

Two types of feature sets have been used in previous research: author-group level and individual-author level. Most previous research used author-group-level sets where one set of features is applied across all authors. In contrast, individual-author-level sets consist of a feature set for each author (e.g., 10 authors = 10 feature sets). For instance, Peng et al. [2003] created a feature set of the 5,000 most frequently used character *n*-grams for each author, based on that author's usage. Similarly, Chaski [2001] developed author-level feature sets for misspelled words, where each author's feature set consisted of words they commonly misspelled. Individual-author-level feature sets can be effective when using feature categories with large potential feature spaces, such as *n*-grams or misspellings [Peng et al. 2003]. However, the use of individual-author-level sets requires techniques that can handle multiple feature sets. Standard machine learning techniques typically build a classifier using only a single feature set.

2.3.1 *Individual-Level Techniques.* Two multiple feature-set techniques that have been utilized for pattern recognition and stylistic analysis are ensemble classifiers and the Karhunen-Loeve transform. Ensemble classifiers consist of supervised techniques that can be incorporated for the stylometric identification task. They use multiple classifiers with each built using different techniques, training instances, or feature subsets [Dietterich 2000]. Ensembles are effective for analyzing large data streams [Wang et al. 2003]. Particularly, the feature-subset classifier approach has been shown effective for analysis of style and patterns. Stamatatos and Widmer [2002] used an SVM ensemble for music performer recognition. They used multiple SVMs, each trained using different feature subsets. Similarly, Cherkauer [1996] used a neural-network ensemble for imagery analysis. The ensemble of the two aforementioned papers consisted of 32 neural networks trained on 8 different feature subsets. The intuition behind using an ensemble is that it allows each classifier to act as an "expert" on its particular subset of features [Cherkauer 1996; Stamatatos and Widmer 2002], thereby improving performance over simply using a single classifier. For stylometric analysis, building a classifier trained using a particular author's features could allow it to become an "expert" on identifying that author against others.

Karhunen-Loeve (K-L) transforms are a supervised form of principal-component analysis (PCA) that allows inclusion of class information in the transformation process [Webb 2002]. K-L transforms have been used in several pattern recognition studies (e.g., Kirby and Sirovich [1990], Uenohara and Kanade [1997]). Like PCA, the K-L transforms consist of a dimensionality reduction technique where the transformation is done by deriving the basis matrix (set of eigenvectors) and then projecting the feature usage matrix into a lower-dimension space. PCA captures the variance across a set of authors (interclass variance) using a single feature set and basis matrix. In contrast, K-L transforms can be applied to each individual author (intraclass variance) by only considering that author's feature set and basis matrix. Thus, K-L transforms can be used as an individual-level similarity detection technique where identity A's variance pattern (extracted using A's feature set and basis matrix) can be compared against identity B's variance pattern (extracted using B's feature set and basis matrix). However, when comparing identity A to identity B, we must evaluate A using B's features and basis matrix, and B using A's features and basis matrix. Two comparisons are necessary due to the use of different feature sets for each individual identity.

3. RESEARCH GAPS AND QUESTIONS

Based on our review of previous literature, we have identified several important research gaps.

3.1 Similarity Detection

Most studies have focused on the identification task, with less emphasis on similarity detection. Similarity detection is important for cyberspace, since class definitions are often not known a priori. There is a need for techniques that can perform identification and similarity detection.

3.2 Richer Feature Sets

Previous feature sets lack either the necessary breadth (number of categories) or depth (number of features). It is difficult to apply such feature sets to larger numbers of authors with a high level of accuracy. Consequently, previous research has typically used less then 20 author classes in experiments. However, application of stylometric methodologies to cyber-content necessitates the ability to discriminate authorship across larger sets of authors.

3.3 Individual-Author-Level Feature Sets

Few online studies have incorporated multiple individual-author-level feature subsets, despite their effective application to other areas of style- and pattern recognition. The use of such feature sets along with techniques that can support individual-author-level attributes could improve authorship categorization performance and scalability.

3.4 Scalability Across Domains

Little work has been done to assess the effectiveness of features and techniques across domains. Prior work mostly focused on a single domain (e.g., email or documents). Furthermore, we are unaware of any studies applied to synchronous communication (e.g., instant messaging). Analysis across domains is important in order to evaluate the robustness of stylometric techniques for various modes of CMC.

3.5 Research Questions

Based on the gaps described, we propose the following research questions.

(1) Which authorship analysis techniques can be successfully used for online identification and similarity detection tasks?

(2) What impact will the use of a more holistic feature set have on online classification performance?

(3) Will the use of multiple individual-author-level feature subsets improve online attribution accuracy as compared to using a single author-group-level feature set?

(4) How scalable are these features and techniques with respect to the various domains and in terms of number of authors?

4. RESEARCH DESIGN: AN OVERVIEW

In order to address these questions, we propose the creation of a stylometric analysis technique that can perform ID-level identification and similarity detection. Furthermore, a more holistic feature set consisting of a larger number of features across several categories is utilized in order to improve our representational richness of authorial style. We plan to utilize two variations of this extended feature set: at the author-group and individual-author levels. Our approach will be evaluated across multiple domains in comparison with benchmark techniques and feature sets. The proposed technique, feature sets, and feature types, as well as comparison benchmarks, are discussed in the following.

4.1 Techniques

We propose the development of the Writeprints technique, which is an unsupervised method that can be used for identification and similarity detection. Writeprints is a Karhunen-Loeve-transforms-based technique that uses a sliding window and pattern disruption to capture feature usage variance at a finer level of granularity. A sliding window was incorporated, since it has been shown effective in previous authorship studies [Kjell et al. 1994; Abbasi and Chen 2006]. The technique uses individual-author-level feature sets where a Writeprint is constructed for each author using the author's key features. The use of individual-author-level feature sets is intended to provide greater scalability as compared to traditional machine learning techniques that only utilize a single author-group-level set (e.g., SVM, PCA). For all features that an author uses, Writeprints patterns project usage variance into a lower-dimension space, where each pattern point reflects a single window instance. All key attributes in an author's feature set that the author never uses are treated as pattern disruptors, where the occurrence of these features in an anonymous identity's text decrease the similarity between the anonymous identity and the author.

For the identification task, we plan to compare the Writeprints method against SVM and the Ensemble SVM classifier. SVM is a benchmark technique used in several previous online stylometric identification studies (e.g., De Vel et al. [2001], Zheng et al. [2006], Li et al. [2006]). A single classifier is built using an author-group-level feature set. In contrast, ensemble classifiers provide flexibility for using multiple individual-author-level feature sets [Cherkauer 1996; Dietterich 2000]. SVM ensembles with multiple feature subsets have been shown effective for stylistic classification [Stamatatos and Widmer 2002].

For the similarity detection task, we plan to compare the Writeprints method against PCA and Karhunen-Loeve transforms. PCA has been used in numerous previous stylometric analysis studies (e.g., Baayen et al. [2002], Abbasi and Chen [2006]). In PCA, the underlying usage variance across a single author-group-level feature set is extracted by deriving a basis of eigenvectors that are used to transform the feature space to a lower-dimensional representation/pattern [Binogo and Smith 1999]. The distance between two identities' patterns can be used to determine the degree of stylistic similarity. K-L transforms are a PCA variant often used in pattern recognition studies [Watanbe 1985; Webb 2002] and provide a mechanism for using multiple individual-author-level feature sets. As previously mentioned, the use of different feature sets and basis matrices for each author in K-L transforms entails two comparisons for each set of identities (A using B's features and basis, and vice versa). Specific details about Writeprints against comparison identification and similarity detection techniques are provided in the system design discussion in Section 5.

