Mining Writeprints from Anonymous E-mails for Forensic Investigation

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

Computer Security Laboratory
Faculty of Engineering and Computer Science
Concordia University, Montreal, Quebec, Canada.

Abstract

Many criminals exploit the convenience of anonymity in the cyber world to conduct illegal activities. E-mail is the most commonly used medium for such activities. Extracting knowledge and information from e-mail text has become an important step for cybercrime investigation and evidence collection. Yet, it is one of the most challenging and time-consuming tasks due to special characteristics of e-mail dataset. In this paper, we focus on the problem of mining the writing styles from a collection of e-mails written by multiple anonymous authors. The general idea is to first cluster the anonymous e-mails by the stylometric features and then extract the writeprint, i.e., the unique writing style, from each cluster. We emphasize that the presented problem together with our proposed solution is different from the traditional problem of authorship identification, which assumes training data is available for building a classifier. Our proposed method is particularly useful in the initial stage of investigation, in which the investigator usually have very little information of the case and the true authors of suspicious e-mails collection. Experiments on a real-life dataset suggest that clustering by writing style is a promising approach for grouping e-mails written by the same author.

Key words:
e-mail, writing styles, writeprint, forensic investigation, clustering, classification, stylometric features, authorship analysis

* Corresponding author and his contact address: C/O CIISE (EV7.628) Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec Canada H3G 1M8. Phone: (514) 576-6460, Fax: (514) 848-3171

Email addresses: iqbal_f@ciise.concordia.ca (Farkhund Iqbal), h_binsal@ciise.concordia.ca (Hamad Binsalleeh), fung@ciise.concordia.ca (Benjamin C. M. Fung), debbabi@ciise.concordia.ca (Mourad Debbabi).

1 Introduction

The cyber world provides a convenient platform for criminals to anonymously conduct their illegal activities, such as spamming and phishing. E-mail is the most commonly used communication medium that results in financial as well as moral loss to the victims of cybercrimes. In spamming, for instance, a culprit may attempt to hide his true identity. Likewise, in phishing, an intriguer may impersonate a banker to trick bank clients to disclose their personal sensitive information. Terrorist groups and criminal gangs also use e-mail as a safe channel for their secret communication.

Authorship analysis techniques, for identifying the true author of disputed anonymous online messages to prosecute cybercriminals in the court of law, are focused in recent cyber forensic investigation cases. These techniques are used to build a classification model based on the stylometric features, extracted from the example writings of potential suspects, and then use the model to identify the true author of anonymous documents in question [2][16][9]. Most authorship studies assume the true author of the disputed anonymous message must be among the given potential suspects. Another assumption is the availability of training data that is enough to build a classification model. Similarly, authorship characterization techniques are applied to collect cultural and demographic characteristics such as gender, age, and education background, of the author of an anonymous document. These techniques, however, need sufficiently large training data of sample population to classify the author to one of the categories of gender, age, etc.

In our study we focus the worst case scenario where neither the candidate suspects' list nor training examples are available to the investigator. For instance, during the initial stage of an investigation, a crime investigator may not have any clue about the potential suspects of the given disputed e-mails. Given a collection of suspicious anonymous e-mails E, presumably written by a group of unknown suspects say $\{S_1, \ldots, S_k\}$ with no example writings. A forensic investigator, however, may or may not know the actual number of authors in E. Both scenarios will be addressed in this paper.

In this situation, the investigator may apply our method to first identify the major groups of e-mails based on the writing style features. Depending on the purity of clusters, it is assumed that each e-mail collection is written by one suspect. The extracted writeprint from each e-mails group, by using our method, can be used for authorship attribution and authorship characterization at a later stage. Intuitively, the extracted writeprint (c.f. fingerprint) represents one author's writing style that is specific enough to distinguish his/her written e-mails from others [19].

The major objective of this paper is to illustrate that clustering by writing style is a promising approach for grouping e-mails written by the same author. Our method provides the crime investigator a deep insight on the writing styles found in the given anonymous e-mails, in which the clusters and the extracted writeprint could serve as input information for higher-level data mining. To investigate the relative discriminating power of stylometric features, clustering is applied separately to each type (lexical, syntactic, structural and content-specific). In our experiments, we gauge the effects of varying the number of authors and the size of training set on the purity of clustering. Using visualization and browsing features of our developed tool, the investigator is able to explore the process of clusters formation and to evaluate clusters quality.

More explicitly the brief summary of our main contributions are listed below.

Clustering based on stylometric features: traditionally content-based clustering is in use since long to identify the topic of discussion from a collection of documents. The current study dictates that stylometry-based clustering can be used to identify major groups of writing styles from an anonymous e-mail dataset. The claim is supported by calculating *recall* and *F-measure* using Enron e-mail corpus [7].

Preliminary information: sometime, the investigator is provided with just a bunch of anonymous suspicious e-mails and is asked to collect forensically relevant evidence from those unknown messages. Our proposed method can be used to initiate the investigation process by identifying groups of stylistics. The hypothesis is that every author has a unique (or nearly unique) writing style and clustering by stylometric features can group together e-mails of the same author. This hypothesis is supported by extensive experimental results on a real-life dataset in Section 5.

Cluster analysis: we propose a method and develop a tool for the investigator to visualize, browse, and explore the writing styles that are extracted from a collection of anonymous e-mails. The relative strength of different clustering algorithms is evaluated. Our study reveals the relative discriminating power of four different categories of stylometric features. Effects of the number of suspects as well as the number of messages per suspect on the clustering accuracy is addressed in the present study.

Leading to authorship analysis: The suspects' writeprints, extracted in our approach, can be used for authorship attribution (discussed in [16]) of disputed anonymous e-mails.

The rest of the paper is organized as follows: Section 2 reviews the literature. Section 3 formally defines the problem. Section 4 presents the framework for clustering the e-mails by writing styles and mining writeprints. Section 5 examines the effectiveness of our method on a real-life dataset. Section 6 concludes the paper.

