available at www.sciencedirect.com

ScienceDirect

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

ELSEVIER

Digital Investigation

Mining writeprints from anonymous e-mails for forensic investigation

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

Faculty of Engineering and Computer Science, Concordia University, Montreal, Quebec, Canada H3G 1M8

ARTICLE INFO

Article history:
Received 18 December 2009
Accepted 19 March 2010

Keywords:
E-mail
Writing styles
Writeprint
Forensic investigation
Clustering
Classification
Stylometric features
Authorship analysis

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-mail 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-mail 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.

1. Introduction

The cyber world provides a convenient platform for criminals to anonymously conduct their illegal activities, such as spamming and phishing, etc. 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 use e-mail system 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, has received some attention in recent cyber forensic investigation cases (Zheng et al., 2003a). These techniques 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 (Abbasi and Chen, 2008; Iqbal et al., 2008; de Vel, 2000). 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 (Koppel et al., 2009). These techniques, however, need

[*] *Corresponding author*. Tel.: +1 514 848 2424x5919; fax: +1 514 848 3171.
E-mail addresses: iqbal_f@ciise.concordia.ca (F. Iqbal), h_binsal@ciise.concordia.ca (H. Binsalleeh), fung@ciise.concordia.ca (B.C.M. Fung), debbabi@ciise.concordia.ca (M. Debbabi).

1742-2876/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2010.03.003

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 on the worst case scenario where neither the 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 anonymous e-mails. Given a collection of suspicious anonymous e-mails *E*, presumably written by a group of unknown suspects with no sample writings. Furthermore, a forensic investigator, may or may not know the actual number of authors in *E*. To get a deeper insight into the writing styles of the given anonymous e-mails, the investigator can apply our proposed method. First, cluster the e-mails by the writing styles. Our study suggests that clustering is an effective approach for identifying e-mails written by the same authors. Then, extract the writeprint (c.f. fingerprint) (Li et al., 2006a) from each group of e-mails. Writeprint is a set of writing style features that is specific enough to distinguish one author's e-mails from others (Li et al., 2006b).

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.

We summarize the contributions of this paper as follows.

1.1. *Clustering based on stylometric features*

Traditionally content-based clustering is used to identify the topic of discussion from a collection of documents. In contrast, this paper suggests that clustering the anonymous e-mail messages by writing styles is very effective for identifying e-mails written by the same author.

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 experimental results performed on real-life e-mail corpus (Enron e-mail corpus (Cohen, 2004)).

1.2. *Preliminary information*

Often, the investigator is provided with just a collection 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.

1.3. *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 this study.

1.4. *Leading to authorship analysis*

The suspects' writeprints extracted by our approach can be used for authorship attribution (Iqbal et al., 2008) of disputed anonymous e-mails.

The rest of the paper is organized as follows: Section 2 reviews the literature in authorship analysis. 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 dataset in Section 2.2. State-of-the-art techniques developed for clustering e-mails are elaborated in Section 2.3.

2.1. *Stylometric features*

Traditionally, fingerprints were used to uniquely identify criminals. In the present era of computer and world wide web, the nature of most crimes as well as the tools used to commit crimes have changed. Traditional tools and techniques may no longer be applicable in prosecuting cyber criminals in a court of law. Stylistics or the study of stylometric features shows that individuals can be identified by their relatively consistent writing styles. 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 (Baayen et al., 1996; Burrows, 1987; de Vel et al., 2001; Zheng et al., 2003b) contain lexical, syntactic, structural and content-specific features. A brief description and the relative discriminating capability of each type of these features are 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

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)

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 (Yule, 1938, 1944; Holmes, 1998). Recent studies on e-mail authorship analysis (de Vel et al., 2001; Zheng et al., 2006) indicate that word-based stylometry such as vocabulary richness is not very effective due to two reasons. First, e-mail messages and online documents 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 (1964) were the first who showed the effectiveness of the so-called function words in addressing the issue of Federalist Papers. Burrows (1987) used 30–50 typical function words for authorship attribution. Subsequent studies (Baayen et al., 1996) validated the discriminating power of punctuation and function words. Zheng et al. (2006) have used more than 300 function words. Stamatatos et al. (2000) 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 de Vel et al. (2001) (Corney et al., 2002) 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 may contain fake signatures.

Content-specific features are used to characterize certain activities, discussion forums or interest groups by a few keywords 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 is not applicable in other domain and even vary from person to person in the same domain. Zheng et al. (2003b, 2006) 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 Abbasi and Chen (2008).