4.2 Feature Sets and Types

The use of an extended set of features could improve the scalability of stylometric analysis by allowing greater discriminatory potential across larger sets of authors. We propose the development of a holistic feature set containing more feature categories (breadth) and numbers of features (depth) intended to improve performance and scalability. Our extended feature (EF) set contains several static and dynamic feature categories across various groups (i.e., lexical, syntactic, structural, content-specific, and misspellings). Static features include well-defined context-free categories such as function words, word-length distributions, vocabulary richness measures, etc. In contrast, dynamic feature categories are context-dependent attributes, such as *n*-grams (e.g., word-, character-, POS tag-, and digit-level) and misspelled words. These categories have infinite potential feature spaces, varying based on the underlying text corpus. As a result, dynamic feature categories usually include some form of feature selection in order to extract the most important style markers for particular authors and text [Koppel and Schler 2003]. We utilized the information-gain heuristic due to its effectiveness in previous text categorization [Efron et al. 2004] and authorship analysis research [Koppel and Schler 2003]. In order to evaluate the effectiveness of our extended feature set (EF), we plan to compare its performance against a baseline feature set (BF) commonly used in previous online stylometric analysis research [De Vel et al. 2001; Corney et al. 2002; Abbasi and Chen 2005; Zheng et al. 2006; Li et al. 2006]. The

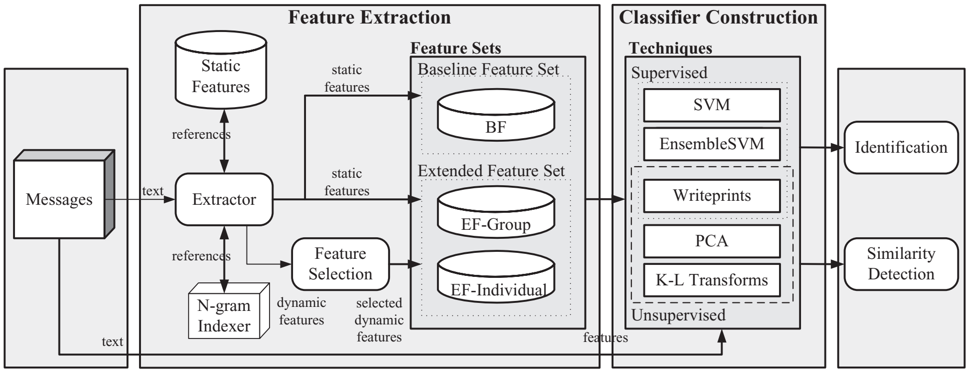

Fig. 2. Stylometric analysis system design.

baseline set (BF) consists of static lexical, syntactic, structural, and content-specific features used for categorization of up to 20 authors. Further details about the two feature sets (EF and BF), extraction, and feature selection procedures are discussed in the system design section.

4.2.1 *Feature-Set Types.* Based on the success of multiple feature-subset approaches, we propose to compare the effectiveness of the author-group-level feature-set approach used in most previous studies against the use of multiple, individual-identity-level feature sets. Thus, our extended feature set (EF) will be used as a single group-level set (EF-Group), or multiple individual-level subsets will be selected (EF-Individual).

5. SYSTEM DESIGN

We propose the following system design (shown in Figure 2). Our design has two major steps: feature extraction and classifier construction. These steps are used to carry out identification and similarity detection of online texts.

5.1 Feature Extraction

The extraction phase begins with a data preprocessing step where all message signatures are initially filtered out in order to remove obvious identifiers [De Vel et al. 2001]. This step is particularly important for email data where authors often include signatures such as name, job title, address, position, contact information, etc. The next step involves extraction of the static and dynamic features, resulting in the creation of our feature sets. We included two feature sets: a baseline feature set (BF) consisting of static author-group-level features and an extended feature set (EF) consisting of static and dynamic features. For static features, extraction simply involves generating the feature usage statistics (feature vectors) across texts; however, dynamic feature categories such as *n*-grams require indexing and feature selection. The feature extraction procedures for the two feature sets (BF and EF) are described next, while Table III provides a description of the two feature sets. For dynamic feature categories,

Table III. Baseline and Extended Feature Sets

		Quantity		
Group	Category	Baseline (BF)	Extended (EF)	Description
Lexical	Word-Level	5	5	total words, % char. per word
	Character-Level	5	5	total char., % char. per message
	Letters	26	26	count of letters (e.g., a, b, c)
	Character Bigrams	—	<676	letter bigrams (e.g., aa, ab, ac)
	Character Trigrams	—	<17,576	letter trigrams (e.g., aaa, aab, aac)
	Digits	—	10	digits (e.g., 1, 2, 3)
	Digit Bigrams	—	<100	2 digit number frequencies (e.g., 10, 11)
	Digit Trigrams	—	<1,000	frequency of 3 digit numbers (e.g., 100)
	Word Length Dist.	20	20	frequency of 1–20 letter words
	Vocab. Richness	8	8	richness (e.g., hapax legomena, Yule's K)
	Special Characters	21	21	occurrence of special char. (e.g., @#$%ˆ)
Syntactic	Function Words	150	300	frequency of function words (e.g., of, for)
	Punctuation	8	8	occurrence of punctuation (e.g., !;:,.?)
	POS Tags	—	<2,300	frequency of POS tags (e.g., NP, JJ)
	POS Tag Bigrams	—	varies	POS tag bigrams (e.g., NP VB)
	POS Tag Trigrams	—	varies	POS tag trigrams (e.g., VB JJ)
Structural	Message-Level	6	6	e.g., has greeting, has url, quoted content
	Paragraph-Level	8	8	e.g., no. of paragraphs, paragraph lengths
	Technical Structure	50	50	e.g., file extensions, fonts, use of images
Content	Words	20	varies	bag-of-words (e.g., "senior", "editor")
	Word Bigrams	—	varies	word bigrams (e.g. "senior editor")
	Word Trigrams	—	varies	word trigrams (e.g., "editor in chief")
Idiosyncratic	Misspelled Words	—	<5,513	misspellings (e.g., "beleive", "thougth")

the number of features varies depending on the indexing and feature selection for a specific dataset, as well as whether the author-group (EF-Group) or individual-author (EF-Individual) level is being used for feature selection. For some such categories, the upper limit of features is already known (e.g., number of character bigrams is less than 676).

5.1.1 *Baseline Feature Set (BF).* This feature set contains 327 lexical, syntactic, structural, and content-specific features. Variants of this feature set have been used in numerous previous studies (e.g., De Vel et al. [2001], Corney et al. [2002], Abbasi and Chen [2005], Li et al. [2006], Zheng et al. [2006]). Since this feature set is devoid of any dynamic feature categories (e.g., *n*-grams, misspellings), it has a fairly straightforward extraction procedure.

5.1.2 *Extended Feature Set (EF).* The extended feature set is comprised of a mixture of static and dynamic features. The dynamic features include several *n*-gram feature categories and a list of 5,513 common word misspellings taken from various websites, including Wikipedia (www.wikipedia.org). The *n*-gram categories we utilized include character-, word-, POS tag-, and digit-level *n*-grams. The POS tagging was conducted using the Arizona noun-phrase extractor [McDonald et al. 2005], which uses the Penn Treebank tag set and also performs noun-phrase chunking and named entity recognition and tagging. These *n*-gram-based categories require indexing, with the number of initially indexed features varying depending on the dataset. The indexed features are then sent forward to the feature selection phase. Use of such an indexing and feature selection/filtering procedure for *n*-grams is quite necessary and common in stylometric analysis research (e.g., Peng et al. [2003], Koppel and Schler [2003]).