2 Related Work

We provide a literature review of stylometric features in Section 2.1 followed by a description of special characteristics of e-mail datasets in Section 2.2. State-of-the-art techniques developed for clustering e-mails are elaborated in Section 2.3.

2.1 Stylometric Features

Often, investigators sometimes finger prints to uniquely identify criminals. In the present era of computer and world wide web, the nature of some crimes as well as the tools used to commit crimes have changed. Traditional tools and techniques may no longer be applicable in prosecuting criminals in a court of law. Stylistics or the study of stylometric features shows that individuals can be identified by their relatively consistent writing style. The writing style of an individual is defined in terms of word usage, selection of special characters, composition of sentences and paragraphs and organization of sentences into paragraphs and paragraphs into documents.

Though, there is no such features set that is optimized and is applicable equally to all people and in all domains. However, previous authorship studies [5,6,10,30] contain lexical, syntactic, structural and content-specific features. A brief description and the relative discriminating capability of each type of these features is given below.

Lexical features are used to learn about the preferred use of isolated characters and words of an individual. Some of the commonly used character-based features are indexed 1-8 in Table 1. These include frequency of individual alphabets (26 letters of English), total number of upper case letters, capital letters used in the beginning of sentences, average number of characters per word, and average number of characters per sentence. The use of such features indicates the preference of an individual for certain special characters or symbols or the preferred choice of selecting certain units. For instance, some people prefer to use '$' symbol instead of word 'dollar', '%' for 'percent', and '#' instead of writing the word 'number'.

Word-based features including word length distribution, words per sentence, and vocabulary richness were very effective in earlier authorship studies [26,27,14]. Recent studies on e-mail authorship analysis [10,28] indicate that word-based stylometry such as vocabulary richness is not very effective due to two reasons. First, e-mail documents and online messages are very short compared to literary and poetry works. Second, word-oriented features are mostly context dependent and can be consciously controlled by people.

Syntactic features, called style markers, consist of all-purpose function words such as 'though', 'where', 'your', punctuation such as '!' and ':', parts-of-speech tags and hyphenation (see Table 1). Mosteller and Wallace [20] were the first who showed the effectiveness of the so-called function words in addressing the issue of Federalist Papers. Burrows [6] used 30-50 typical function words for authorship attribution. Subsequent studies [5] validated the discriminating power of punctuation and function words. Zheng et al. [28] have used more than 300 function words. Stamatatos et al. [24] have used frequencies of parts-of-speech tags, passive account and nominalization count for authorship analysis and document genre identification.

Structural features are helpful in learning about how an individual organizes the layout and structure of his/her documents. For instance, how are sentences organized within paragraphs and paragraphs within documents. Structural features were first suggested by Vel et al. [10,8] for e-mail authorship attribution. In addition to the general structural features, they used features specific to e-mails such as the presence/absence of greetings and farewell remarks and their position within the e-mail body. Moreover, some people use first/last name as a signature while others prefer to include their job title and mailing address as well within e-mails. Malicious e-mails contain no signature and in some cases fake signatures.

Content-specific features are used to characterize certain activities, discussion forums or interest groups by a few key words or terms. For instance people involving in cybercrimes (spamming, phishing and intellectual property theft) commonly use (street words) 'sexy', 'snow', 'download', 'click here' and 'safe' etc. Usually term taxonomy built for one domain are not applicable in other domain and even vary from person to person in the same domain. Zheng et al. [30,28] used around 11 keywords (such as 'sexy', 'for sale', and 'obo' etc.) from the cybercrime taxonomy in authorship analysis experimentations. A more comprehensive list of stylistic features including idiosyncratic features is used in [2].

Idiosyncratic Features include common spelling mistakes such as transcribing 'f' instead of 'ph' say in phishing and grammatical mistakes such as sentences containing incorrect form of verbs. The list of such characteristics varies from person to person and is difficult to control. Gamon [13] claims to have achieved high accuracy by combining certain features including parts-of-speech trigrams, function word frequencies and features derived from semantic graphs.

2.2 E-mail Characteristics

The application of authorship analysis techniques to e-mail datasets is more challenging than historical and literary documents [10]. Literary works are large collections, usually comprising of several sections, subsections and paragraphs. They follow definite grammatical rules and composition styles. They are usually written in a formal template. E-mails on the other hand are short in length usually contain a few sentences or words. Therefore, it is hard to learn about the writing habits of people from their e-mails. Ledger and Merriam [17], for instance, established that authorship analysis results, would not be significant for texts containing less than 500 words.

E-mails are often informal in contents and interactive in style. While writing especially informal e-mails, people may not pay attention to their spelling and grammatical mistakes. Therefore, analytical techniques that are successful in authorship analysis of literary and historic collections may not have the same analytical power on e-mail datasets.

Certain aspects of e-mail documents are rich sources of information. An e-mail has a header, subject, and body. Headers contain information about the path traveled by the e-mail, time stamps, e-mail client information, sender and recipient addresses and recipient responses. Some messages are accompanied by one or more attachments. Such additional information are mostly helpful in learning about the writing styles and behavior of a user. Vel et al. [10] discovered that when applied together with other stylometric features, structural features are very successful in discriminating the writing styles of their authors.

2.3 E-mail Cluster Analysis

To collect creditable evidence against a cybercriminal, a forensic investigator would need to perform several different kinds of analysis. For instance, he/she may want to retrieve all those e-mails which talk about certain crimes say drug, pornography, hacking or terrorism etc. This could be achieved by simple keyword searching or more efficiently by using traditional content-based clustering technique [18]. Similarly, an investigator may want to visualize the general communication patterns of a suspect within his/her community. This could be achieved by using the techniques of social networking and behavior modeling [25]. To identify the true author of a disputed anonymous e-mail, different machine learning techniques (e.g. discussed in [16,2]) can be used.

