12-31-2007

A Framework for Stylometric Similarity Detection in Online Settings

Ahmed Abbasi
University of Arizona

Hsinchun Chen
University of Arizona

A Framework for Stylometric Similarity Detection in Online Settings

Ahmed Abbasi and Hsinchun Chen
Department of Management Information Systems, Eller College of Business
The University of Arizona, Tucson, Arizona 85721, USA
{aabbasi@email.arizona.edu, hchen@eller.arizona.edu}

Abstract

Online marketplaces and communication media such as email, web sites, forums, and chat rooms have been ubiquitously integrated into our everyday lives. Unfortunately, the anonymous nature of these channels makes them an ideal avenue for online fraud, hackers, and cybercrime. Anonymity and the sheer volume of online content make cyber identity tracing an essential yet strenuous endeavor for Internet users and human analysts. In order to address these challenges, we propose a framework for online stylometric analysis to assist in distinguishing authorship in online communities based on writing style. Our framework includes the use of a scalable identity-level similarity detection technique coupled with an extensive stylistic feature set and an identity database. The framework is intended to support stylometric authentication for Internet users as well as provide support for forensic investigations. The proposed technique and extended feature set were evaluated on a test bed encompassing thousands of feedback comments posted by 100 electronic market traders. The method outperformed benchmark stylometric techniques with an accuracy of approximately 95% when differentiating between 200 trader identities. The results indicate that the proposed stylometric analysis approach may help mitigate the effects of online anonymity abuse.

Introduction

One of the problems associated with online anonymity is that it facilitates opportunistic behavior, thereby hindering social accountability, much to the detriment of the overall online community. 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). Furthermore, Internet-based communication is swarming with fraudulent scams. One well-known fraudulent scheme is the 4-1-9 scam (Airoldi and Malin, 2004) which has been around for over a decade, generating billions of dollars in fraudulent revenues (Sullivan, 2005). Electronic market places are another area susceptible to deception stemming from easy identity changes and reputation rank inflation (Dellarocas, 2003; Josang et al., 2007). In this scheme online sellers create fake sales transactions to themselves in order to improve reputation rank (Josang et al., 2007). While such behavior may simply serve as a mechanism to garner accreditation, it is often employed in order to defraud unsuspecting fellow traders.

The aforementioned forms of Internet misuse all involve text-based modes of computer mediated communication. Hence, those responsible often leave behind textual traces of their identity (Li et al., 2006). Keselj et al. (2002) refer to an author's unique stylistic tendencies as an "author profile." 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), program code (Gray et el., 1997), and group support system comments (Hayne and Rice, 1997; Hayne et al., 2003). 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 apriori (which is more practical for cyberspace).

There is also a need to apply stylometric techniques to cyber content in such a manner that improves online accountability while protecting people's privacy. Tools providing greater informational transparency in cyberspace are necessary to counter anonymity abuses and garner increased accountability (Erickson and Kellogg, 2000; Sack, 2000). However, such systems must also protect the privacy of individuals. Erickson and Kellogg (2000) suggested the development of socially translucent systems that can monitor behavior in online communities without infringing upon people's privacy. Such systems are intended to proactively deter deviant behavior in cyberspace via authentication mechanisms. Based on the aforementioned limitations of prior studies we are motivated to address the following research issues:
(1) We intend to develop a framework for stylometric analysis that supports social translucence in cyberspace. Our proposed framework will address some of the current limitations of online stylometric analysis. (2) We will incorporate a larger, more holistic feature set than those used in previous research. (3) We will also develop the Writeprint technique, which is intended to improve stylometric analysis scalability across authors and domains for the similarity detection task.

Research Background

In this section we present a summary of previous stylometry research. Stylometric analysis techniques have been used for analyzing and attributing authorship of literary texts for numerous years (e.g., Mosteller and Wallace, 1964). Four important characteristics of stylometry are the analysis tasks, writing style features used, techniques incorporated to analyze these features, and stylometric parameters (Zheng et al., 2006). These characteristics are discussed below.

