Journal of Computing and Security

October 2014, Volume 1, Number 4 (pp. 319–327)

http://www.jcomsec.org

A Novel Solution for Author Attribution Problem in Anonymous E-mail

Farahnaz Rezaeian Zadeh [a,*]
Shohreh Ajoudanian [b]

[a] *Department of Electronic, Computer and Information Technology engineering, Foolad Institute of Technology, Fooladshahr, Isfahan 8491663763, Iran.*
[b] *Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan, Iran.*

ARTICLE INFO.

Article history:
Received: 30 January 2014
Revised: 11 October 2014
Accepted: 08 November 2014
Published Online: 27 February 2015

Keywords:
Anonymous, Cybercriminals, Authorship Identification, Writeprint, Frequent Pattern, Similarity Measure.

ABSTRACT

Due to increasing criminal activities by anonymous E-mails in the cyber world, it is a challenging task to extract beneficial knowledge from E-mail systems. This problem in cyber world attracts many researchers in cyber-crime domain. Recent studies in this area concentrate on traditional classification approaches such as Decision Tree and Support Vector Machines (SVM). These approaches are employed to identify the author. The main goal of these researches is increasing the accuracy of identification, but the quality of evidence is ignored and also it is hard to be traced. So, in this paper, we propose a new approach based on data mining methods for improving the quality of evidence which leads to boost the accuracy of identification. We use writeprints as the evidence and extract them from each E-mail of individuals. The next step for author identification, is matching the writeprints with anonymous E-mails by applying Earth Mover Distance (EMD) criterion to identify the plausible author. In addition to high accuracy, EMD can help cybercrime investigators in making decision about anonymous intruder. Experiments with real data in both English and Persian languages, demonstrate the proposed approach can effectively identify the author and capture strong evidence to prove the identification.

1 Introduction

Nowadays, e-mail is a common way for textual communications on the web environment. Every day, millions of business letters, financial transactions and friendship messages are exchanged through e-mail systems. So the criminal activities incident by means of e-mail shouldn't be ignored [1]. Some examples of these malicious activities are spamming, phishing, threatening, e-mail bombing and racial vilification. E-mail is also used as a safe channel for some groups of terrorists and criminals to contact each other. The criminals do their suspicious activities with an unknown identity, for example in phishing they attempt to acquire sensitive information such as usernames, passwords, and credit card details (and sometimes, indirectly money) by masquerading as a trustworthy entity in an electronic communication.

Tree attributes of e-mail systems, make them prone to be applied as tools for illegal activities. The first attribute is related to anonymous servers, which assist to route an e-mail and hide the right information

* Corresponding author.
Email addresses: fr.rezaeeian@gmail.com (F. Rezaeian Zadeh), shajoudanian@pco.iaun.ac.ir (Sh. Ajoudanian)
ISSN: 2322-4460

about its origin. The Second attribute is the capability of e-mail for transferring executive files, hyperlinks, trojan horses and scripts. The last attribute is availability of internet and e-mail services in public places which makes anonymous problem more complicated. In recent years, several approaches have been used to prevent the abuse of this communication channel, but these methods are neither enough nor satisfactory. The criminal analysis of the e-mail by considering the authorship attribute can be useful for identifying the guilty author who is the owner of illegal e-mails.

The problem of authorship attribution in cybercrime domain can be defined as follows: An investigator wants to determine the author of an anonymous e-mail φ among a set of suspect $S=\{S_1,\ldots,S_n\}$. In other word, the problem is finding a plausible author from a group of suspected authors and also gathering firm evidence to confirm real culprit identity. In forensic science each person is distinguished by his/her fingerprint; this can be extended to the cybercrime world to reveal the anonymous guilty from his/her writing style. An investigator who works on cyber-forensic records could extract the writing style (writeprint) of each suspect from his/her e-mails and apply it to identify the real author. Note that there is no claim for uniqueness of a writeprint among all people as fingerprint, but we prove this solitary through a group of suspected authors.

The writeprint of an individual is the combinations of the features that repeat frequently in his/her written e-mails. The attributes that usually are used in this context are lexical, syntactical and structural features. The real author can be recognized by matching a writeprint with malicious e-mails. A writeprint is valuable in cyber forensic context as it can provide firm evidence to prove and support the identification results. There are several surveys in this context which focus on stylistic and structural features separately [1–3] while a few of them have studied the combination of the features to form a writeprint.