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 (2004) 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 (de Vel et al., 2001). 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 (1994), 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

Please cite this article in press as: Iqbal F, et al., Mining writeprints from anonymous e-mails for forensic investigation, Digit. Investig. (2010), doi:10.1016/j.diin.2010.03.003

are accompanied by one or more attachments. Such additional information is mostly helpful in learning about the writing styles and behavior of a user. de Vel et al. (2001) discovered that when applied together with other stylometric features, structural features are very successful in discriminating the writing styles of the 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 (Li et al., 2006a). 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 (Stolfo et al., 2006). To identify the true author of a disputed anonymous e-mail, different machine learning techniques (e.g. discussed in Iqbal et al., 2008; Abbasi and Chen, 2008) can be used.

Holmes and Forsyth (1995) and Ledger and Merriam (1994) were among the pioneers who applied multivariate clustering techniques to text dataset. Later Baayen et al. (1996) performed stylometric clustering in authorship attribution. They considered merely data-driven features, the term used in Aaronson (1999), which include word frequency, letter frequency and sentence length, etc. Aaronson (1999) 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 (2008) 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 (Li et al., 2006a), 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 (1995) applied Principal Component Analysis (PCA) for stylometry-based clustering. Later on Ledger and Merriam (1994) performed clustering for authorship analysis on text datasets.

Li et al. (2006a) 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. (2006a) 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 aliases of the same individual. Novak et al. (2004) 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 et al., (2008), Abbasi and Chen (2008) have proposed a novel technique called *Writeprints* for authorship identification and similarity detection. They have used an 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 Fig. 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. Then 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

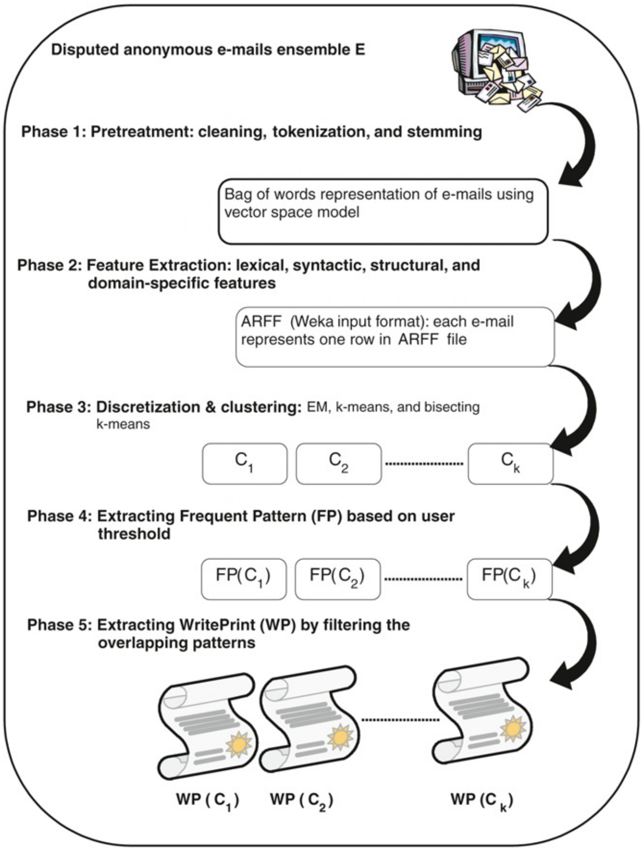
Disputed anonymous e-mails ensemble E
Phase 1: Pretreatment: cleaning, tokenization, and stemming
Bag of words representation of e-mails using vector space model
Phase 2: Feature Extraction: lexical, syntactic, structural, and domain-specific features
ARFF (Weka input format): each e-mail represents one row in ARFF file
Phase 3: Discretization & clustering: EM, k-means, and bisecting k-means
C1
C2
Ck
Phase 4: Extracting Frequent Pattern (FP) based on user threshold
FP(C1)
FP(C2)
FP(Ck)
Phase 5: Extracting WritePrint (WP) by filtering the overlapping patterns
WP (C1)
WP (C2)
WP (Ck)

Fig. 1 – Mining WritePrints {WP_1,…,WP_k} from anonymous e-mails *E*.

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.

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 (Li et al., 2006a), 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 (Porter, 1980; Paice, 1990) 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 so far exceeds 1000 (Abbasi and Chen, 2008). 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 or not ?. 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 focus 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

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)

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.