Holmes and Forsyth [15] and Ledger and Merriam [17] were among the pioneers who applied multivariate clustering technique to text datasets. Later

Baayen et al. [5] performed stylometric clustering in authorship attribution. They considered merely data-driven features, the term used in [1], which include word frequency, letter frequency and sentence length etc. S. Aaronson [1] studied the effects of data-driven features, syntactic features and combination of them. By syntactic features they mean the grammar rules that are extracted by using language parser. They claimed that the clustering accuracy is significantly better than the previous studies.

Abbasi and Chen [2] studied the effects of stylometric features on similarity detection by employing Principal Component Analysis (PCA) and their newly proposed technique, called *Writeprints*. To the best of our understanding we have not seen any study which addresses all the questions stated in the problem statement.

The traditional content-based clustering [18], where each e-mail is represented as a 'bag of words', is not appropriate in the context of the problem studied in this paper. Initially, Holmes and Forsyth [15] applied Principal Component Analysis (PCA) for stylometry based clustering. Later on Ledger and Merriam [17] performed clustering for authorship analysis on text datasets.

Li et al. [18] applied content-based clustering on e-mails by employing their proposed algorithm. They used to feed e-mail subject to a Natural Language (NL) parser. Output of the parser is then given to their proposed algorithm to generalize them to what they called meaningful Generalized Sentence Patterns (GSP). Using GSP as a false class label, clustering is performed in a supervised manner. Work of Li et al. [18] was limited to the e-mail subject and it suffered from GSP redundancy.

Internet-based reputation system, used in online market, is manipulated by the use of multiple alias of the same individual. Novak et al. [21] have proposed a new algorithm to identify when two aliases belong to the same individual while preserving the privacy. The technique was successfully applied to postings of different bulletin boards with achieving more than 90% accuracy. To address the same issue of anonymity Abbasi and Chen [3,2] have proposed a novel technique called *writeprints* for authorship identification and similarity detection. They have used a very extended feature list including idiosyncratic features in their experimentations. In similarity detection part, they take an anonymous entity and compare it with all other entities and then calculate a score. If the score is above a certain predefined value the entity in hand is clustered with the matched entity.

3 The Problem

The problem addressed in this paper is stated as: a forensic investigator has a collection of suspicious anonymous e-mails E. The e-mails are (presumably) written by K suspects, but the investigator may or may not know the number of suspects in advance. The investigator wants to get an insight into the writing styles of an e-mail collection E, and wants to identify major groups of writing styles called writeprints $\{WP_1, \ldots, WP_k\}$ in E. Our objective is to develop a framework that allows the investigator to extract stylometric features from E and group e-mails E into clusters by stylometric features. In this paper, we propose a method and develop a tool for the investigator to visualize, browse, and explore the writing styles, found in a collection of anonymous e-mails E.

We measure discriminating capabilities of different stylometric features in e-mail data clustering. For example, if different collections of e-mails are written on distinct topics, content-specific features may give better clustering results than style markers. This study also focuses on evaluating different state-of-the-art clustering algorithms and determining which algorithm is more suitable in a specific scenario. For instance, EM may be a better option if an investigator does not have any clue about the number of authors contributing to a dataset. Likewise, our study will help users understand the internal structure of an e-mail corpus in terms of different writing style features and to decide on how to narrow down the investigation.

4 Our Method

The general idea of our proposed method, depicted in Figure 1, can be summarized in five phases: (1) Pretreatment: includes extracting e-mail body and applying standard preprocessing techniques of cleaning, tokenization and stemming. At the end of first phase, a list of all the tokens including stemmed words is obtained. (2) Stylometric features extraction: is employed to identify the pertinent writing style features found in the anonymous e-mail dataset. Thus each e-mail is converted into a vector of numbers. (3) Stylometry-based clustering: is applied to identify major groups of stylistics belonging to different authors. (4) Frequent patterns mining: is applied to unveil hidden association among different stylometric features. (5) Writeprint mining: provided that each cluster of e-mails obtained in phase three is written by the same author, we can extract the writeprint from each cluster that represent the unique writing style of one author.

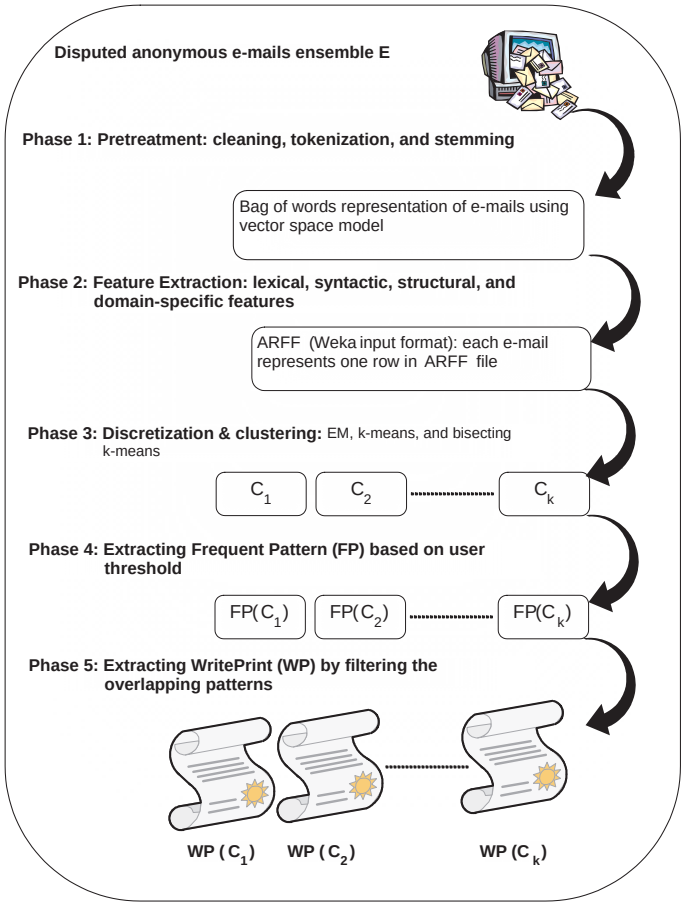

Fig. 1. Mining WritePrints $\{WP_1, \ldots, WP_k\}$ from anonymous e-mails E

4.1 Pre-Treatment

Each e-mail is converted into a stream of characters. Using Java tokenizer API, each character stream is converted into tokens or words. Unlike content-based clustering [18], in which syntactic features are usually dropped, we calculate these features. In our experiments, we have used more than 300 function words that are listed in Table 1. A word may appear in different forms which usually increase dimensionality of the features set. To converge all such variations of the same word to its root, stemming algorithms are applied. Porter2 [23,22] is a popular stemming algorithm used by data mining and Natural Language Processing (NLP) community. We modified Porter2 by adding some more rules to fit it into our proposed approach.