Feature selection is applied to all the *n*-gram and misspelled word categories using the information-gain (IG) heuristic. IG has been used in many text categorization studies as an efficient method for selecting text features (e.g., Forman [2003], Efron et al. [2004], Koppel and Schler [2003]). Specifically, it is computationally efficient compared to search-based techniques [Dash and Liu 1997; Guyon and Elisseef 2003] and good for multiclass text problems [Yang and Pederson 1997]. IG is applied at the author-group and individual-author levels. The information gain for feature j across a set of classes c is derived as $\mathrm{IG(c,j)} = \mathrm{H(c)} - \mathrm{H(c|j)}$, where H(c) is the overall entropy across author classes and H(c|j) is the conditional entropy for feature j. For the author-group-level feature set (EF-Group), IG is applied across all author classes (size of $c =$ no. authors). For individual-identity-level feature sets (EF-Individual), IG is applied using a 2-class (one-against-all) setup (size of $c = 2$, = identity, = rest). The EF-Group feature set is intended for utilizing the set of features that can best discriminate authorship across all authors, while each EF-Individual feature set attempts to find the set of features most effective at differentiating a specific author against all others.

5.2 Classifier Construction

5.2.1 *Writeprints Technique.* The Writeprints technique has two major parts: creation and pattern disruption. The creation part is concerned with those steps relating to the construction of patterns reflective of an identity's writing-style variation. In this step, Karhunen-Loeve transforms are applied with a sliding window in order to capture stylistic variation with a finer level of granularity. The pattern disruption part describes how zero usage features can be used as red flags to decrease the level of stylistic similarity between

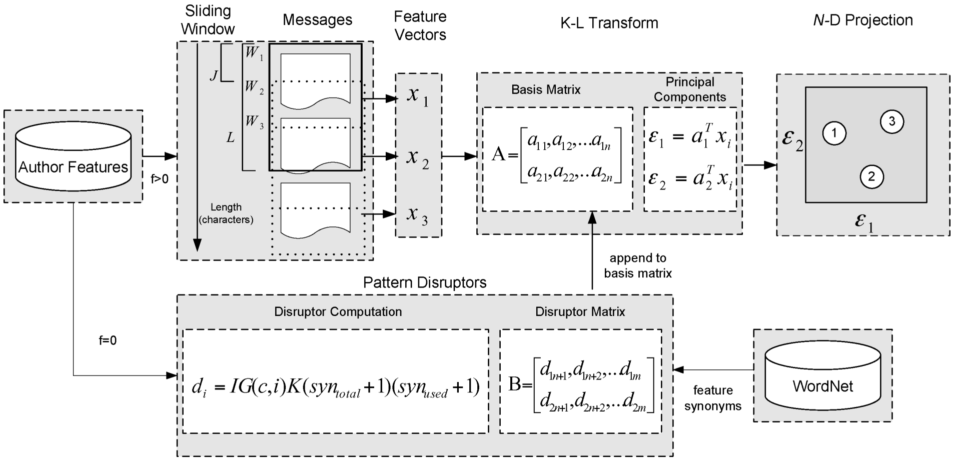

Fig. 3. Writeprints creation illustration.

identities. The two major steps, which are repeated for each identity, are shown next.

Algorithm. Writeprint Steps

```
1) For all identity features with occurrence frequency $> 0$.
        a) Extract feature vectors for each sliding window instance.
        b) Derive basis matrix (set of eigenvectors) from feature usage covariance
           matrix using Karhunen-Loeve transforms.
        c) Compute window instance coordinates (principal components) by
           multiplying window feature vectors with basis. Window instance points
           in $n$ dimensional space represent author Writeprint pattern.
2) For all author features with occurrence frequency $= 0$.
        a) Compute feature disruption value as product of information gain, synonymy
           usage, and disruption constant $K$.
        b) Append features' disruption values to basis matrix.
        c) Apply disruptor based on pattern orientations.
3) Repeat steps 1-2 for each identity.
```

Figure 3 presents an illustration of the Writeprints process, while these steps are described in greater detail in the following.

Step 1 (*Writeprints Creation*). A lower-dimensional usage variation pattern is created based on the occurrence frequency of the identity's features (individual-level feature set). For all features with usage frequency greater than zero, a sliding window of length L with a jump interval of J characters is run over the identity's messages. The feature occurrence vector for each window is projected to an n-dimensional space by applying the Karhunen-Loeve transform. The Kaiser-Guttman stopping rule [Jackson 1993] is used to select the number of eigenvectors in the basis. The formulation for step 1 is presented next.

Let $\Omega = \{1, 2, \ldots, f\}$ denote the set of f features with frequency greater than 0 and let $\Phi = \{1, 2, \ldots, w\}$ represent the set of w text windows. Let X

denote the author's feature matrix, where x_{ij} is the value of feature $j \in \Omega$ for window $i \in \Phi$.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1f} \\ x_{21} & x_{22} & \dots & x_{2f} \\ \dots & \dots & \dots & \dots \\ x_{w1} & x_{w2} & \dots & x_{wf} \end{bmatrix}$$

Extract the set of eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ for the covariance matrix Σ of the feature matrix X by finding those points where the characteristic polynomial of Σ equals 0.

$$p(\lambda) = \det(\Sigma - \lambda I) = 0$$

For each eigenvalue $\lambda_m > 1$, extract its eigenvector $a_m = (a_{m1}, a_{m2}, \dots, a_{mf})$ by solving the following system, resulting in a set of n eigenvectors $\{a_1, a_2, \dots, a_n\}$.

$$(\Sigma - \lambda_m I)a_m = 0$$

Compute an n-dimensional representation for each window i by extracting principal component scores ε_{ik} for each dimension $k \leq n$.

$$\varepsilon_{ik} = a_k^T x_i$$

Step 2 (*Pattern Disruption*). Since Writeprints uses individual-author-level feature sets, an author's key set of features may contain attributes that are significant because the author never uses them. However, features with no usage will currently be irrelevant to the process, since they have no variance. Nevertheless, these features are still important when comparing an author to other anonymous identities. The author's lack of usage of these features represents an important stylistic tendency. Anonymous identity texts containing these features should be considered less similar (since they contain attributes never used by this author).

As previously mentioned, when comparing two identities' usage variation patterns, two comparisons must be made since both identities used different feature sets and basis matrices in order to construct their lower-dimensional patterns. The dual comparisons are illustrated in Figure 4. We would need to construct a pattern for identity B using B's text with A's feature set and basis matrix (Pattern B) as a comparison against identity A's Writeprint (and vice versa). The overall similarity between identities A and B is the sum of the average n-dimensional Euclidean distance between Writeprint A and pattern B and Writeprint B and pattern A. When making such a comparison we would like A's zero-frequency features to act as "pattern disruptors," where the presence of these features in identity B's text decreases the similarity for the particular A-B comparison. It's less likely that identity A wrote text containing features that identity A never uses.

Such disruption can be achieved by appending a nonzero coefficient d in identity A's basis matrix for such features. Let $\Psi = \{f + 1, f + 2, \dots, f + g\}$ denote the set of g features with zero frequency. For each feature $p \in \Psi$, append the value d_{kp} to each eigenvector a_k, where $k \leq n$. Let's assume that one of identity A's key attributes is the word "folks," which is important because identity A

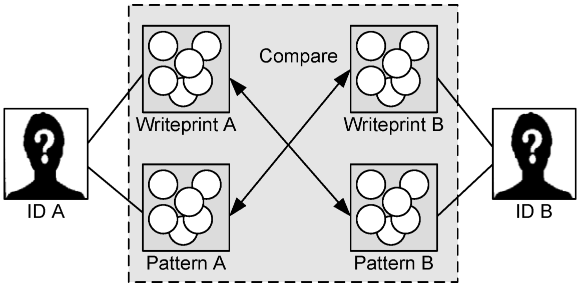

Fig. 4. Writeprints comparisons.