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 Maximization (EM) algorithm*, first proposed in Dempster et al. (1977), 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 (Fung et al., 2003). 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.

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

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

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

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, etc.). 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* (Agrawal et al., 1993), in a way similar to the one described in Iqbal et al. (2008). 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.

Table 3 – Feature items extracted from e-mail clusters of ensemble *E*.

Cluster C	Message μ	Feature F_1			Feature F_2			Feature F_3		
		$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	$F_{3,1}$	$F_{3,2}$	$F_{3,3}$
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

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* (Agrawal et al., 1993), 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 our previous work (Iqbal et al., 2008) that discretized feature values into equal number of intervals, we ask the user to specify maximum number of occurrences per interval. For each numerical feature, we divide the value into two groups G1 and G2. 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.

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}	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$

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 $P = \{F_{1,2}, F_{2,3}, F_{2,3}\}$, 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 use a running example to explain the above concepts of our proposed writing style mining approach. 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. Its 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 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.

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 contains 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 Iqbal et al. (2008).

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	$\{F_{1,2}, F_{2,1}, F_{3,1}\}$

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	$\{F_{2,1}, F_{3,1}\}$

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 we 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-specific features respectively.

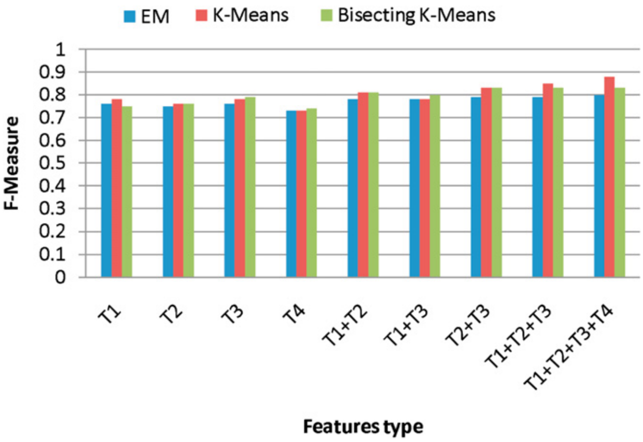

Fig. 2 – F-Measure vs. Feature Type and Clustering Algorithms (*Authors* = 5, *Messages* = 40).

We used a real-life e-mail data: Enron E-mail Dataset (Cohen, 2004), 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. 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) indicate 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 for upto 40 e-mails per author. By increasing e-mails beyond 40 for each author the accuracy of bisecting k-means started 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 any 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 Figs. 3 and 4. 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.

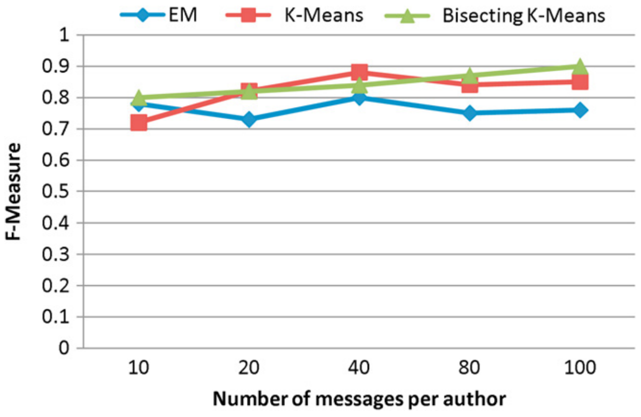

Fig. 3 – F-Measure vs. Features Type and Clustering Algorithms (*Authors* = 5, *Features* = $T_1 + T_2 + T_3 + T_4$).

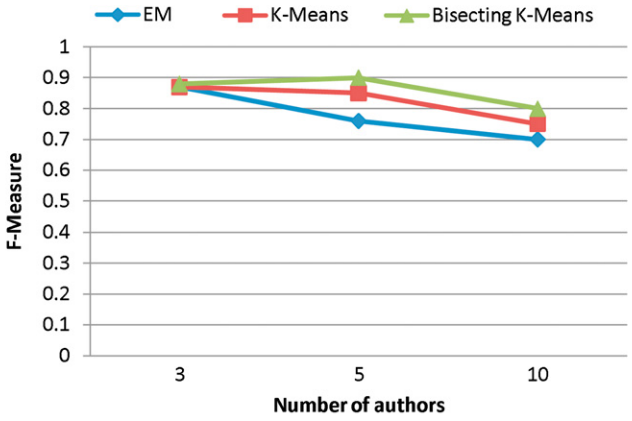

Fig. 4 – F-Measure vs. Number of authors and Clustering Algorithms (*Messages* = 100, *Features* = $T_1 + T_2 + T_3 + T_4$).