Certain word sequences like 'United States of America' and 'United Arab Emirates' etc. often appear together. Therefore, we developed a module to automatically scan those sequences and treat them as single tokens. This help in reducing the features dimensionality. Using vector space model representation, each e-mail μ_i is converted into an n-dimensional vector of features $\mu_i =$

$\{F_1, \ldots, F_n\}$. Once all e-mails are converted into feature vectors, normalization is applied to the columns as needed. The purpose of normalization is to limit values of a certain feature to [0,1] and thus avoid overweighing some attribute by others.

4.2 Features Extraction

The total number of stylometric features discovered exceeds 1000 [2]. In our experiments we have used 419 features (listed in Table 1 and Table 2). In general, there are two types of features. The first type is a numerical value, e.g., the frequencies of some individual characters and punctuation. Numerical values are normalized to $[0, 1]$ by dividing all the occurrences of a feature item by the maximum. Normalization is applied across the entire collection of e-mails. The second type is a binary value, e.g., whether an e-mail has greetings. Certain features are calculated by applying certain functions like Yule's K measure to compute vocabulary richness.

Some features are extracted by calculating the ratios of other known features. For instance, computing ratio of word-length frequency distribution to total number of words (W) is considered as a separate feature. Once feature extraction is done, each e-mail is represented as a vector of feature values. In this study we focused more on using structural features as they play significant role in distinguishing writing styles.

Features indexed at 1-8 involves calculation of frequencies of individual characters. Upper case letters appearing in the beginning of a sentence are counted separately. Different words of length 1-3 characters (such as 'is', 'are', 'or', 'and' etc.) are mostly context-independent and are considered as a separate feature. Frequencies of Words of various lengths 1-20 characters (indexed at 14) are counted separately. Hepax Legomena and Hapax dislegomena are the terms used for once-occurring and twice-occurring words. As mentioned earlier, we have used more than 300 function words (indexed at 20).

Structural feature, given at index 21 in Table 2, is of type boolean. It checks whether an e-mail has welcoming and farewell greetings. Paragraph separator can be a blank line or just a tab/indentation or there may be no separator between paragraphs. For content-specific features, we selected about 13 high frequency words from the Enron e-mail dataset. The words are listed at index 34 in Table 2.

Table 1
Lexical and Syntactic Features

Features Type	**Features**
Lexical: Character-based	1. Character count (N)
	2. Ratio of digits to N
	3. Ratio of letters to N
	4. Ratio of uppercase letters to N
	5. Ratio of spaces to N
	6. Ratio of tabs to N
	7. Occurrences of alphabets (A-Z) (26 features)
	8. Occurrences of special characters: < > % | { } [ ] / \ @ # ~ + - * $ ^ & _ ÷ (21 features)
Lexical: Word-based	9. Token count(T)
	10. Average sentence length in terms of characters
	11. Average token length
	12. Ratio of characters in words to N
	13. Ratio of short words (1-3 characters) to T
	14. Ratio of word length frequency distribution to T (20 features)
	15. Ratio of types to T
	16. Vocabulary richness(Yule's K measure)
	17. Hapax legomena
	18. Hapax dislegomena
Syntactic Features	19. Occurrences of punctuations , . ? ! : ; ' " (8 features)
	20. Occurrences of function words (303 features)

4.3 Clustering

Clustering is the process of grouping similar objects together. Intuitively, the resulting cluster solution should have high intra-cluster similarity, but low inter-cluster similarity. In the context of our studied problem, e-mails in the same cluster should have similar writing styles, but e-mails in different clusters should have different writing styles.

Our proposed method is evaluated by using three clustering algorithms: Expectation Maximization (EM), k-means, and bisecting k-means. k-means and bisecting k-means are the most commonly used formulae. *Expectation Max-*

Table 2
Structural Features

Features Type	**Features**
Structural Features	
	21. lines in an e-mail
	22. Sentence count
	23. Paragraph count
	24. Presence/absence of greetings
	25. Has tab as separators between paragraphs
	26. Has blank line between paragraphs
	27. Presence/absence of separator between paragraphs
	28. Average paragraph length in terms of characters
	29. Average paragraph length in terms of words
	30. Average paragraph length in terms of sentences
	31. Use e-mail as signature
	32. Use telephone as signature
	33. Use URL as signature
Domain-specific Features	34. agreement, team, section, good, parties, office, time, pick, draft, notice, questions, contracts, day (13 features)

imization (EM) algorithm, first proposed in [11], is often employed where it is hard to predict the value of K (number of clusters). For instance, during forensic analysis of anonymous e-mails, the investigator may not know the total number of authors (or different writing styles) within that collection. In a more common scenario, a user may want to validate the results obtained by other clustering algorithms say k-means, or bisecting k-means.