Tasks

Two major stylometric analysis tasks are identification and similarity detection (Gray et al., 1997; De Vel et al., 2001). Identification entails comparing anonymous texts against those belonging to identified entities, where anonymous text is known to be written by one of those entities. Since all possible classes are known apriori, the identification tasks can use supervised or unsupervised techniques. However, this "known class" assumption is not practical (Juola and Baayen, 2005), especially for online settings. In cyberspace, author classes are rarely known in advance, and hence require the use of unsupervised clustering based approaches. Such a similarity detection task requires the comparison of anonymous texts against other anonymous texts in order to assess the degree of similarity. For instance, in online forums, where there are numerous anonymous identities (i.e., screen names, handles, email addresses) one can only use unsupervised stylometric analysis techniques since no class definitions are available.

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 (Argamon et al., 2003), vocabulary richness (De Vel et al., 2001) 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). 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, for example, the use of braces and comments, are utilized (Krsul and Spafford, 1997).

Content-specific features are 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, e.g., use of the word "centre" 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, 1998). 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. For instance, the use of feature sets containing lexical, syntactic, structural, and syntactic features has been shown to be more effective for online identification than feature sets containing only a subset of these feature groups (Abbasi and Chen, 2005; Zheng et al., 2006).

Techniques

Several techniques have been used for stylometric identification. These can broadly be classified as supervised and unsupervised methods. However, only unsupervised techniques are suitable for online settings, since class definitions are unknown apriori. We discuss previous unsupervised methods, since they can support the online similarity detection task. These techniques include principal component analysis (PCA), N-Gram Models, Markov Models, and Cross Entropy.

PCA is a popular stylometric identification technique that has been used in numerous previous studies (Burrows, 1987; Kjell et al, 1994; Baayen et al., 1996; Abbasi and Chen, 2006). PCA's ability to capture essential variance across large amounts of features in a reduced dimensionality makes it attractive for text analysis problems, which typically involve large feature sets. The essence of PCA can be described as follows: given a feature matrix with each column representing a feature and instance vector rows for the various authors' texts, project the matrix into a lower dimensional space by plotting principal component scores (which are the product of the component weights and instance feature vectors). The similarity between authors can be compared based on visual proximity of patterns (Kjell et al., 1994) or computation of average distance (Abbasi and Chen, 2006).

Proposed by Keselj et al. (2003) and Peng et al. (2003), N-Gram models require the construction of a profile for each author, where a profile is the set of the *n* most frequently used character n-grams. Keselj et al. (2003) used in between 20-5,000 as the value for *n*, with the best accuracy attained using 5,000 n-grams. They attained optimal results using 4-8 character n-grams. Using this approach, they computed the dissimilarity between two authors as the normalized difference in usage frequency for all unique features occurring in either profile. Keselj et al. (2003) and Peng et al. (2003) were able to attain good performance using this approach on test beds consisting of up to 8 authors.

Markov Models (Khmelev, 2000; Khemelev and Tweedie, 2001) entail the creation of a Markov model for each author, using bi-grams of letters and the space character. Khmelev (2000) removed all other characters and ignored words beginning with a capitalized letters, resulting in a fixed (27 x 27 = 729) feature space for each author. Using this approach, the similarity between two authors can be computed by taking the difference between their letter/space transition probabilities. The technique has performed well on larger test beds of 45 and 82 authors (Khmelev, 2000; Khmelev and Tweedie, 2001) however these data sets consisted of literary texts which tend to be longer and more stylistically consistent due to contextual independence.

Cross Entropy (Juola, 1997; 2003; Juola and Baayen, 2005) is based on the concept of match length where:

$$\text{The match length } L_n(x) \text{ of a sequence } x_1, x_2, \ldots x_k \text{ is the length of the longest prefix of}$$
$$\text{the sequence } x_{n+1}, x_{n+2}, \ldots x_k \text{ that matches a contiguous substring of } x_1, x_2, \ldots x_n$$

For cross entropy, simply compute the average match length for author B's text compared against author A's database and vice versa. Texts written by the same author should result in higher match lengths. Juola (1997) used *n*=2,000 characters for each author's database size. The cross entropy method has performed well in prior studies, outperforming PCA on a test bed consisting of 8 students' essays (Juola and Baayen, 2005).