The techniques that have been widely used for author attribution problem can be divided in two main categories. The similarity based approach use some metric to measure the difference between two documents. An author whom his documents have the most similarity with the target document is known as the plausible author. The main contributions in the similarity-based approach are related to design the best presentation for document feature, the strategy for dimensionality reduction of the feature space and the selection of a similarity measure. Some of the most popular measures used in this context includes Burrows's Delta [4] and Delta [5], which most of these techniques suffer from the processing time. The second category is the machine learning technique. In this approach a training set is used (which is constricted from known writings of each candidate author) to create a classifier. Decision Tree [6] is one of the most popular machine learning techniques in which each decision node is constructed by considering only the local information on a feature. This is the main problem of this technique which makes it impossible to use multiple features concurrently, therefore the results could not be as accurate as required in cyber forensic context. The other approach which is widely used to address the authorship attribution problem [7] is Support Vector Machine (SVM) method. Although SVM has proven its ability to provide accurate recognition, it operates like a black box and the way to obtain the result is inexplicit, therefore it is almost impossible to trace the evidence to prove the recognition. In other words, SVM captures the input (the malicious e-mails and the e-mail of suspected author) and generates the output (the plausible author of malicious e-mails), but the process of generating output from input is hidden from the user. Therefore, due to these specifications, we can conclude that SVM doesn't have enough competence to assist cyber forensic investigators in tracing the anonymous author and isn't also appropriate in this context where collecting credible evidence is considered as a major objective. The other technique which is recently used to solve the authorship attribution is data mining technique. The accuracy of these techniques is approximately equal to SVM, but the main advantage of them is traceability which makes it easier to deal with cyber forensic issues. The main deficiencies of the studies which applied data mining approaches [8–10] to solve author attribution problem are pattern matching and similarity measure. These parameters (pattern matching and similarity measure) in the context of cyber criminal research help the investigators making better and stronger decision. To our knowledge, there is a lack of a good technique to cover these shortcomings. Therefore, in this paper, we propose a new approach to address the deficiencies of current algorithms which are combinations of machine learning and similarity based techniques. This approach is comprised of two main steps. At first, a unique writeprint is extracted for each author using Data Mining methods. Frequent pattern techniques [11] assist to provide an accurate writeprint model. The combination of multiple features appearing frequently in the suspect's e-mails is extracted as a writeprint. Frequent pattern mining is proven as a successful method to find hidden patterns in DNA sequences, customer purchasing habits, security intrusion and the other applications of pattern recognition. The second step of the proposed approach is to compare the writeprints of each suspected author with malicious e-mail's pattern by means of a simi-

larity measure called Erath' Mover Distances (EMD) [12]. This measure has been successfully used in a large number of similarity search settings including image retrieval [12, 13], transportation problem [14], document retrieval [15] and many more. EMD also has many applications in Machine Vision [16]. By relying on our survey; this is the first use of EMD in the authorship attribution context which EMD presents good results in determining the similarity and identifying the anonymous author. The proposed approach can help the cyber forensic investigators to have an accurate analysis of results and also to trace the evidence (writeprint).

The rest of this paper is organized as follows. In Section 2 the proposed approach is described. The experimental results to evaluate the proposed algorithm are discussed in Section 3. Concluding result is presented in Section 4.

2 Proposed Algorithm

In this section, we first present authorship attribution problem. A set of stylometric attributes is then described based on lexical, structural, content specific features. Finally, the proposed algorithm is introduced in two phases involved extracting writeprints and pattern matching in order to identify the real author.

2.1 Authorship attribution problem

Let $S=\{s_1,\ldots,s_n\}$ be a set of the suspected author of malicious e-mail φ. For each author s_i considers m messages as $E_i=\{e_1,\ldots,e_m\}$. The main objective of the authorship attribution problem is to find the most plausible author S_p which has the maximum similarity with the patterns belonging to malicious e-mail φ. In other words, a collection of e-mails in φ is matched with φ, if they share similar patterns of vocabulary usage, structural and stylometric features. A set of stylometric features which extracted for each suspected author is called writeprint and is used to present evidence in the criminal courts. These writeprints are generated from the frequent features FP_i repeated in the e-mails contained in E_i and then the pattern of φ is matched with all writeprints using EMD similarity measure.