To measure the purity of resultant clusters and validate our experimental results, the commonly used formula called F-measure is applied [12]. F-measure is derived from *precision* and *recall*, which are the accuracy measures commonly employed in the field of Information Retrieval (IR). The three functions are shown by the following mathematical equations.

$$recall(N_p, C_q) = \frac{O_{pq}}{|N_p|} \tag{1}$$

$$precision(N_p, C_q) = \frac{O_{pq}}{|C_q|} \tag{2}$$

Table 3
Feature Items Extracted from E-mail Clusters of Ensemble E

		Feature F_1			**Feature** F_2			**Feature** F_3		
Cluster C	**Message** μ	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	$F_{3,1}$	$F_{3,2}$	$F_{3,3}$
C_1	μ_1	0	1	0	0	0	1	0	0	1
C_1	μ_2	0	1	0	0	0	1	0	0	1
C_1	μ_3	0	1	0	0	1	0	0	0	1
C_1	μ_4	1	0	0	0	0	1	0	0	1
C_2	μ_5	1	0	0	0	1	0	0	1	0
C_2	μ_6	1	0	0	0	1	0	0	0	1
C_2	μ_7	1	0	0	1	0	0	0	0	1
C_3	μ_8	0	1	0	1	0	0	1	0	0
C_3	μ_9	0	0	1	1	0	0	1	0	0
C_3	μ_{10}	0	1	0	1	0	0	0	1	0
C_3	μ_{11}	0	1	0	1	0	0	1	0	0

$$F(N_p, C_q) = \frac{2 * recall(N_p, C_q) * precision(N_p, C_q)}{recall(N_p, C_q) + precision(N_p, C_q)} \tag{3}$$

where O_{pq} is the number of members of actual (natural) class N_p in cluster C_q, N_p is the actual class of a data object O_{pq} and C_q is the assigned cluster of O_{pq}.

We have developed a software toolkit that can be used to perform the entire writing style mining process. Its GUI interface helps a user in features selection, algorithm selection, and parameter selection (such as the number of clusters). This will help gauge the relative strength of each type of writing style features in discriminating the styles of different people. Our software tool has the capability to compare different clustering algorithms and select an appropriate algorithm for particular e-mail dataset by determining which algorithm perform better within a certain context.

4.4 Mining Frequent Patterns (FP)

Once clusters $\{C_1, \ldots, C_k\}$ are formed, each of these clusters is used to determine the writing style contained in that particular cluster C_i. Intuitively, the "writing style" in an ensemble of e-mails E is a combination of a subset of feature items that *frequently* occurs together in certain e-mails $\{\mu_1, \ldots, \mu_n\} \in E$.

For instance, a person may use certain formal words with nearly the same proportion in most of his formal e-mails. By feature items we mean the discretized value of a feature, discussed in next paragraph. We capture such frequently occurred patterns by concept of *frequent itemset* [4], in a way similar to the one described in [16]. The process consists of two major steps. (1) Patterns (P) extraction, and (2) Frequent Patterns (FP) calculation. Below, we first define what exactly writing styles and frequent patterns mean.

Let $F = \{F_1, \ldots, F_n\}$ be a set of features as shown in Table 1 and Table 2. To fit into the method of *frequent itemset* [4], we discretize each feature F_i into some intervals $\{F_{i,1}, \ldots, F_{i,j}\}$, where each $F_{i,b} \in \{F_{i,1}, \ldots, F_{i,j}\}$ denotes a feature item b of a feature F_i (as shown in Table 3). Unlike [16], who discretized feature values into equal number of intervals, we ask the user to specify maximum number of occurrences per interval. Applying binary division we divide feature values into two groups G_1 and G_2. Each group is in turn divided into two subgroups subject to the condition that number of occurrences (within a group) exceeds the threshold. The process is repeated until all the feature occurrences are grouped. With the proposed method the interval size as well as the total number of intervals is determined dynamically for each feature.

Let $P \subseteq F$ be a set of feature items called a *pattern*. An e-mail μ contains a pattern P if $P \subseteq \mu$. A pattern that contains q feature items is a *q-pattern*. For example, as depicted in Table 4, pattern $F = \{F_{1,2}, F_{2,3}, F_{2,3}\}$, as extracted from e-mail μ_1 is a 3-pattern. The *support* of a pattern P is the percentage of e-mails in E_i that contain P. A pattern P is a *frequent pattern* in a set of e-mails E_i if the support of P is greater than or equal to some user-specified minimum support (threshold). The writing pattern, found in a cluster C_i, is represented as a set of frequent patterns, denoted by $FP(C_i) = \{F_{1,1}, \ldots, F_{m,n}\}$, extracted from e-mails E_i contained in cluster C_i. Where integers m and n represent feature number and interval number, respectively.

We trying to explain the above mentioned concepts in the context of our proposed approach *writing style mining*, by using a running example. Suppose at the end of clustering phase we have three clusters, C_1 with $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ e-mails, C_2 with $\{\mu_5, \mu_6, \mu_7\}$ e-mails, and C_3 containing $\{\mu_8, \mu_9, \mu_{10}, \mu_{11}\}$ e-mails, as shown in Table 4. The presence of a feature item within an e-mail is indicated by a '1' in the respective cell and vice versa. The extracted patterns of each e-mail μ_i and the associated cluster C_i are shown in Table 4. It's worth mentioning that discretization of the extracted features $\{F_1, F_2, F_3\}$ into respective *feature items* is done after the clustering phase.