Most techniques, such as N-Gram and Markov models were designed to be used with character n-grams. Word based features are too sparse to be used accurately with these techniques (Peng et al., 2003). It is unclear if such methods can be effectively applied to online settings, where techniques capable of handling larger feature sets are typically required (Abbasi and Chen, 2005; Zheng et al., 2006). It is therefore especially important to assess the efficacy of these techniques for online analysis in order to gauge their applicability for stylometric similarity detection of reputation system feedback comments.

Stylometric Analysis Parameters

An important stylometric analysis parameter for online authentication is scalability. Scalability refers to the impact of the number of author classes on classification performance. Typically, there has been a noticeable drop in performance for prior online message level identification research as the number of authors increased. Zheng at al. (2006) noted a 14% drop in accuracy when increasing the number of author classes from 5 to 20. Argamon et al. (2003) observed as much as a 23% drop in accuracy when increasing the number of authors from 5 to 20. Given the large number of traders in online markets, it is important to assess the impact of the number of traders and identities per trader on stylometric identification performance in electronic market reputation systems.

Proposed Research

Based on the previously identified research gaps, we aim to answer the following questions:

- Which authorship analysis techniques can be successfully used for the online similarity detection task?
- What impact will the use of a more holistic feature set have on online classification performance?
- How scalable are these features and techniques with respect to the number of authors?

In order to address these questions, we propose the creation of a stylometric analysis framework that can perform ID level similarity detection in online settings. Our approach (shown in Figure 1) utilizes a more holistic feature set consisting of a larger number of features across several categories to improve our representational richness of authorial style. The framework has four major components: an identity database, an extensive set of stylistic features, and the Writeprint technique which can be used for identity-level similarity detection. The framework is intended to provide Internet users with stylometric authentication to support social translucence and also aid cybercrime investigators for forensic applications without infringing on individuals' privacies. Further details about the framework are provided in the ensuing sections.

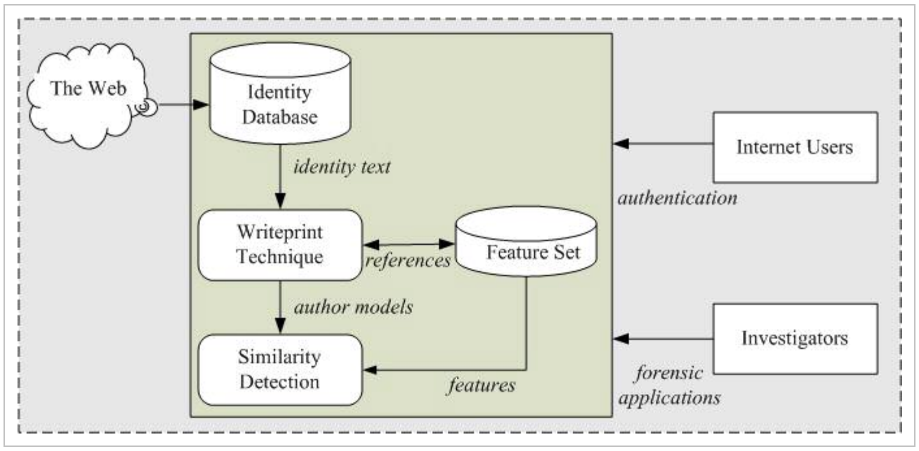

Figure 1: A Framework for Online Stylometric Analysis

Identity Database

We propose the creation of an identity database with two types of identities and their associated text: (1) identities and text for individuals known to be involved in fraudulent activities (2) publicly available computer mediated communication archives for large numbers of identities.

Many public online databases/archives have been developed containing conversation logs for cyber scammers and criminals. For instance, EBay provides an archive of fraudulent buyers and sellers with their associated transactions and feedback comments. Several web sites provide galleries containing thousands of fraudulent emails written by scammers (Airoldi and Malin, 2004). CyberWatch (www.cyberwatch.com) and several others provide instant messaging chat logs for cyber criminals. There are also databases for fraudulent web sites. For instance, escrow-fraud.com has a database containing over 3,000 fraudulent web sites claiming to provide escrow services, while the Artists Against 4-1-9 database (URL) contains over 13,000 fraudulent financial web sites. These fake sites are constantly rehashed with different URLs by the same set of individuals, often with similar writing.