2.2 Stylometric features

Written patterns of each individual can be defined with word usage, word arrangement, misspelling and grammatical mistakes. Chen and Abbasi widely investigated on the structural and stylistic features [14, 17]. In this study by surveying the previous works, we choose a set of stylometric features which are proper and also most effective for forensic analysis of anonymous e-mail. These features have been mostly applied to English text, but for the first time in this paper, we use these features in Persian language, by adopting them to be applicable in this language. The stylometric features chosen for solving authorship attribution problem, are grouped into three categories: lexical, structural and content specific features. The details of these features are described as following:

Lexical features are divided into alphabetic based and word based characteristic. Alphabetic based features contain frequency of individual alphabet (including English and Persian alphabets), number of alphabets appearing in each word, the incident rate of capital and small alphabets (just for English language) and the number of alphabets in each sentence. The most significant word based features include the count of the words in a sentence and distribution of word length. The features are considered for syntactic attributes including function words (auxiliary verb, preposition, conjunction, and pronoun). These features are defined separately for each language. The other feature used as syntactic attributes is punctuations which has the effective role in authorship attribution. Structural features are used to evaluate the layout of written texts include the average of the paragraph's length, number of paragraphs, present/absence of greeting and the position of them. The last feature is referred to content specific attribute which contains a collection of keywords in a certain context and it may be different for each person in various domains (friendly, official, academic and etc.) Zheng et.al [2] applied eleven keywords in cybercrime taxonomy for authorship attribution domain and this paper has followed this taxonomy.

2.3 Writeprint extraction

In order to extract writeprint of each individual, we need to detect stylometric features in each e-mail e_i and then normalize and discretize them in order to present them in vector form. After that we apply frequent pattern mining algorithm and filter common pattern to obtain a unique writeprint for each author S_i from their e-mail set (e_i). In following, the details of writeprint extraction process is described in three phases: feature discretization, frequent pattern extraction, unique pattern extraction.

2.3.1 Feature discretization

Let E_i be a set of e-mails written by suspected author $S_i \in \{S_1,\ldots,S_m\}$. At first, the stylometric features are extracted from each e-mail. Note that in following section when using feature word, we refer to all features described in Section 2.2. The normalization

process is applied to each feature value and then the values are discretized in a certain interval. For example, these intervals can be [0.00-0.25], [0.25-0.50], [0.50-0.75], [0.75-1.00]. Each interval is called a feature item. The normalized feature frequency is then matched with these intervals. Then assign value 1 to the feature item if the interval contains the normalized feature frequency; otherwise assign value 0. The most common techniques used for discretization include Equal-width, Equal frequency and Clustering Based Discretization. The proper technique for discretization of values based upon the feature attribute and values distribution in author attribution problem is cluster based methods. The effectiveness of this method was proved by comprehensive experiments which were done for the values with these specifications [18–20]. In this problem Attribute Distribution Clustering Orthogonality (ADCO) measure is chosen which cluster the features based on the information distribution and density [20]. The main reason for choosing this measure is the power of ADCO to prepare efficient data and prevent removing or ignoring essential values to construct an accurate writeprint. An example of discretizing feature is described as follows:

Suppose that A ,B and$C$ are the three features which are extracted from 5 e-mails. The value of feature items are normalized in [0 1] range. Then By applying the discretization methods the following results are obtained for each feature:A is divided into 4 feature item A_1=[0.00-0.25], A_2=[0.25-0.50], A_3=[0.75-0.50], A_4=[0.75-1.00];B contains 2 feature items B_1=[0.00-0.66], B_2=[0.66-1]; the feature items for$C$ are C_1=[0.00-0.75], C_2=[0.75-1.00]. After the intervals are calculated, it is the time for determining the feature values for ϵ_i included in each e-mail E_i, for example these value for ϵ_1 corresponds to A=0.34, B=0.12, C=0.50 and its vector is presented by <0,1,0,0,1,0,0,1,0> (Table 1). Note that, another vector is required to save the real values of each feature item for using in pattern matching phase.

2.3.2 Frequent pattern extraction

Written pattern of a set of e-mails E_i (written by author E_i) is a combination of feature items which frequently occur in S_i. In order to accurately detect and model frequent pattern used approach presented by Agrawal et.al [11]. In following, by means of an example, describe the details of frequent pattern mining algorithm which used for pattern extraction.