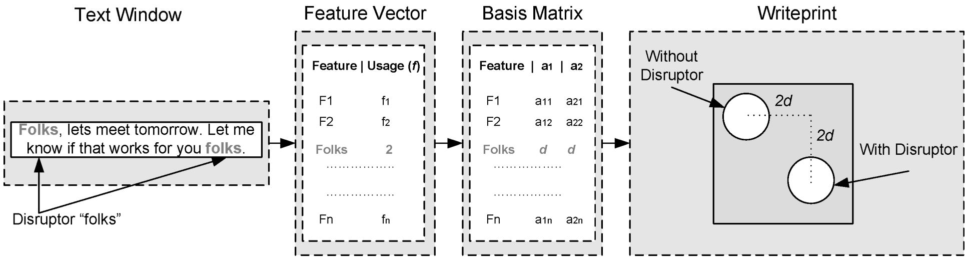

Fig. 5. Illustration of pattern disruption

never uses it. Figure 5 shows how pattern disruption can reduce the similarity between identities A and B by shifting away identity B's pattern points for text windows containing the word "folks". In this example, the value d is substituted as the coefficient for feature number 3 ("folks") in identity A's primary two eigenvectors (a_{13}, a_{23}). The direction of a window point's shift is intended to reduce the similarity between the Writeprint and comparison pattern. This is done by making d_{kp} positive or negative for a particular dimension k, based on the orientation of the Writeprint (WP) and comparison pattern (PT) points along that dimension, as follows.

$$d_{kp} = \begin{cases} -d_{kp}, & \text{if } \sum_{i=1}^{w} \frac{WP_{ik}}{w} > \sum_{i=1}^{w} \frac{PT_{ik}}{w} \\ d_{kp}, & \text{if } \sum_{i=1}^{w} \frac{WP_{ik}}{w} < \sum_{i=1}^{w} \frac{PT_{ik}}{w} \end{cases}$$

For instance, if identity A's Writeprint is spatially located to the left of identity B's pattern for dimension k, the disruptor d_{kp} will be positive in order to ensure that the disruption moves the comparison pattern away from the Writeprint, as opposed to towards it.

The magnitude of d signifies the extent of disruption for a particular feature. Larger values of d will cause pattern points representing text windows containing the disruptor feature to be shifted further away. However, not all features are equally important discriminators. For example, lack of usage of the word "Colorado" is less significant than lack of usage of the word "folks," because "Colorado" is a noun conveying topical information. Lack of usage of "Colorado" simply means this author doesn't talk about Colorado, and is not indicative of stylistic choice. It is more reflective of context than style. In contrast, lack of use of "folks" (a function word used to address people) is a stylistic tendency. It is possible and likely that the author uses some other word (synonym of "folks") to address people, or doesn't address them at all. Koppel et al. [2006] developed a machine-translation-based technique for measuring the degree of feature "stability." Stability refers to how often a feature changes across authors and documents for a constant topic. They found nouns to be more stable than function words and argued that function words are better stylistic discriminators than nouns, since use of function words involves making choices between a set of synonyms. Based on this intuition we devised a formula for the disruptor coefficient d for feature p. Our formula considers the feature's information gain (IG) and synonymy usage. Specifically,

$$d_p = IG(c,p)K(syn_{total}+1)(syn_{used}+1),$$

where $IG(c,p)$is the information gain for feature p across the set of classes, c, and syn_{total} and syn_{used} are the total number of synonyms and the number used by the author, respectively, for the disruptor feature. The feature synonym information is derived from Wordnet [Fellbaum 1998]. Synonym information is only used for word-based features (e.g., word n-grams, function words). For other feature category disruptors, syn_{total} and syn_{used} will equal 0. Moreover, K is a disruptor constant used to control the magnitude and aggressiveness of the pattern disruption mechanism. We used integer values between 1 and 10 for K and generally attained the best results using a value of 2. As previously mentioned, each disruptor is applied in such a manner as to shift the comparison print further away from the Writeprint.

5.2.2 *Comparison Identification and Similarity Detection Techniques.* For all comparison techniques, feature vectors are derived for nonoverlapping 1,500-character blocks of text from each identity's text. This particular length was used since it corresponds to approximately 250 words, the minimum text length considered effective for authorship analysis [Forsyth and Holmes 1996].

In addition to the Writeprints method, SVM and Ensemble SVM are utilized as comparison identification techniques. SVM is run using a linear kernel with a sequential minimal optimization (SMO) algorithm [Platt 1999]; these are the same settings as in numerous previous studies (e.g., Zheng et al. [2006], Li et al. [2006]. For Ensemble SVM we build multiple classifiers (one for each identity's features). Anonymous identities are assigned by aggregating results across classifiers.

For similarity detection, PCA and K-L transforms are used. For PCA we extract the basis matrix for a single author-group-level feature set where the feature matrix contains vectors across identity classes. Thus, PCA captures the interauthor feature usage variation for a common set of features. In contrast, for K-L transforms the basis matrix is extracted for each individual identity using the identity's feature set and occurrence vectors. Each author basis matrix thus captures the intraauthor feature usage variation.

Table IV. Details for Datasets in Testbed

Data Set	Domain	No. Authors	Words (per Author)	Time Duration	Noise
Enron Email	Asynchronous (D1)	100	27,774	10/98–09/02	Yes
EBay Comments	Asynchronous (D1)	100	23,423	02/03–04/06	No
Java Forum	Program Code (D4)	100	43,562	04/03–05/06	Yes
CyberWatch Chat	Synchronous (D2)	100	1,422	05/04–08/06	No

6. EVALUATION

In order to evaluate the effectiveness of the Writeprints technique and extended feature set (EF), two experiments were conducted. The experiments compared the extended features (EF) and Writeprints technique against our comparison techniques and baseline feature set (BF). The experiments were conducted for the identification and similarity detections tasks across testbeds from various domains. The testbeds and experiments are described next.

6.1 Testbed

The testbed consists of four datasets spanning asynchronous, synchronous, and program code domains. This first dataset is composed of email messages from the publicly available Enron email corpus. The second test set consists of buyer/seller feedback comments extracted from eBay (www.ebay.com). The third dataset contains programming code snippets taken from the Sun Java Technology Forum (forum.java.sun.com), while the fourth set of data consists of instant messaging chat logs taken from CyberWatch (www.cyberwatch.com). Table IV provides some details about the testbed. For each dataset, we randomly extracted 100 authors. The datasets also differ in terms of average amount of text per author, time span, and amount of noise. The email and forum datasets have greater noise due to the presence of requoted and forwarded content (which is not always easy to filter out). CyberWatch chat logs contain the least amount of text, since each author's text is only a single conversation.