The objective of the aforementioned web sources is to create greater user awareness to hinder future cyber fraud. However, browsing such archives (which contain text from hundreds or thousands of identities) manually can quickly result in information overload. The use of stylometric methods coupled with such information sources could provide invaluable authentication capabilities. There are also many useful public corpora and archives that can provide invaluable large scale research test beds. Some relevant data sets include the Enron email corpus and programming web forums containing thousands of code snippets (e.g., the Sun Java Forum). Application of stylometric analysis techniques to the first information source can be useful for providing online user authentication (i.e., using authorship analysis to confirm that the user is not a known scammer). In order to protect the privacy of individuals, we do not believe that common Internet users need access to stylometric analysis of non-fraudulent identities as such analysis is not necessary for authentication. These identities may however be useful for forensic investigations, analogous to a biometric database.

Collection of such materials from public sources off the web requires the use of spidering programs that can collect such information periodically and wrappers that can parse out the important meta-data (e.g., author, text, date, etc.) based on the various web sites' formats. We have already identified and collected several such information sources into our identity database. Under the proposed system, we intend to continue expanding our database to include additional identities from various CMC modes. The identity database provides text samples that may serve as an important part of the online stylometric analysis framework for stylometric similarity detection tasks as well as for evaluation of the feature set and Writeprint technique.

Feature Set

The proposed feature set (shown in Table 1) is a mixture of static and dynamic features. The dynamic features include several n-gram feature categories and a list of 5,513 common word misspelling taken from various websites including Wikipedia. N-gram categories utilized include character, word, POS tag, and digit level n-grams. These categories require indexing with the number of initially indexed features varying depending on the data set. 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., Keselj et al., 2002; Peng et al., 2002; Koppel and Schler, 2003).

Table 1: Proposed Extended Feature Set

Group	Category	Quantity	Description
Lexical	Word-Level	5	total words, % char. per word
	Character-Level	5	total char., % char. per message
	Character N-Grams	< 18,278	count of letters, bigrams, and trigrams (e.g., a, b, aa, ab, aab)
	Digits N- Grams	< 1,110	frequency of one to three digit numbers (e.g., 1, 2, 11, 101)
	Word Length Dist.	20	frequency distribution of 1-20 letter words
	Vocabulary Richness	8	richness (e.g., hapax legomena, Yule's K)
	Special Characters	21	occurrences of special char. (e.g., @#$%^and*+)
Syntactic	Function Words	300	frequency of function words (e.g., of, for, to)
	Punctuation	8	occurrence of punctuation marks (e.g., !;:,.?)
	POS Tag N-Grams	varies	counts of part-of-speech n-grams (e.g., "NP," "NP VB")
Structural	Message-Level	6	e.g., has greeting, has url, requoted content
	Paragraph-Level	8	e.g., no. of paragraphs, sentences per paragraph
	Technical Structure	50	e.g., file extensions, fonts, use of images
Content	Words N-Grams	varies	words, bigrams and trigrams (e.g., "senior," "editor in chief")
Idiosyncratic	Misspelled Words	< 5,513	common misspellings (e.g., "beleive", "thougth")

Writeprint Technique

The Writeprint technique has two major components: creation and comparison. The creation steps are concerned with the construction of Writeprint patterns reflective of an identities' writing style variation, based on the occurrence of common identity features as well as lack of occurrence of style markers prevalent in other identities' text. The comparison steps describe how created Writeprints for various trader identities are compared against one another to assess the degree of stylistic similarity. The two components are described below.

Writeprint Creation

The Writeprint creation component can be further decomposed into two steps. In the first step, Karhunen-Loeve transforms (KL-Transforms) are applied with a sliding window in order to capture stylistic variation with a finer level of granularity. The second step, pattern disruption, uses zero usage features as red flags intended to decrease the level of stylistic similarity between identities when one identity contains important features not occurring in the other. The two major steps, which are repeated for each identity, are described below:

Step 1: Sliding Window and KL-Transforms

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) was used to select the number of eigenvectors in the basis. The formulation for step 1 is presented below:

a) Let $\Omega = \{1,2,...,f\}$ denote the set of *f* features with frequency greater than 0 and $\Phi = \{1,2,...,w\}$ 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 $i \in \Phi$.