Suppose that each e-mail ϵ in E_i is defined by a set of feature items and U=$\{f_1,\ldots,f_m\}$ and $\epsilon \subseteq$U. an e-mail ϵ contains a feature item f_i if the numerical feature value of the e-mail ϵ falls within the interval of f_i. For example e-mail ϵ_1 is a collection of feature items which are presented by ϵ_1=A2, B1, C1 as shown in the Table 1.

Table 1. Feature vector

e-mails	Feature A				Feature B		Feature C	
	A1	A2	A3	A4	B1	B2	C1	C2
ϵ_1	0	1	0	0	1	0	1	0
ϵ_2	0	1	0	0	1	0	1	0
ϵ_3	0	1	0	0	1	0	1	0
ϵ_4	1	0	0	0	1	0	1	0
ϵ_5	0	0	0	1	1	0	1	0

Let $F \subseteq U$ be a pattern which contains a set of feature items. Each e-mail pattern ϵ contains F, if $F \subseteq \epsilon$. A pattern containing k feature items, is called a k-pattern. The support value for pattern F is defined by the percentage of e-mails which contain pattern F. The pattern F is considered as a frequent pattern, if and only if the support value for F is greater than or equal to minimum threshold support (MTS). The value of MTS is defined by the user based upon the needs of the problem.

Frequent pattern definition: Suppose that E_i is a collection of e-mails written by suspected author S_i and $support(F \mid E_i)$ represents the percent of e-mail in E_i which contain the pattern $F(F \subseteq U)$. The pattern F is a frequent pattern, if $support(F \subseteq E_i) \geq min\_sup$. MST (min_sup) is a real value in range of [0, 1]. Therefore, the stylistic pattern for suspected author S_i is composed of a set of frequent pattern presented by fp_i=$F_1, \ldots, F_k$.

The popular Data Mining algorithms used for obtaining frequent pattern include Apriori [21], FP-growth [22] and ECLAT. In this paper, the process of frequent pattern mining is done by Apriori algorithm. The performance of the Apriori algorithm for detecting frequent pattern is proven by comprehensive survey which is done in this context [8–10].

Apriori is a level-wise iterative search base algorithm which uses a frequent k-pattern to explore frequent (k+1)pattern [23]. In order to apply Apriori algorithm, at first all frequent 1-pattern is detected through a set of E_i and then by means of the output constructed in this level, the frequent 2-patterns are explored. This process is continued until frequent k-patterns are detected.

Definition of Apriori algorithm: ″All nonempty subsets of a frequent pattern must also be frequent ″ [23]. With regard to this definition, F' isn't considered

as a frequent pattern, if $support(F' \mid E_i) \leq min\_sup$. This property refers to that f_i would never be a frequent pattern if it was added to infrequent pattern F'. Therefore, it isn't required to generate (k+1)pattern from k-pattern F' because it is considered as an infrequent pattern.

2.3.3 Generate writeprint

When extracting frequent pattern, several suspected authors may share common patterns which violate uniqueness of the pattern. So to create unique patterns which are called the writeprints for each individual, we need to remove common patterns.

Writeprint definition: Writeprint WP_i is a collection of patterns in which each pattern F satisfies the MST condition, i.e. $support(F \mid E_i) \leq min\_sup$, $support(F \mid E_j) \leq min\_sup$ and no pairs of author's writeprints share the same patterns ($i \neq j$). The model of writeprint WP_i for author S_i can be formulated as $WP_i \subseteq FP_i$, $WP_i \cap WP_j = \emptyset$ if $j \leq n$, $1 \leq i$ and $j \leq n$. FP_i indicates all of the frequent patterns for author S_i. So WP_i can be defined as a unique writeprint which is extracted from a set of e-mail E_i for author S_i.

2.4 Identifying anonymous author

In order to accomplish the last phase of author attribution problem, the algorithm compares the writeprints of each individual with the pattern of the malicious e-mail φ (which is sent to the victim). So in this phase, the degree of similarity between captured evidence and the pattern of φ is detected and then the last decision about the most plausible author is made. Note that to apply EMD, we use the real values of feature items included in writeprint. In the following, we describe EMD measure to calculate the similarity of patterns.