Now, to calculate frequent patterns for each cluster, we assume that the user defined $min\_sup = 0.4$. It means that a pattern $P = \{F_{1,1}, \ldots, F_{m,n}\}$ is frequent if at least 40% of e-mails within a cluster C_i contain all feature items in P. For instance, pattern $\{F_{1,2}, F_{2,3}, F_{3,3}\}$ is a frequent pattern because at

Table 4
Patterns Extracted from Ensemble E

Cluster(C)	**E-mail(μ)**	**Pattern(P)**
C_1	μ_1	$\{F_{1,2}, F_{2,3}, F_{3,3}\}$
	μ_2	$\{F_{1,2}, F_{2,3}, F_{3,3}\}$
	μ_3	$\{F_{1,2}, F_{2,2}, F_{3,3}\}$
	μ_4	$\{F_{1,1}, F_{2,3}, F_{3,3}\}$
C_2	μ_5	$\{F_{1,1}, F_{2,2}, F_{3,2}\}$
	μ_6	$\{F_{1,1}, F_{2,2}, F_{3,3}\}$
	μ_7	$\{F_{1,1}, F_{2,1}, F_{3,3}\}$
C_3	μ_8	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$
	μ_9	$\{F_{1,3}, F_{2,1}, F_{3,1}\}$
	μ_{10}	$\{F_{1,2}, F_{2,1}, F_{3,2}\}$
	μ_{11}	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$

Table 5
Frequent Patterns (FP) Extracted from Ensemble E

Cluster (C)	**Frequent Patterns (FP)**
C_1	$\{F_{1,2}, F_{2,3}, F_{3,3}\}$
C_2	$\{F_{1,1}, F_{2,2}, F_{3,3}\}$
C_3	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$

least 3 and/or 4 e-mails of cluster C_1 contain this pattern. On the other hand pattern $\{F_{2,2}\}$ is contained in only one e-mail of the same cluster and therefore is not a frequent pattern. Similarly, pattern $\{F_{1,2}, F_{2,1}, F_{3,1}\}$ appears in at least three out of four e-mails of cluster C_3 and so is a frequent pattern.

In contrast, each of the patterns $\{F_{1,3}\}$ and $\{F_{3,2}\}$ appears in only one e-mail of the associated cluster and thus are not frequent patters. $\{F_{1,2}, F_{2,1}, F_{3,1}\}$ and $\{F_{1,3}\}$ are 3-frequent patterns and 1-frequent patterns, respectively. In our example, applying $min\_sup = 0.4$ means that a pattern is a frequent pattern if it is contained in at least two out of three and/or four e-mails. All the frequent patterns and their associated e-mails/clusters, extracted from ensemble E, are shown in Table 5.

Table 6
Writing Styles (WS) Mined from Ensemble E

Cluster (C)	**Writing Styles (WS)**
C_1	$\{F_{2,3}\}$
C_2	$\{F_{1,1}, F_{2,2}\}$
C_3	$\{F_{2,1}, F_{3,1}\}$

4.5 Writing Styles

A writeprint should uniquely identify an individual. Patterns that are shared by more than one clusters are dropped. For instance in our example $F_{1,2}$ is shared by cluster C_1 and C_3 while $\{F_{3,3}\}$ is common among C_1 and C_2. Therefore, both patterns $\{F_{1,2}\}$ and $\{F_{3,3}\}$ are deleted from concerned clusters. The remaining frequent patterns constitute the unique (or near to unique) writeprints $\{WP_1, WP_2, WP_3\}$ as mined from clusters C_1, C_2 and C_3, as shown in Table 6. From these results we conclude that the e-mail ensemble E contained e-mails of 3 suspects. The distinct writeprints $\{WP_1, \ldots, WP_k\}$ are used for identifying the true author of a malicious e-mail, as described in [16].

5 Experiments and Evaluation

Our goal in this section is to evaluate our proposed method and to analyze whether it can precisely identify the different writing styles of an e-mail collection. The set of experiments need to be designed such that to find answers to the following questions. Which of the clustering algorithm perform better than others for a given e-mail dataset? What is the relative strength of each of the four different types of writing features? What is the effect of varying the number of authors on the experimental results? In our experiments, we also investigate the effects of varying the number of e-mail messages per author on clusters quality.

We have performed three sets of experiments. (1) To evaluate stylometric features in terms of F-measure we applied clustering over nine different combinations of these features. (2) Varying the number of authors while keeping other parameters (messages per author and features) constant. (3) In the third set of experiments is to check the effects of number of messages per author.

In all the three set of experiments three different clustering algorithms, namely EM, k-means and bisecting k-means were applied. Different feature combinations are $\{T_1, T_2, T_3, T_4, T_1+T_2, T_1+T_3, T_2+T_3, T_1+T_2+T_3, T_1+T_2+T_3+T_4, \}$, where T_1, T_2, T_3 and T_4 represent lexical, syntactic, structural and content-

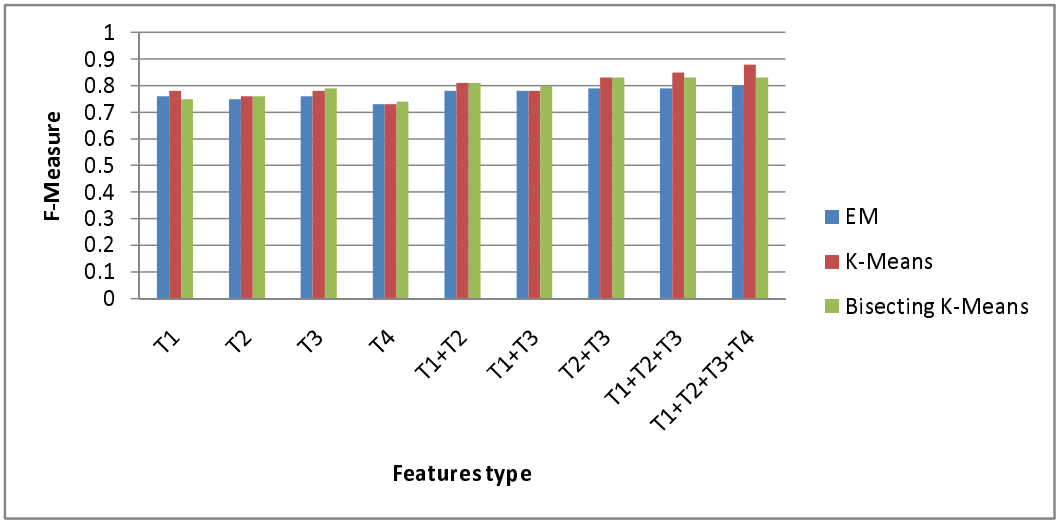

Fig. 2. F-Measure vs. Feature Type and Clustering Algorithms ($Authors = 5$, $Messages = 40$)