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

b) Extract the set of eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ for the covariance matrix Σ of the feature matrix *X* by finding the 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 $a_m = (a_{m1}, a_{m2}, ..., a_{mf})$ by solving the following system, resulting in a set of *n* eigenvectors $\{a_1, a_2, ..., a_n\}$:

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

c) Compute an *n*-dimensional representation for each window *i* by extracting principal component scores ε_{ik} for each dimension $k \le n$:

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

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 by the identity of interest will currently be irrelevant to the process since they have no variance. Nevertheless these features are still important when comparing a trader identity to other anonymous trader identities. The trader'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). When comparing two trader identities A and B, we would like A's zero frequency features to act as pattern disruptors, where the presence of these features in identity B's feedback comments decreases the similarity for the particular A – B comparison (and vice versa for the B – A comparison).

The magnitude of a disruptor signifies the extent of the disruption for a particular feature. Larger values of for the disruptor will cause pattern points representing text windows containing the disruptor feature to be shifted further away. However, not all features are equally important discriminators. 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 noun phrases to be more stable than function words and argued that function words are better stylistic discriminators than noun phrases since use of function words involves making choices between a set of synonyms. Based on this intuition, we used the disruptor feature's information gain and synonymy information to assign them a weight (disruptor coefficient), which was appended to the identity's basis matrix (set of eigenvectors).

a) Let $\Psi = \{f+1, f+2, ..., f+g\}$ denote the set of *g* features with zero frequency. For each feature $p \in \Psi$ compute the disruptor coefficient d_p:

$$d_p = IG(c,p)K(\mathrm{syn}_{total} + 1)(\mathrm{syn}_{used} + 1)$$

where $IG(c,p)$ is the information gain for feature *p* across the set of classes *c*, syn_{total} and syn_{used} are the total synonyms and the number used by the author, respectively, for the disruptor feature, and *K* is a disruptor constant.

b) For each feature $p \in \Psi$ append the value d_{kp} to each eigenvector a_k where $k \le n$.

Writeprint Comparisons

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. We would need to construct a pattern for identity B using B's text with A's feature set and basis matrix (Pattern B) to be compared against identity A's Writeprint (and vice versa). The overall similarity between Identity A and B is the sum of the average distance between Writeprint A and Pattern B and Writeprint B and Pattern A.

As previously mentioned, the pattern disruptors are intended to assess the degree of stylistic dissimilarity based on important features only found in one of the two identities' feedback comments. Disruptors shift pattern points further away from the Writeprint they're being compared against, thereby increasing the average distance between patterns (and reducing the similarity score). The direction of a pattern 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 (towards the right in this case) as opposed to towards it.

Evaluation

In order to evaluate the effectiveness of the proposed system, which includes the Writeprint technique and extended feature set, experiments were conducted that compared the system against previous unsupervised stylometric identification techniques described, including PCA, N-Gram Models, Markov Models, and Cross Entropy. The test bed, experimental design, and parameter settings for the Writeprint and comparison techniques are described below.

Test Bed

The test bed consisted of buyer/seller feedback comments extracted from eBay's online reputation system. We extracted 100 eBay members selling electronic goods. For each trader, 3,000 feedback comments posted by that author were included. Table 2 provides summary statistics of the test bed while example feedback comments are listed below:

- "Another quick and easy transaction, thanks for your biz!"
- "Excellent e-bayer!! fast payment, great to deal with, many thanks!!!"
- "PLEASURE doing business with you and thanks for making this business a PLEASURE!"

Table 2: eBay Test Bed Statistics

# Authors (i.e., traders)	Words (per author)	Comments (per author)	Ave. Comment Length (words)	Time Duration
100	23,423	3,000	7.81	02/2003 – 04/2006

Experimental Setup

All comparison techniques were run using the best parameter settings determined by tuning these parameters on the actual test bed data. This was done in order to allow the best possible comparison against the proposed Writeprint technique. Most of the parameter values were consistent with prior research. PCA was run using the extended feature set. We extracted feature vectors for 1,500 character text blocks, consistent with prior research (Abbasi and Chen, 2006). The Kaiser-Guttman stopping rule was used (i.e., extract all eigenvectors with an eigenvalue greater than 1). For the N-gram Models, we used character level n-grams, with profile sizes of 5,000 n-grams per identity. For each identity we used 4-8 character n-grams since this configuration garnered the best results, also consistent with Peng et al. (2003) and Keselj et al. (2003). Markov Models were built using letters and space bigrams. We removed all other characters and ignored words beginning with capital letters, as done by Khemelev (2001) and Khemelev and Tweedie (2001). For Cross Entropy we used a database size of 5,000 characters for each identity as this size provided the best performance.