2.4.1 Calculate similarity of different patterns

Two different patterns are compared using EMD to determine the degree of similarity between them. EMD is a distance measure which can compare the patterns with various frequencies. In fact, EMD calculates the minimum cost required to transform one pattern to the other. Let $WP[S_\Upsilon]=\{(\mu_{fp_1},\Sigma_{fp_1}, \omega_{fp_1}), \ldots, (\mu_{fp_s},\Sigma_{fp_s}, \omega_{fp_s})\}$ be a collection with s frequent patterns where $(\mu_{fp_i}, \Sigma_{fp_i}$ and $\omega_{fp_i})$indicate mean, covariance and weight of the frequent patterns involved in writeprint $WP[S_\Upsilon]$ for suspected author S_Υ. Similarly, let $P_\varphi=\{(\mu_{p_1},\Sigma_{p_1},\omega_{p_1}), \ldots, (\mu_{fp_t},\Sigma_{p_t},\omega_{p_t})\}$ be the property of *it* patterns in the malicious e-mail P_φ. Suppose that $d_{fp_i,p_j})$ is the distance between pattern fp_i in writeprint and p_j which is the pattern of malicious mail P_φ and is calculate by the following equation:

$$d_{fp_i,p_j} = \frac{\sum fp_i}{\sum p_j} + \frac{\sum p_j}{\sum fp_i} + (\mu_{fp_i} + \mu_{p_j})^2 \cdot \left(\frac{1}{\sum fp_i} + \frac{1}{\sum p_j}\right) \quad (1)$$

where f_{fp_i,p_j} represents the flow between fp_i and p_j. This flow indicates the cost of moving probability mass (analogous to piles of earth) from one pattern to another. In proposed approach the flow is calculated based on equation (2).

$$f_{fp_i,p_j} = \frac{1}{\sup(fp_i \mid \varphi) \bullet \sup(p_j \mid WP[S_\gamma])} \quad (2)$$

$sup(fp_i|\varphi)$ determines the support degree of pattern fp_i in malicious e-mail φ). The support degree of pattern p_j in proportion to the writeprint $WP[S_\Upsilon]$ is calculated by $sup(p_j|WP[S_\gamma])$. The distance measure between patterns and the cost of transformation obtained by EMD simplify the process of computing the similarity degree. If the distance between patterns which is shown by d_fp_i,p_j, defines the neighborhood of the feature values and f_fp_i,p_j indicates the cost of transformation with regard to the concept of MST, EMD measure can be calculated based upon these two parameters. A low degree of support value shows the higher cost of transformation which leads to lower similarity between patterns. The EMD measure is calculated by the following equation:

$$EMD(WP[S_\gamma], P_\varphi) = \frac{\sum_{i=1}^{s}\sum_{j=1}^{t} d_{fp_i,p_j} f_{fp_i,p_j}}{f_{fp_i,p_j}} \quad (3)$$

3 Experimental Evaluation

In this section, we conduct an experimental study to evaluate the accuracy of proposed method, prove the uniqueness of the writeprint, show the strength of detected evidence to support the result and analysis the efficiency of the evidence to present in the forensic court. Since the proposed method consists of two main phases in two different contexts (data mining and pattern matching), we decided to implement the approach by means of Mweka Tool. This GUI runs the Weka classifiers and displays the results in MATLAB. This tool facilitates the use of data mining algorithm in proposed approach with pattern matching method. In the following, we apply the proposed method in English and Persian languages in two sections.

Note that in all experiments the accuracy is defined as the proportion of e-mails whose author was correctly identified the total number of messages.

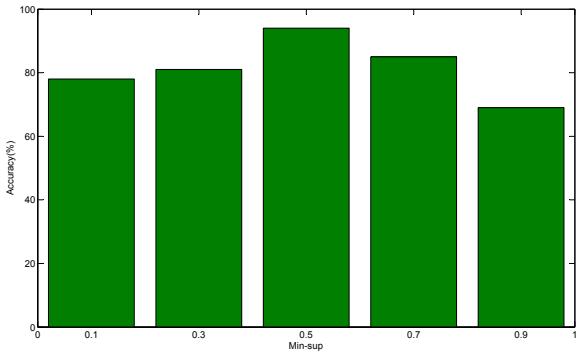

Figure 1. Accuracy of proposed method across various min_sup (m=15, n=10)

3.1 English Language

In order to design the experiments, we used m e-mails for n employees which were randomly selected from the Enron e-mail Dataset. These n employees are considered as suspected authors in this problem which is denoted by S= $\{S_1, \ldots, S_n\}$. For each suspected author S_i, m e-mails E_i=$\{e_1, \ldots, e_m\}$ were selected, from 1\3 is applied for testing and 2\3 is allocated for learning. Let E=$\{E_1, \ldots, E_n\}$ be a set of e-mails belonging to n suspected author.