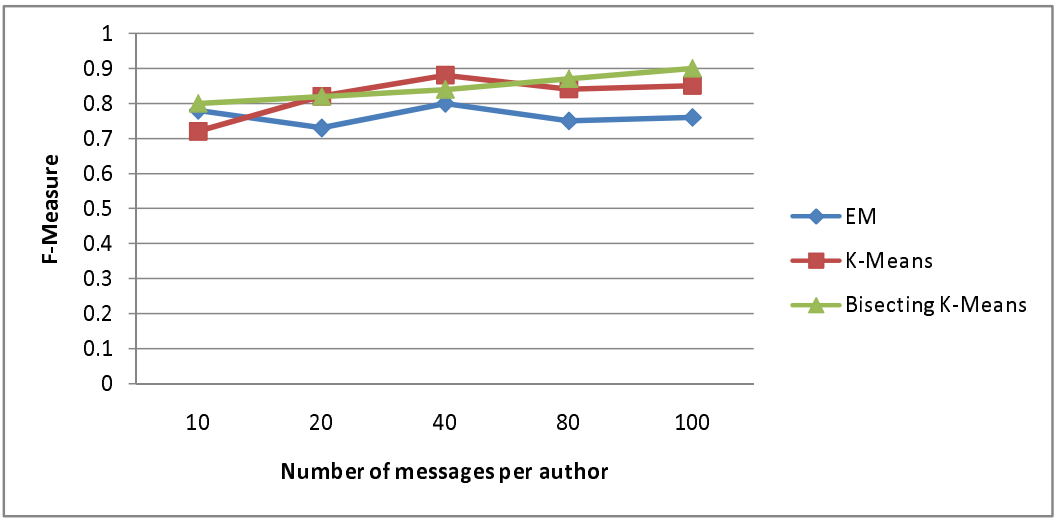

Fig. 3. F-Measure vs. Features Type and Clustering Algorithms ($Authors = 5$, $Features = T_1 + T_2 + T_3 + T_4$)

specific features respectively.

We used a real-life e-mail data: Enron E-mail Dataset [7], which contains 200,399 e-mails of about 150 employees of Enron corporation (after cleaning). We randomly selected h employees from the Enron E-mail Dataset, representing h authors $\{A_1, \ldots, A_h\}$. For each author A_i, we selected x of A_i's e-mails. Where h varies from three to ten while value of x is selected from {10, 20, 40, 80, 100}.

In the first set of experiments, we have selected 40 e-mails from each one of the five authors. The results of the three clustering algorithms are shown in

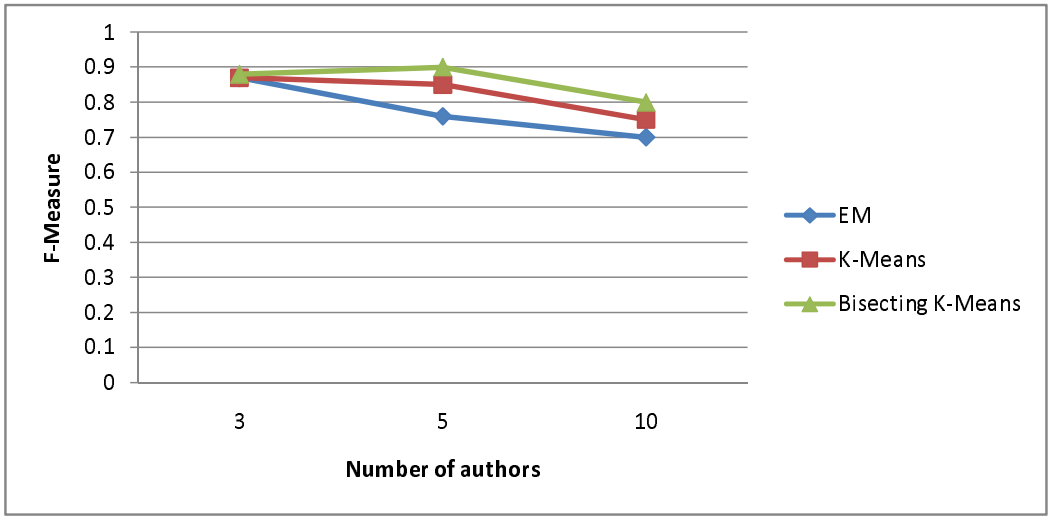

Fig. 4. F-Measure vs. Features Type and Clustering Algorithms ($Messages = 100$, $Features = T_1 + T_2 + T_3 + T_4$)

Figure 2. It illustrates that the value of F-measure spans from 0.73 to 0.80 for EM, from 0.73 to 0.88 for k-means, and from 0.75 to 0.83 for bisecting k-means. The better results of k-means and bisecting k-means over EM (in this set of experiments) indicates that knowing the number of clusters K, one can obtain better results. Results of k-means are better than bisecting k-means. Initially these results seemed unexpected which were later on validated after completing all sets of experiments. K-means performed better as compared to bisecting k-means upto 40 e-mails per author. By increasing e-mails beyond 40 for each author the accuracy of bisecting k-means was increasing. It seems that bisecting k-means is more scalable than EM and k-means.

Looking at the individual features, T_4 (content-specific features) performed poorly while T_3 (structural features) produced very good results. These two trends are matching to the previous stylometric studies. The best results are obtained by applying k-means on $T_1 + T_2 + T_3 + T_4$, combination of all four types of features. By adding contents-specific features to $T_1 + T_2 + T_3$, we do not see any noticeable improvement in the results of EM and bisecting k-means. The selected keywords are probably common among e-mails of the selected authors. Another important observation is that $\{T_2 + T_3\}$ results are better than other two features combination (such as $T_1 + T_2$ and $T_1 + T_3$).