In the third set of experiments (depicted in Fig. 4), we considered $T_1 + T_2 + T_3 + T_4$ features and picked up 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 to the experimental design.

The best accuracy is 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 are 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 be 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 the 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 (Gamon, 2004)).

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

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

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

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

Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: Proc. of the 1993 ACM international conference on Management of Data (SIGMOD). Washington, D.C., United States; 1993. p. 207–216.

Baayen RH, Van Halteren H, Tweedie FJ. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing 1996;2:110–20.

Burrows JF. Word patterns and story shapes: the statistical analysis of narrative style. Literary and Linguistic Computing 1987;2:61–7.

Cohen WW. Enron email dataset; 2004.

Corney M, de Vel O, Anderson A, Mohay G. Gender-preferential text mining of e-mail discourse. In: Proc. of the 18th Annual computer security applications conference; 2002. p. 282.

de Vel O. Mining e-mail authorship. In: Proc. of the Workshop on text mining in ACM international conference on knowledge discovery and data mining (KDD); 2000.

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

Dempster P, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 1977;39(1):1–38.

Fung CM, Wang K, Ester M. Hierarchical document clustering using frequent itemsets. In: Proc. of the 3rd SIAM international conference on data mining (SDM). San Francisco, CA; May 2003. p. 59–70.

Gamon M. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. of the 20th international conference on computational linguistics. Geneva, Switzerland; 2004. p. 611.

Holmes DI. The evolution of stylometry in humanities. Literary and Linguistic Computing 1998;13(3):111–7.

Holmes DI, Forsyth RS. The federalist revisited: new directions in authorship attribution. Literary and Linguistic Computing 1995;10(2):111–27.

Iqbal F, Hadjidj R, Fung BCM, Debbabi M. A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digital Investigation 2008;5:42–51.

Koppel M, Schler J, Argamon S. Computational methods in authorship attribution. Journal of American Society Information Science Technology 2009;60(1):9–26.

Ledger GR, Merriam TVN. Shakespeare, Fletcher, and the two Noble Kinsmen. Literary and Linguistic Computing 1994;9: 235–48.

Li H, Shen D, Zhang B, Chen Z, Yang Q. Adding semantics to email clustering. In: Proc. of the 6th international conference on data mining (ICDM). Washington, DC, USA: IEEE Computer Society; 2006a. p. 938–42.

Li J, Zheng R, Chen H. From fingerprint to writeprint. Communications of the ACM 2006b;49(4):76–82.

Mosteller F, Wallace DL. Inference and disputed authorship: the federalist. In: behavioral science:quantitative methods edition. Massachusetts: Addison-Wesley; 1964.

Novak J, Raghavan P, Tomkins A. Anti-aliasing on the web. In: Proc. of the 13th international conference on world wide web (WWW); 2004. p. 30–39.

Paice D. Another stemmer. SIGIR Forum 1990;24(3):56–61.

Porter MF. An algorithm for suffix stripping. Program October 1980;3(14):130–7.

Stamatatos E, Kokkinakis G, Fakotakis N. Automatic text categorization in terms of genre and author. Computational Linguistics 2000;26(4):471–95.

Stolfo SJ, Creamer G, Hershkop S. A temporal based forensic analysis of electronic communication. In: Proc. of the 2006 international conference on Digital Government Research. San Diego, CA, 2006. p. 23–24.

Yule GU. On sentence length as a statistical characteristic of style in prose. Biometrika 1938;30:363–90.

Yule GU. The statistical study of literary vocabulary. Cambridge, UK: Cambridge University Press; 1944.

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

Zheng R, Qin Y, Huang Z, Chen H. Authorship analysis in cybercrime investigation. Presentation. Retrieved February 07, 2009, http://www.isiconference.org/2003/presentation.htm; 2003a.

Zheng R, Qin Y, Huang Z, Chen H. Authorship analysis in cybercrime investigation. In: Proc. of the 1st International symposium on Intelligence and security informatics (ISI). Tucson, Arizona; 2003.

Please cite this article in press as: Iqbal F, et al., Mining writeprints from anonymous e-mails for forensic investigation, Digit. Investig. (2010), doi:10.1016/j.diin.2010.03.003