In order to analysis the influence of MST, the proposed algorithm is examined with different values for *min_sup*. As shown in Figure 1, the accuracy of proposed method is kept robust for various *min_sup*. The number of authors selected for this assessment is 10 and the number of e-mails for each author is 15 (i.e., a total of 150 e-mails).

Since this domain is in infancy, there aren't too many algorithms which are applied for this context. In order to show the efficiency of the proposed method in contrast to the other algorithms, we choose an approach which is closer to our proposed method and also have a good performance in compare to the other methods such as END [24], J48 [6], RBFNetwork [25], NaiveBays [26] and BaysNet [27]. This algorithm is called AuthorMiner algorithm which is applied in [8]. This algorithm is considered as one of the few methods which use data mining technique to solve author attribution problem in cyber-forensic domain. As shown in Figure 2, the accuracy of proposed method is higher than the AuthorMiner for different number of suspected authors. As shown in Figure 2, by increasing the number suspected authors, the accuracy of the proposed method doesn't have considerable reduction. The influence of various number of the authors in accuracy, demonstrate the stability of the proposed algorithm, since the degree of accuracy reduction is only 19% by increasing the number of authors to 16.

In the cyber forensic domain the efficient data will have more effect on the quality of evidence as well

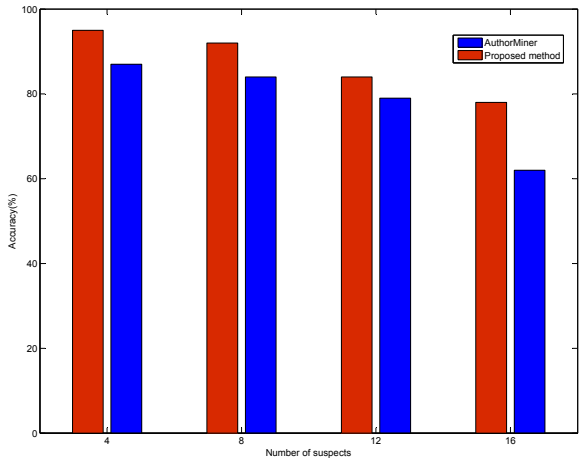

Figure 2. Comparison of proposed method with AuthorMiner Algorithm [6] (m=15).

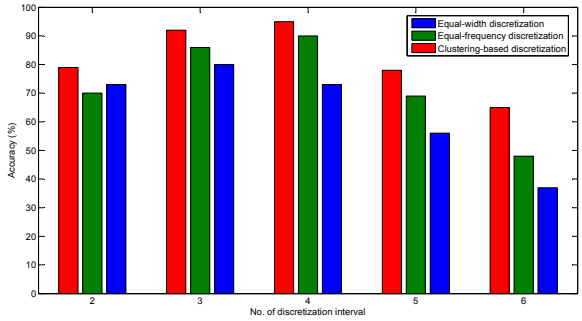

Figure 3. Accuracy of different discretization method in proposed algorithm

as accuracy of identification. One of the important phases which could prepare the suitable data is the discretization process and determining the disjoint point for clustering data. As depicted in Figure 3, three kinds of discretization methods are compared. The clustering based method, which used the ADCO measure for clustering data, presents the higher accuracy than Equal-width and Equal frequency discretization based algorithm.

The EMD measure is applied in proposed algorithm to identify the anonymous author among a group of suspected authors. One of the advantages of EMD measure is that we can use EMD as a measure to evaluate the quality of writeprints in addition to its strength for detecting the plausible author based upon the most similarity degree. The results of the experiment are shown in Table 2. The short distance between writeprint of each individual and their e-mail's patterns show the quality of writeprint and also prove the correctness of the proposed method. The long distance between patterns of the various authors shows the uniqueness of captured writeprints. As shown in Table 2, the short distance reflects the lower value which means the higher degree of similarity and the long distance refers to lower similarity. These results