For the experiments, we created two identities for each of the 100 eBay traders by splitting the traders' feedback comment text into two parts. The objective of the experiments was to see how well the proposed Writeprint method and comparison techniques could match up the different trader identities based on their comment texts. Each trader's text was split into 12 parts. If two identities were to be created for a single trader, 6 parts were randomly assigned to each identity. For example, parts 1, 5, 7, 8, 9, 11 (identity 1), parts 2, 3, 4, 6, 10, 12 (identity 2). In order to test the statistical significance of the techniques' performance, bootstrapping was performed 30 times for each technique, where each iteration the 12 trader text parts were randomly split into the desired number of identities. A trial and error method was used to find the optimal similarity threshold for matching for each technique. The same threshold was used throughout the experiments for the Writeprint method. A dynamic threshold yielding optimal results for the particular experimental settings was used for each comparison technique. This was done in order to compensate for differences in performance attributable to thresholds instead of techniques. All identity-identity scores above a techniques' threshold were considered a match. The F-Measure was used to evaluate performance.

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

Using these experimental settings, two sets of experiments were conducted. The first assessed the scalability of the proposed stylometric similarity detection system and comparison approaches in terms of number of traders and number of identities' comments. The second attempted to evaluate the effectiveness of these stylometric methods against intentional stylistic alteration and forging/copycatting. Details about the two experiments are presented in the ensuing sections.

Experimental Results

We conducted experiments to analyze the scalability across traders. Each trader's text was split into two anonymous identities. We used 25, 50, and 100 traders (i.e., 50, 100, and 200 identities). Figure 2 shows the F-measure percentages for 25, 50, and 100 traders (with 2 identities per trader), intended to assess the scalability across traders. Overall all the techniques except PCA performed well. As expected, doubling the number of authors and identities decreased performance, however the decrease was gradual. Writeprint had the best performance for all three trader/identity levels. The technique only had approximately a 3% decrease when going from 100 to 200 identities. In contrast the performance of N-Gram Models and Cross Entropy fell 6%-7%.

Techniques	**# Traders / #Identities**		
	25/50	**50/100**	**100/200**
Writeprint	**100.00**	**97.88**	**94.59**
PCA	81.19	77.49	72.25
N-Gram Models	97.96	97.60	89.39
Markov Models	98.04	90.91	87.86
Cross Entropy	**100.00**	96.15	90.34

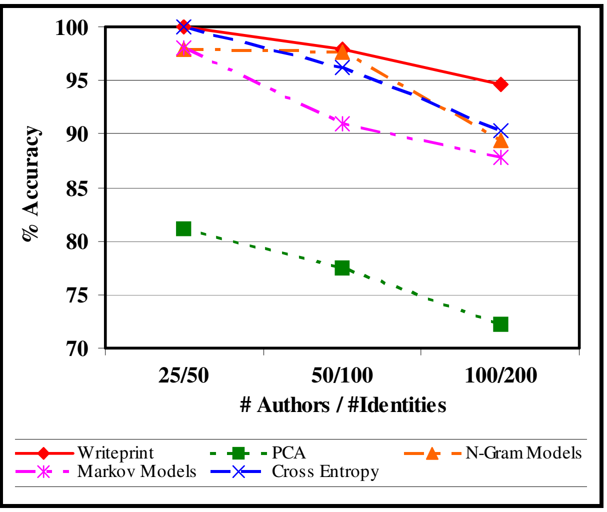

Figure 2: Experimental Results (scalability across traders)

Table 3 shows the p-values for the pair wise t-tests on F-measure. Writeprint significantly outperformed all comparison techniques. The N-gram and Markov models and Cross Entropy techniques significantly outperformed PCA for all three settings. Furthermore, Cross Entropy significantly outperformed N-Gram and Markov Models when using 200 identities.