6.2 Experiment 1: Identification Task

6.2.1 *Experimental Setup.* For the identification task, each author's text was split into two identities: one known and one anonymous identity. All techniques were run using tenfold cross-validation by splitting author texts into 10 parts (5 for known entity, 5 for anonymous identity). For example, in fold 1, parts 1–5 are used for the known entity, while parts 6–10 are for the anonymous identity; in fold 2, in parts 2–6 are used for the known entity while parts 1 and 7–10, are for the anonymous identity. The overall accuracy is comprised the average classification accuracy across all 10 folds of where the classification accuracy is computed as follows.

$$\text{Classification Accuracy} = \frac{\text{Number of Correctly Classified Identities}}{\text{Total Number of Identities}}$$

Four combinations of feature sets, feature types, and techniques were used (shown in Table V). As shown in the fifth row in Table V, a baseline was included

Table V. Techniques/Feature Sets for Identification Experiment

Label	Technique	Feature Set Type	Feature Set
Writeprint	Writeprint	Individual	EF
Ensemble	Ensemble SVM	Individual	EF
SVM/EF	SVM	Group	EF
Baseline	SVM	Group	BF

which featured the use of SVM with the group-level baseline feature set (BF). This particular baseline consisted of the same combination of features and techniques used in numerous previous studies (e.g., De Vel et al. [2001], Abbasi and Chen [2005], Li et al. [2006], Zheng et al. [2006]). The baseline was intended to be compared against the use of SVM with the group-level extended feature set (SVM/EF, as shown in-row the fourth) in order to assess the effectiveness of a more holistic feature set for online identification (fourth row versus fifth row in Table V). We also wanted to evaluate the effectiveness of individual-author-level feature sets by comparing Ensemble SVM using EF-Individual against the SVM/EF method, which uses a single group-level feature set (third row versus fourth row in Table V). Finally, the Writeprints technique was included with the extended feature set in order to evaluate the effectiveness of this technique in comparison with Ensemble SVM and SVM/EF (second row versus third and fourth row in Table V).

6.2.2 *Hypotheses.*

—*H*1*a* (*Feature Sets*). The use of a more holistic feature set with a larger number of features and categories (EF) will outperform the baseline feature set (BF). Thus, SVM/EF will outperform the baseline.

—*H*1*b* (*Feature-Set Types*). The use of individual-author-level feature subsets (EF-Individual) will outperform the use of a single author-group-level feature set (EF-Group). Thus Ensemble SVM will outperform SVM/EF.

—*H*1*c* (*Techniques*). The Writeprint technique will outperform SVM (SVM/EF and Ensemble SVM).

6.2.3 *Experimental Results.* Table VI shows the experimental results for all four combinations of features and techniques across the four datasets. The Writeprints technique had the best performance on the email, comments, and chat datasets. Furthermore, individual-author-level feature set techniques (Writeprints and Ensemble) had higher accuracy on these datasets than author-group-level feature set methods (SVM/EF and the baseline). However, Writeprints performed poorly on the programming forum dataset. This is attributable to the inability of the variation patterns and disruptors to effectively capture programming style. The extended feature set (EF) had better performance than the benchmark feature set (BF), as reflected by the fact that all techniques using EF had higher accuracy than the baseline on all datasets.

6.2.4 *Hypotheses Results.* Table VII shows the p-values for the pair-wise *t*-tests conducted on the classification accuracies in order to measure the

Table VI. Experimental Results (% accuracy) for Identification Task

		No. Authors		
Test Bed	Techniques/Features	**25**	**50**	**100**
Enron Email	Writeprint	**92.0**	**90.4**	**83.1**
	Ensemble	88.0	88.2	76.7
	SVM/EF	87.2	86.6	69.7
	Baseline	64.8	54.4	39.7
eBay Comments	Writeprint	**96.0**	**95.2**	**91.3**
	Ensemble	**96.0**	94.0	90.9
	SVM/EF	95.6	93.8	90.4
	Baseline	90.6	86.4	83.9
Java Forum	Writeprint	88.8	66.4	52.7
	Ensemble	92.4	85.2	**53.5**
	SVM/EF	**94.0**	**86.6**	41.1
	Baseline	84.8	60.2	23.4
CyberWatch Chat	Writeprint	**50.4**	**42.6**	**31.7**
	Ensemble	46.0	36.6	22.6
	SVM/EF	40.0	33.3	19.8
	Baseline	37.6	30.8	17.5

statistical significance of the results. Values in boldface indicate statistically significant outcomes that are in line with our hypotheses. Values with a plus sign indicate significant outcomes contradictory to our hypotheses.

—*H*1*a* (*Feature Sets*). The extended feature set (EF) outperformed the baseline feature set (BF) across all datasets ($p < 0.01$) based on the better performance of SVM/EF as compared to the baseline.

—*H*1*b* (*Feature-Set Types*). Individual-author-level feature subsets (EF-Individual) significantly outperformed the group-level feature set (EF-Group) on the Enron and CyberWatch datasets ($p < 0.01$). This outcome is based on the better performance of the Ensemble technique as compared to SVM/EF. EF-Individual also outperformed the EF-Group feature set on the eBay dataset, but not significantly.

—*H*1*c* (*Techniques*). The Writeprints technique significantly outperformed SVM (Ensemble and SVM/EF) on the Enron and CyberWatch datasets ($p < 0.01$). Writeprints also outperformed SVM on the eBay dataset, but not significantly.

6.2.5 *Results Discussion.* The Enron email dataset feature set sizes and SVM techniques' performance are shown in Figure 6. The number of features for EF-Individual is the average of each author's feature set. The increased number of authors caused the EF-Group feature set to grow at an increasing rate. This resulted in a decreased number of relevant features per author in EF-Group, as evidenced by the widening gap between EF-Individual and EF-Group as the number of authors grew to 50 and 100.

Consequently, the ensemble SVM technique significantly outperformed SVM/EF for experiments involving a larger number of authors (50 and 100). This is shown in Table VIII, which presents results for the Enron dataset. We can see that when using only 25 authors, the EF-Individual feature set marginally outperformed EF-Group, as illustrated by the slightly better

Table VII. P-Values for Pair-Wise t-Tests on Accuracy

		No. Authors		
Test Bed	Techniques/Features	25	50	100
Enron Email	Writeprint vs. Ensemble	<0.001**	0.002**	<0.001**
	Writeprint vs. SVM/EF	<0.001**	<0.001**	<0.001**
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	Ensemble vs. SVM/EF	0.330	0.049*	<0.001**
	Ensemble vs. Baseline	<0.001**	<0.001**	<0.001**
	SVM/EF vs. Baseline	<0.001**	<0.001**	<0.001**
eBay Comments	Writeprint vs. Ensemble	0.500	0.100	0.134
	Writeprint vs. SVM/EF	0.673	0.167	0.101
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	Ensemble vs. SVM/EF	0.673	0.772	0.339
	Ensemble vs. Baseline	<0.001**	<0.001**	<0.001**
	SVM/EF vs. Baseline	<0.001**	<0.001**	<0.001**
Java Forum	Writeprint vs. Ensemble	0.002+	< 0.001+	0.309
	Writeprint vs. SVM/EF	<0.001+	<0.001+	<0.001**
	Writeprint vs. Baseline	0.005**	<0.001**	<0.001**
	Ensemble vs. SVM/EF	0.097	0.166	<0.001**
	Ensemble vs. Baseline	<0.001**	<0.001**	<0.001**
	SVM/EF vs. Baseline	<0.001**	<0.001**	<0.001**
CyberWatch Chat	Writeprint vs. Ensemble	<0.001**	0.052	0.004**
	Writeprint vs. SVM/EF	<0.001**	<0.001**	<0.001**
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	Ensemble vs. SVM/EF	<0.001**	0.155	0.008**
	Ensemble vs. Baseline	<0.001**	0.064	<0.001**
	SVM/EF vs. Baseline	<0.001**	<0.001**	<0.001**

*P-values significant at alpha = 0.05.
**P-values significant at alpha = 0.01.
+P-values contradict hypotheses.