In the next set of experiments the number of authors (five) and features set ($T_1 + T_2 + T_3 + T_4$) were kept constant. The value of F-measure increases with increasing the number of e-mails per author, as shown in Figure 3. K-means and bisecting k-means achieve 90% purity for 40 messages per author while EM results are inconsistent. Increasing the number of messages per author beyond 40 negatively affect all the three algorithms. Among the three EM drops faster than the other two, and bisecting k-means is more robust compared to simple

k-means. These results explain the relative behavior of these algorithms in terms of scalability.

In the third set of experiments (depicted in Figure 3), we considered ($T_1 + T_2 + T_3 + T_4$ features and picked 100 e-mails for each author. Value of F-measure reaches 0.91 for bisecting k-means for all the combinations in this set of experiments. Accuracy of all the three clustering models drops as more authors are added.

The best accuracy was achieved by applying k-means over a combination of all four feature types when e-mails per user is limited to 40. Bisecting k-means is a better choice when there more authors and the training set is larger. Taking into account the topic of discussion better results can be obtained by selecting domain-specific words carefully. One way could to identify author-specific keywords by apply content-based clustering on e-mails of each author separately. Results of EM are insignificant and are hard to improve by parameter tuning.

6 Conclusion

We have developed an e-mail analysis framework to extract different writing styles from a collection of anonymous e-mails. Our proposed method first clusters the given anonymous e-mails based on their stylometric features and then extracts unique (near to unique) writing styles from each resultant cluster. This will help the investigator to learn about the potential authors of anonymous e-mail dataset. The writing styles in terms of feature patterns provide more concrete evidence than producing some statistical numbers. Our experimental results show that clustering is an appropriate technique for grouping e-mails on the basis of stylometric features.

The decreased accuracy of the three clustering techniques due to increase in the number of candidate authors and sample size indicates scalability issues. Therefore, the need is to investigate more robust clustering techniques. Moreover, existing features list need to be expanded by including idiosyncratic features and using combined features approach (see [13]).

Existing research studies show that content-specific keywords can play a more important role in style mining when used in specific contexts like cybercrime investigation. Therefore, it is imperative to develop a sound technique for keywords selection. Features optimization will certainly be helpful in determining authors' style that is a true representative. Furthered, human behavior changes from context to context and from person to person. The need is to develop methods for capturing style variations for better authorship results. Addressing language multiplicity is another research direction. The research

of stylometric forensics is still in its infancy stage. There is still a long way to develop a comprehensive, reliable authorship analysis approach before it can be widely accepted in courts of law.

References

[1] S. Aaronson. Stylometric clustering, a comparative analysis of data-driven and syntactic features. Technical report, UC Berkeley, 1999.

[2] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29, 2008.

[3] A. Abbasi, H. Chen, and J. Nunamaker. Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, 5(1):49–78, 2008.

[4] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM International Conference on Management of Data (SIGMOD)*, pages 207–216, Washington, D.C., United States, 1993.

[5] R. H. Baayen, H. Van Halteren, and F. J. Tweedie. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 2:110–120, 1996.

[6] J. F. Burrows. Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing*, 2:61–67, 1987.

[7] W. W. Cohen. Enron Email Dataset, 2004.

[8] M. Corney, O. de Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In *Proc. of the 18th Annual Computer Security Applications Conference*, page 282, 2002.

[9] O. de Vel. Mining e-mail authorship. In *Proc. of the Workshop on Text Mining in ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2000.

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

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[12] B. C. M. Fung, K. Wang, and M. Ester. Hierarchical document clustering using frequent itemsets. In *Proc. of the 3rd SIAM International Conference on Data Mining (SDM)*, pages 59–70, San Francisco, CA, May 2003.

[13] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proc. of the 20th International Conference on Computational Linguistics*, page 611, Geneva, Switzerland, 2004.

[14] D. I. Holmes. The evolution of stylometry in humanities. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[15] D. I. Holmes and R. S. Forsyth. The federalist revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127, 1995.

[16] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:42–51, 2008.

[17] G. R. Ledger and T. V. N. Merriam. Shakespeare, Fletcher, and the two Noble Kinsmen. *Literary and Linguistic Computing*, 9:235–248, 1994.

[18] H. Li, D. Shen, B. Zhang, Z. Chen, and Q. Yang. Adding semantics to email clustering. In *Proc. of the 6th International Conference on Data Mining (ICDM)*, pages 938–942, Washington, DC, USA, 2006. IEEE Computer Society.

[19] J. Li, R. Zheng, and H. Chen. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82, 2006.

[20] F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist.* behavioral science:quantitative methods edition. Addison-Wesley, Massachusetts, 1964.

[21] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. In *Proc. of the 13th International Conference on World Wide Web (WWW)*, pages 30–39, 2004.

[22] C. D. Paice. Another stemmer. *SIGIR Forum*, 24(3):56–61, 1990.

[23] M. F. Porter. An algorithm for suffix stripping. *Program*, 3(14):130–137, October 1980.

[24] E. Stamatatos, G. Kokkinakis, and N. Fakotakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495, 2000.

[25] S. J. Stolfo, G. Creamer, and S. Hershkop. A temporal based forensic analysis of electronic communication. In *Proc. of the 2006 International Conference on Digital Government Research*, pages 23–24, San Diego, CA, 2006.

[26] G. U. Yule. On sentence length as a statistical characteristic of style in prose. *Biometrika*, 30:363–390, 1938.

[27] G. U. Yule. *The statistical study of literary vocabulary.* Cambridge University Press, Cambridge, UK, 1944.

[28] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):1532–2882, February 2006.

[29] R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercrime investigation. In *Proc. of the 1st International Symposium on Intelligence and Security Informatics (ISI)*, pages 59–73. Springer-Verlag, 2003.

[30] R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercrime investigation. In *Proc. of the 1st International Symposium on Intelligence and Security Informatics (ISI)*, Tucson, Arizona, 2003.