Table 3: P-Values for Pair Wise t-tests on F-measure (n=30)

Techniques	# Traders / #Identities		
	25/50	50/100	100/200
Writeprint vs. PCA	<0.001*	<0.001*	<0.001*
Writeprint vs. N-Gram	<0.001*	0.109	<0.001*
Writeprint vs. Markov	<0.001*	<0.001*	<0.001*
Writeprint vs. Cross	0.852	<0.001*	<0.001*
PCA vs. N-Gram	<0.001*	<0.001*	<0.001*
PCA vs. Markov	<0.001*	<0.001*	<0.001*
PCA vs. Cross	<0.001*	<0.001*	<0.001*
N-Gram vs. Cross	<0.001*	<0.001*	0.138
N-Gram vs. Markov	0.464	<0.001*	<0.001*
Markov vs. Cross	<0.001*	<0.001*	<0.001*

* P-values significant at alpha = 0.01

Writeprint had the best performance for all trader/identity levels. The performance gap widened as the number of traders and identities increased, suggesting that the extended feature set and pattern disruption mechanism incorporated by Writeprint allowed improved scalability. The enhanced representational richness of Writeprint allowed it to outperform the n-gram based techniques (N-Gram and Markov Models) while the pattern disruption component enabled improved performance over PCA. With respect to the comparison techniques, Cross Entropy had the best performance, outperforming PCA, N-Gram and Markov Models (significantly outperforming them in many instances).

Conclusions and Future Directions

In this research we proposed a framework for online stylometric analysis. Our framework includes the use of an identity database, the novel Writeprint technique, and an extensive feature set suitable for stylometric analysis across various online text based media. We evaluated our approach on a test bed encompassing feedback comments from 200 eBay identities in comparison with several benchmark techniques. The combination of the Writeprint method and our extended feature set significantly outperformed the comparison methods, achieving over 94% accuracy when differentiating between 200 online traders. We believe the results are quite promising, suggesting that the proposed framework may help decrease online anonymity abuse on the Internet.

We have identified several future directions. We intend to evaluate the framework on other CMC modes, including email, forum postings, chat room logs, and web sites. We also plan to conduct experiments using even larger numbers of authors and identities. We also mean to explore the impact of contextual factors on writing style, and develop models that take into account factors such as the communication genre, recipient, emotion, topic, and period in time.

References

Abbasi, A., and Chen, H. “Identification and Comparison of Extremist-Group Web Forum Messages using Authorship Analysis,” IEEE Intelligent Systems (20:5), 2005, pp. 67-75.

Abbasi, A. and Chen, H. "Visualizing Authorship for Identification", In the 4th IEEE Symposium on Intelligence and Security Informatics (ISI 2006), San Diego, CA, 2006.

Airoldi, E., and Malin, B. “Data Mining Challenges for Electronic Safety: The Case of Fraudulent Intent Detection in E-Mails,” Workshop on Privacy and Security Aspects of Data Mining, 2004.

Argamon, S., Saric, M., and Stein, S. “Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results,” In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

Baayen, R. H., Halteren, H. V., and Tweedie, F. J. “Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution,” Literary and Linguistic Computing, (2), pp. 110-120, 1996.

Burrows, J. F. “Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style,” Literary and Linguistic Computing, (2), 1987, pp. 61-67.

Chaski, C. E. “Empirical Evaluation of Language-Based Author Identification Techniques,” Forensic Linguistics, (8:1), 2001, pp. 1-65.

Dellarocas, C. “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms,” Management Science, (49:10), 2003, pp. 1407-1424.

De Vel, O., Anderson, A., Corney, M., and Mohay, G. “Mining E-mail Content for Author Identification Forensics,” SIGMOD Record, (30:4), pp. 55-64, 2001.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. “Authorship Attribution with Support Vector Machines,” Applied Intelligence (19), 2003, pp. 109-123.

Erickson, T. and Kellogg, W. A. “Social Translucence: An Approach to Designing Systems that Support Social Processes,” ACM Transactions on Computer-Human Interaction (7:1), 2000 pp. 59-83.