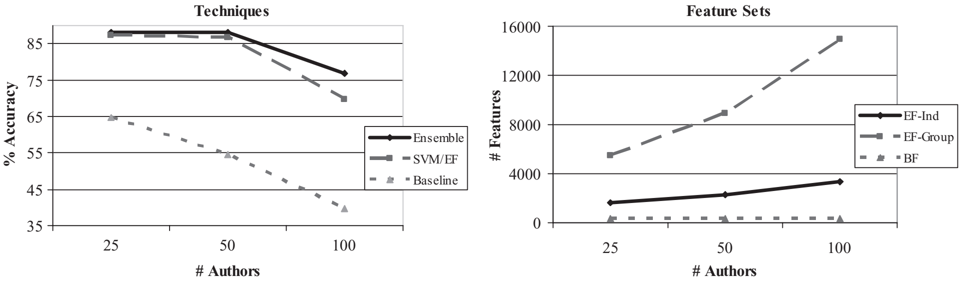

Fig. 6. Enron dataset feature set sizes and SVM technique performances.

performance of the Ensemble over SVM/EF. However, when the number of authors increased to 100, the widening gap in terms of number of features in each feature set caused the Ensemble technique to significantly outperform SVM/EF.

The Writeprints technique significantly outperformed SVM (Ensemble, SVM/EF, and the baseline) on the email and chat datasets. For most datasets, the Writeprints technique also had a smaller dropoff in accuracy as the number of authors increased. This is shown in Figure 7, which presents the performance accuracy for each technique across datasets and authors. The Writeprints technique appears to be more scalable as the number of authors increases, based

Table VIII. Performance Comparison of Ensemble SVM and SVM on Enron Dataset

No. Authors	Ensemble	SVM/EF	Difference
25	88.0%	87.2%	0.8%
50	88.2%	86.6%	1.6%*
100	76.7%	69.7%	7.0%**

*P-values significant at alpha = 0.05.
**P-values significant at alpha = 0.05.

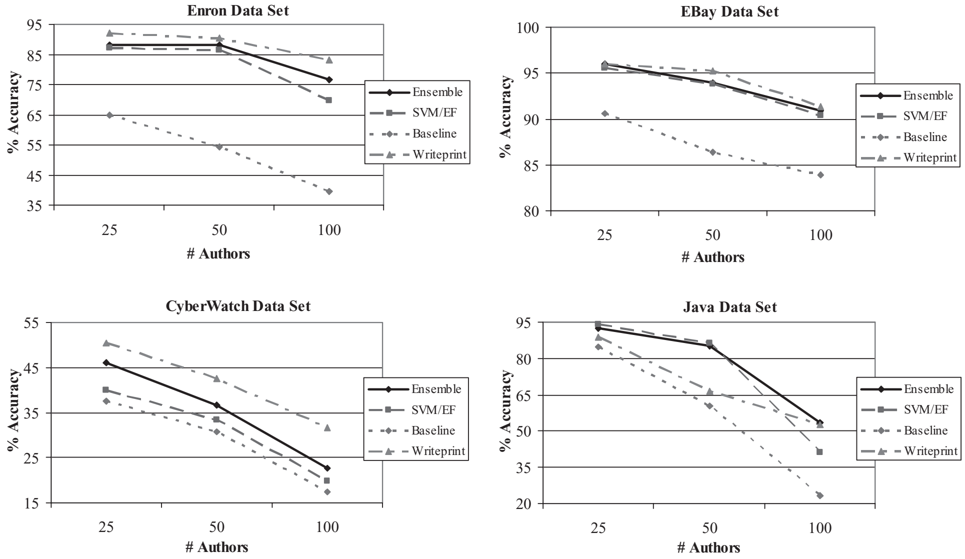

Fig. 7. Performance for identification techniques across testbeds.

on the fact that the slope of its accuracy line typically remains consistent. In contrast, other techniques' accuracy decreases more sharply as the number of authors goes from 25 to 50 or 100. We believe this is attributable to the disruptors effectively differentiating authorship across larger numbers of identities in the Writeprints technique. For the programming dataset, however, the disruptors were less effective due to the differences in program code as opposed to other forms of text. These differences are expounded upon in the similarity detection experimental results discussion.

6.3 Experiment 2: Similarity Detection Task

6.3.1 *Experimental Setup.* For the similarity detection task, each author's text was split into two anonymous identities. All techniques were run using tenfold cross-validation in the same manner as in the previous experiment. A trial-and-error method was used to find a single optimal similarity threshold for matching. The same threshold was used for all techniques. All identity-identity pair scores above the predefined threshold were considered a match. Trial-and-error methods for finding optimal thresholds are common for stylometric similarity detection tasks (e.g., Peng et al. [2002]). The average F-measure

Table IX. Performance Comparison of Ensemble SVM and SVM on Enron Dataset

Label	Technique	Feature Set Type	Feature Set
Writeprint	Writeprint	Individual	EF
K-L	K-L Transforms	Individual	EF
PCA/EF	PCA	Group	EF
Baseline	PCA	Group	BF

across all 10 folds was used to evaluate performance, where the F-measure for each fold was computed as follows.

$$\text{F-Measure} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}}$$

Similar to the identification experiment (see the results Table V), four combinations of feature sets, feature types, and techniques were used (shown in Table IX). A baseline was included which featured the use of PCA with the baseline feature set (BF). The baseline was intended to be compared against the use of PCA with the group-level extended feature set (PCA/EF) in order to assess the effectiveness of a more holistic feature set for online similarity detection (fourth row versus fifth row in Table IX). We also wanted to evaluate the effectiveness of individual-author-level feature sets by comparing Karhunen-Loeve transforms (which use EF-Individual) against the PCA/EF method, which uses a single group-level feature set (third row versus fourth row in Table IX). Finally, the Writeprints technique was included with extended feature set (EF-Individual) in order to evaluate the effectiveness of this technique in comparison with standard Karhunen-Loeve (K-L) transforms and PCA/EF (second row versus third and fourth rows, in Table IX). Since the Writeprints technique also utilizes the K-L transform, comparing Writeprint against K-L provided a good method for evaluating the effectiveness of the sliding window and pattern disruption algorithms.

6.3.2 *Hypotheses.*

—*H*1*a* (*Feature Sets*). The use of a more holistic feature set with a larger number of features and categories (EF) will outperform the baseline feature set (BF). Thus PCA/EF will outperform the baseline.

—*H*1*b* (*Feature-Set Types*). The use of individual-author-level feature subsets (EF-Individual) will outperform the use of a single author group-level feature set (EF-Group). Thus K-L transforms will outperform PCA/EF.

—*H*1*c* (*Techniques*). The Writeprints technique will outperform K-L transforms and PCA.

6.3.3 *Experimental Results.* Table X shows experimental results for all four combinations of features and techniques across the four datasets. The Writeprints technique had the best performance on all datasets, with F-measures over 85% for the Enron and eBay datasets when using 100 authors (200 identities). Furthermore, individual-author-level feature set techniques (Writeprints and K-L transforms) had higher accuracy on all datasets than

Table X. Experimental Results (% F-measure) for Similarity Detection Task

		No. Authors		
Test Bed	Techniques/Features	**25**	**50**	**100**
Enron Email	Writeprint	**93.62**	**94.29**	**85.56**
	K-L	75.29	68.23	65.44
	PCA/EF	70.32	56.33	50.82
	Baseline	64.32	48.49	34.33
eBay Comments	Writeprint	**100.00**	**97.96**	**94.59**
	K-L	92.25	84.10	80.93
	PCA/EF	81.19	77.32	72.25
	Baseline	75.65	70.02	60.19
Java Forum	Writeprint	**90.13**	**85.02**	**76.87**
	K-L	77.76	67.63	60.27
	PCA/EF	76.21	66.65	56.10
	Baseline	72.90	60.59	42.45
CyberWatch Chat	Writeprint	**68.43**	**62.88**	**49.91**
	K-L	50.72	42.39	30.77
	PCA/EF	40.0	33.3	19.8
	Baseline	39.43	28.62	20.10

author-group-level feature set methods (PCA/EF and baseline). This gap in performance appears to widen as the number of authors increases (e.g., looking at K-L versus PCA/EF), suggesting that the individual-author-level feature set (EF-Individual) is more scalable than the author-group-level feature set (EF-Group). The extended feature set (EF) had better overall performance than the benchmark feature set (BF), as reflected by the fact that all techniques using EF had higher accuracy than the baseline across all datasets.

6.3.4 *Hypotheses Results.* Table XI shows the p-values for the pair-wise t-tests conducted on the classification F-measures in order to measure the statistical significance of the results. Values in boldface indicate statistically significant outcomes that are in line with our hypotheses.

—*H*1*a* (*Feature Sets*). The extended feature set (EF) outperformed the baseline feature set (BF) across all datasets ($p < 0.01$) based on the better performance of PCA/EF as compared to baseline.

—*H*1*b* (*Feature-Set Types*). Individual-author-level feature subsets (EF-Individual) significantly outperformed the group-level feature set (EF-Group) on most datasets ($p < 0.01$). This outcome is based on the better performance of the K-L transforms technique as compared to PCA/EF.

—*H*1*c* (*Techniques*). The Writeprint technique significantly outperformed K-L transforms and PCA/EF on all datasets ($p < 0.01$).

6.3.5 *Results Discussion.* Overall performance for all techniques was best on the synchronous CMC datasets: Enron email and eBay comments. Once again, the performance was somewhat lower on the Java Forum and considerably lower on the CyberWatch chat datasets. For the Java Forum, we suspect that the feature sets EF and BF are not as effective at capturing programming style. For instance, many of the Writeprints pattern disruptors for the programming dataset were variable names and program methods. While such

Table XI. P-Values for Pair-Wise t-Tests on F-Measure

		No. Authors		
Test Bed	Techniques/Features	25	50	100
Enron Email	Writeprint vs. K-L	<0.001**	<0.001**	0.001**
	Writeprint vs. PCA/EF	<0.001**	<0.001**	<0.001**
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	K-L vs. PCA/EF	<0.001**	<0.001**	<0.001**
	K-L vs. Baseline	<0.001**	<0.001**	<0.001**
	PCA/EF vs. Baseline	<0.001**	<0.001**	<0.001**
eBay Comments	Writeprint vs. K-L	<0.001**	<0.001**	0.001**
	Writeprint vs. PCA/EF	<0.001**	<0.001**	<0.001**
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	K-L vs. PCA/EF	<0.001**	<0.001**	<0.001**
	K-L vs. Baseline	<0.001**	<0.001**	<0.001**
	PCA/EF vs. Baseline	<0.001**	<0.001**	<0.001**
Java Forum	Writeprint vs. K-L	<0.001**	<0.001**	0.001**
	Writeprint vs. PCA/EF	<0.001**	<0.001**	<0.001**
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	K-L vs. PCA/EF	0.094	0.087	<0.001**
	K-L vs. Baseline	<0.001**	<0.001**	<0.001**
	PCA/EF vs. Baseline	<0.001**	<0.001**	<0.001**
CyberWatch Chat	Writeprint vs. K-L	<0.001**	<0.001**	0.001**
	Writeprint vs. PCA/EF	<0.001**	<0.001**	<0.001**
	Writeprint vs. Baseline	<0.001**	<0.001**	<0.001**
	K-L vs. PCA/EF	<0.001**	<0.001**	<0.001**
	K-L vs. Baseline	<0.001**	<0.001**	<0.001**
	PCA/EF vs. Baseline	<0.001**	<0.001**	<0.001**

*P-values significant at alpha = 0.05.
**P-values significant at alpha = 0.01.

disruptors were assigned low values (based on synonymy), they still had a noticeable negative impact on performance. As we previously mentioned, program style analysis requires the use of features specifically geared towards code [Krsul and Spafford 1997]. In many cases, these features are not only tailored towards code, but rather for code in a specific programming language [Berry and Meekings 1985]. Future analysis of programming style should continue to incorporate more program-specific features, such as those used by Oman and Cook [1989] and Krsul and Spafford [1997].

In the case of the CyberWatch chat dataset, the amount of text for each author was insufficient to effectively discriminate authorship. More important than the number of words per author was the fact that we only had a single conversation for each author. It is unlikely that a single conversation would reveal a sufficient portion of an author's spectrum of stylistic variation for effective categorization. Further work is needed on stylometric analysis of chatroom data, including investigating chatroom-specific features and techniques on larger datasets.

7. CONCLUSIONS

In this study we applied stylometric analysis to online texts. Our research contributions are manyfold. We developed the K-L-transforms-based Writeprints technique, which can be used for identity-level identification and similarity detection. A novel pattern disruption mechanism was introduced to help detect authorship dissimilarity. We also incorporated a significantly more comprehensive feature set for online stylometric analysis and demonstrated the effectiveness of individual-author-level feature subsets. Our proposed feature set and technique were applied across multiple domains, including asynchronous CMC, synchronous CMC, and program code. The results compared favorably against existing benchmark methods and other individual-author-level techniques. Specifically, the Writeprints technique significantly outperformed other identification methods across domains such as email messages and chatroom postings. For similarity detection, Writeprints significantly outperformed comparison techniques across all datasets. The extended feature setutilized demonstrated the effectiveness of using richer stylistic representations for improved performance and scalability. Furthermore, the use of individual-author-level feature sets seems promising for application to cyberspace, where the number of authors can quickly become very large, making a single feature set less effective.

In the future we will work on further improving the scalability of the proposed approach to larger numbers of authors in a computationally efficient manner. We also plan to evaluate temporally dynamic individual-author-level feature sets that can gradually change over time as an author's writing style evolves. Another important direction is to assess the impact of intentional stylistic alteration on stylometric categorization performance.

REFERENCES

Abbasi, A. and Chen, H. 2005. Identification and comparison of extremist-group Web forum messages using authorship analysis. *IEEE Intel. Syst. 20*, 5, 67–75.

Abbasi, A. and Chen, H. 2006. Visualizing authorship for identification. In *Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics*, San Diego, CA.

Airoldi, E. and Malin, B. 2004. Data mining challenges for electronic safety: The case of fraudulent intent detection in e-mails. In *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining*.

Argamon, S., Saric, M., and Stein, S. S. 2003. Style mining of electronic messages for multiple authorship discrimination: First results In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Argamon, S., Koppel, M., and Avneri, G. 1998. Routing documents according to style. In *Proceedings of the 1st International Workshop on Innovative Information.*

Bayyen, R. H., Halteren, H. V., Neijt, A., and Tweedie, F. J. 2002. An experiment in authorship attribution. In *Proceedings of the 6th International Conference on Statistical Analysis of Textual Data*.

Bayyen, R. H., Halteren, H. V., and Tweedie, F. J. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Liter. Linguist. Comput. 2*, 110–120.

Berry, R. E. and Meekings, B. A. E. 1985. A style analysis of C programs. *Commun. ACM 28*, 1, 80–88.

Binongo, J. N. G. and Smith, M. W. A. 1999. The application of principal component analysis to stylometry. *Liter. Linguist. Compu. 14*, 4, 445–466.

Burrows, J. F. 1987. Word patterns and story shapes: The statistical analysis of narrative style. *Liter. Linguist. Comput. 2*, 61–67.

Chaski, C. E. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigation. *Int. J. Digit. Evidence 4*, 1, 1–13.