Gray, A., Sallis, P., and MacDonell, S. “Software Forensics: Extending Authorship Analysis Techniques to Computer Programs,” Paper Presented at the Conference of the International Association of Forensic Linguists, 1997.

Hayne, C, S., Pollard, E, C., and Rice, E, R. “Identification of Comment Authorship in Anonymous Group Support Systems,” Journal of Management Information Systems (20:1), 2003, pp. 301-329.

Hayne, C. S., and Rice, E. R. “Attribution Accuracy when using Anonymity in Group Support Systems,” International Journal of Human-Computer Studies (47), 1997, pp. 429-452.

Jackson, D. “Stopping Rules in Principal Component Analysis: A Comparison of Heuristical and Statistical Approaches,” Ecology, (74:8), pp. 2204-2214, 1993.

Josang, A., Ismail, R., and Boyd, C. “A Survey of Trust and Reputation Systems for Online Service Provision,” Decision Support Systems, 2007.

Juola, P. “What can we do with Small Corpora? Document Categorization via Cross-Entropy,” In Proceedings of the Interdisciplinary Workshop on Similarity and Categorization, 1997.

Juola, P. “The Time course of Language Change,” Computers and the Humanities, (37), 2003, pp. 77-96.

Juola, P. and Baayen, H. “A Controlled-Corpus Experiment in Authorship Identification by Cross-Entropy,” Literary and Linguistic Computing, 2005.

Keselj, V., Peng, F., Cercone, N., and Thomas, C. “N-Gram Based Author Profiles for Authorship Attribution,” In Proceeedings of the Pacific Association for Computational Linguistics, 2003.

Khmelev, D. V. “Disputed Authorship Resolution using Relative Entropy for Markov Chains of Letters in Human Language Texts,” Journal of Quantitative Linguistics, (7), 2000, pp. 115-126.

Khmelev, D. V. and Tweedie, F. J. “Using Markov Chains for Identification of Writers,” Literary and Linguistic Computing, (16:3), 2001, pp. 299-307.

Kjell, B., Woods, W.A., and Frieder, O. “Discrimination of Authorship Using Visualization,” Information Processing and Management, (30:1), pp. 141-150, 1994.

Koppel, M. and Schler, J. “Exploiting Stylistic Idiosyncrasies for Authorship Attribution,” In Proceedings of IJCAI’03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, 2003.

Koppel, M., Akiva, N., and Dagan, Ido. “Feature Instability as a Criterion for Selecting Potential Style Markers,” Journal of the American Society for Information Science and Technology (57:11), 2006, pp. 1519-1525.

Krsul, Ivan., and Spafford, H, E. “Authorship Analysis: Identifying the Author of a Program,” Computers and Security (16:3), 1997, pp. 233-257.

Li, J., Zheng, R. and Chen, H. “From Fingerprint to Writeprint,” Communications of the ACM, (49:4), 2006, pp. 76-82.

Martindale, C., and McKenzie, D. “On the Utility of Content Analysis in Author Attribution: The Federalist,” Computers and the Humanities, (29), pp. 259-270, 1995.

Moores, T., and Dhillon, G. “Software Piracy: A View from Hong Kong,” Communications of the ACM, (43:12), pp. 88-93.

Mosteller, F., and Wallace, D. L. Applied Bayesian and Classical Inference: The Case of the Federalist Papers (2 ed.): Springer-Verlag, 1964.

Oman, W, P., and Cook, R, C. “Programming style Authorship Analysis,” In Proceedings of the 17th Annual ACM Computer Science Conference 1989, pp.320-326.

Peng, F., Schuurmans, D., Keselj, V., and Wang, S. “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.

Rudman, J. “The State of Authorship Attribution Studies: Some Problems and Solutions,” Computers and the Humanities (31), 1998, pp. 351-365.

Sack, W. “Conversation Map: An Interface for Very Large-Scale Conversations,” Journal of Management Information Systems (17:3), 2000, pp. 73-92.

Zheng, R., Qin, Y., Huang, Z., and Chen, H. “A Framework for Authorship Analysis of Online Messages: Writing-style Features and Techniques,” Journal of the American Society for Information Science and Technology (57:3), 2006, pp. 378-393.