Chaski, C. E. 2001. Empirical evaluation of language-based author identification techniques. *Forensic Linguist. 8*, 1, 1–65.

CHERKAUER, K. J. 1996. Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks, In *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, P. Chan, ed., 15–21.

CORNEY, M., DE VEL, O., ANDERSON, A., AND MOHAY, G. 2002. Gender-Preferential text mining of email discourse. In 18*th Annual Computer Security Applications Conference*, Las Vegas, NV.

DASH, M. AND LIU, H. 1997. Feature selection for classification. *Intell. Data Anal. 1*, 131–156.

DE VEL, O., ANDERSON, A., CORNEY, M., AND MOHAY, G. 2001. Mining e-mail content for author identification forensics. *ACM SIGMOD Rec. 30*, 4, 55–64.

DIETTERICH, T. G. 2000. Ensemble methods in machine learning. In *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, 1–15.

DIEDERICH, J., KINDERMANN, J., LEOPOLD, E., AND PAASS, G. 2003. Authorship attribution with support vector machines. *Appl. Intell. 19*, 109–123.

DING, H. AND SAMADZAHEH, H. M. 2004. Extraction of Java program fingerprints for software authorship identification. *J. Syst. Softw. 72*, 49–57.

EFRON, M., MARCHIONINI, G., AND ZHIANG, J. 2004. Implications of the recursive representation problem for automatic concept identification in on-line government information. In *Proceedings of the ASIST SIG-CR Workshop*.

ERICKSON, T. AND KELLOGG, W. A. 2000. Social translucence: An approach to designing systems that support social processes. *ACM Trans. Comput. Hum. Interact. 7*, 1, 59–83.

FELLBAUM, C. 1998. *Wordnet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

FORMAN, G. 2003. An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research 3*, 1289–1305.

FORSYTH, R. S. AND HOLMES, D. I. 1996. Feature finding for text classification. *Litera. Linguist. Comput. 11*, 4, 163–174.

GARSON, G. D. 2006. *Public Information Technology and E-Governance: Managing the Virtual State.* Jones and Bartlet, Boston, MA.

GRAY, A., SALLIS, P., AND MACDONNEL, S. 1997. Software forensics: Extended authorship analysis techniques to computer programs. In *Proceedings of the 3rd Biannual Conference on the International Association of Forensic Linguists*.

GUYON, I., AND ELISSEEFF, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

HAYNE, C. S. AND RICE, E. R. 1997. Attribution accuracy when using anonymity in group support systems. *Int. J. Hum. Comput. Studies 47*, 429–452.

HAYNE, C. S., POLLARD, E. C., AND RICE, E. R. 2003. Identification of comment authorship in anonymous group support systems. *J. Manage. Inf. Syst. 20*, 1, 301–329.

HERRING, S. C. 2002. Computer-Mediated communication on the Internet. *Ann. Rev. Inf. Sci. Technol. 36*, 1, 109–168.

HOLMES, D. I. 1992. A stylometric analysis of Mormon scripture and related texts. *J. Royal Statis. Soci. 155*, 91–120.

JACKSON, D. 1993. Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecol. 74*, 8, 2204–2214.

JOSANG, A., ISMAIL, R., AND BOYD, C. 2007. A survey of trust and reputation systems for online service provision. *Decis. Support Syst. 43,* 2, 618–644.

JUOLA, P. AND BAAYEN, H. 2005. A controlled-corpus experiment in authorship identification by cross-entropy. *Liter. Linguist. Comput. 20,* 59–67.

KIRBY, M. AND SIROVICH, L. 1990. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell. 12*, 1, 103–108.

KJELL, B. WOODS, W. A., AND FRIEDER, O. 1994. Discrimination of authorship using visualization. *Inf. Process. Manage. 30*, 1, 141–150.

KOPPEL, M. AND SCHLER, J. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.

KOPPEL, M. AKIVA, N., AND DAGAN, I. 2006. Feature instability as a criterion for selecting potential style markers. *J. Amer. Soc. Inf. Sci. Technol. 57*, 11, 1519–1525.

KRSUL, I. AND SPAFFORD, H. E. 1997. Authorship analysis: Identifying the author of a program. *Comput. Secur. 16*, 3, 233–257.

Li, J., Zheng, R., and Chen, H. 2006. From fingerprint to writeprint. *Commun. ACM 49*, 4, 76–82.

Martindale, C. and McKenzie, D. 1995. On the utility of content analysis in author attribution: The federalist. *Comput. Humanit. 29*, 259–270.

McDonald, D., Chen, H., Hua, S., and Marshall, B. 2004. Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser. *Bioinf. 20*, 18, 3370–3378.

Merriam, T. V. N. and Matthews, R. A. J. 1994. Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Liter. Linguist. Comput. 9*, 1–6.

Moores, T. and Dhillon, G. 2000. Software piracy: A view from Hong Kong. *Commun. ACM 43*, 12, 88–93.

Morzy, M. 2005. New algorithms for mining the reputation of participants of online auctions. In *Proceedings of the 1st Workshop on Internet and Network Economics*, Hong Kong.

Mosteller, F. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* 2nd ed., Springer.

Oman, W. P. and Cook, R. C. 1989. Programming style authorship analysis. In *Proceedings of the 17th Annual ACM Computer Science Conference*, 320–326.

Pan, Y. 2006. ID identification in online communities. Working paper.

Peng, F., Schuurmans, D., Keselj, V., and Wang, S. 2003. Automated authorship attribution with character level language models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics.*

Platt, J. 1999. Fast training on SVMs using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf et al., eds. MIT Press, Cambridge, MA, 185–208.

Rudman, J. 1997. The state of authorship attribution studies: Some problems and solutions. *Comput. Humanit. 31*, 351–365.

Sack, W. 2000. Conversation Map: An interface for very large-scale conversations. *J. Manage. Inf. Syst. 17*, 3, 73–92.

Stamatatos, E. and Widme, R. G. 2002. Music performer recognition using an ensemble of simple classifiers. In *Proceedings of the 15th European Conference on Artificial Intelligence,* Lyon, France.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. 2000. Automatic text categorization in terms of genre and author. *Comput. Linguist 26*, 4, 471–495.

Sullivan, B. 2005. Seduced into scams: Online lovers often duped. *MSNBC,* July 28.

Tweedie, F. J., Singh, S., and Holmes, D. I. 1996. Neural network applications in stylometry: The Federalist papers. *Comput. Humanit. 30*, 1, 1–10.

Uenohara, M. and Kanade, T. 1997. Use of the Fourier and Karhunen-Loeve decomposition for fast pattern matching with a large set of features. *IEEE Trans. Pattern Analy. Mach. Intell. 19*, 8, 891–897.

Wang, H., Fan, W., and Yu, S. P. 2003. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Watanbe, S. 1985. *Pattern Recognition: Human and Mechanical.* John Wiley, New York.

Webb, A. 2002. *Statistical Pattern Recognition.* John Wiley, New York.

Whitelaw, C. and Argamon, S. 2004. Systemic functional features in stylistic text classification. In *Proceedings of the AAAI Symposium on Style and Meaning in Language, Art, Music and Design*, Washington, DC.

Yang, Y. and Pederson, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning,* 412–420.

Yule, G. U. 1944. *The Statistical Study of Literary Vocabulary.* Cambridge University Press.

Yule, G. U. 1938. On sentence length as a statistical characteristic on style prose. *Biometrika 30*.

Zheng, R., Li, J., Huang, Z., and Chen, H. 2006. A framework for authorship analysis of online messages: Writing-style features and techniques. *J. Amer. Soc. Inf. Sci. Technol. 57*, 3, 378–393.

Received November 2006; revised May 2007; accepted May 2007