Table 2. Similarity degree of patterns

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
WP_1	**0.20**	0.64	0.74	0.59	0.75	0.67	0.86	0.64	0.77	0.74
WP_2	0.68	**0.17**	0.69	0.75	0.81	0.59	0.65	0.76	0.74	0.56
WP_3	0.74	0.75	**0.21**	0.83	0.69	0.55	0.78	0.83	0.84	0.69
WP_4	0.89	0.87	0.75	**0.19**	0.73	0.61	0.65	0.61	0.62	0.73
WP_5	0.59	0.68	0.78	0.74	**0.15**	0.57	0.73	0.76	0.70	0.68
WP_6	0.85	0.89	0.57	0.91	0.78	**0.22**	0.79	0.84	0.72	0.85
WP_7	0.87	0.65	0.65	0.71	0.84	0.67	**0.23**	0.90	0.79	0.83
WP_8	0.79	0.64	0.78	0.64	0.71	0.58	0.81	**0.14**	0.74	0.86
WP_9	0.67	0.76	0.89	0.81	0.63	0.61	0.75	0.78	**0.20**	0.73
WP_{10}	0.78	0.84	0.74	0.64	0.77	0.68	0.81	0.62	0.69	**0.24**

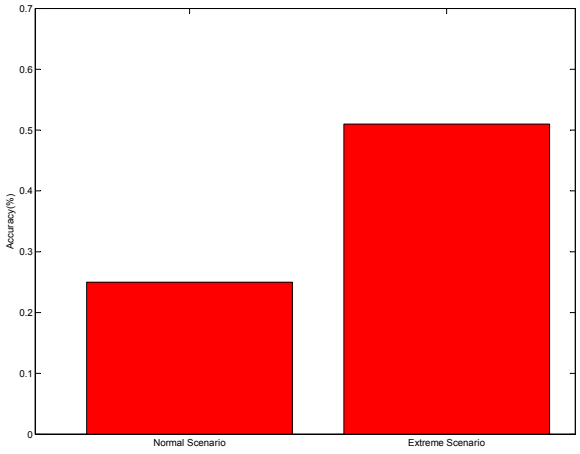

Figure 4. The error rate for Normal (m=40,n =100) and extreme (m=10,n =100) scenario

facilitate the decision making for investigator to detect the most plausible author and present a clear view of the captured evidence with all property by tracing desired evidence. Therefore, we can conclude that the proposed approach reflects a good performance in cyber forensic in order to identify the anonymous author with strong evidence.

In order to evaluate the proposed method in an extreme scenario, consider too many authors with a limited number of sample e-mails for each of them. In this scenario, we want to show how the accuracy is affected by means of the error rate of the author attribution. Although in this situation the probability of generating the write-prints with high similarity is increased, but this method is accurate enough to deal with this challenge. As shown in Figure 4, in extreme scenario the increasing of the error rate is not considerable in contrast to the normal scenario. Therefore, the proposed method can provide reliable attribution when the enough e-mails of the suspects aren't available.

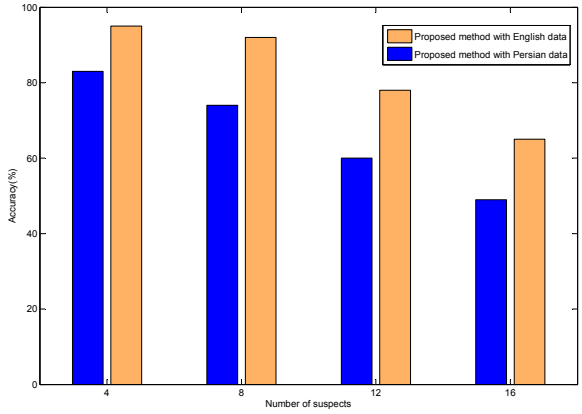

Figure 5. Accuracy of the proposed method in Persian and English Languages

The total runtime is dominated by the first phase for writeprint extraction process and the whole process is also done in less than 10 minutes.

3.2 Persian Language

To evaluate the proposed algorithm in Persian language, we adopt the stylometric feature based on this language and use personal e-mail to test it. These e-mails contain semi-official content such as the e-mails that have been sent to the professor. The framework of these experiments is designed like English (as mentioned in previous section). By comparing the accuracy of proposed algorithm in Persian and English language it is shown that the accuracy of identification in Persian is lower than in English but it's not so considerable (Figure 5). The experimental results in Persian language reflect **good** efficiency of proposed algorithm. It is required to analysis the specialized features that is related to Persian language since the features which is used in this paper is the same as English and we only adopt them to be usable in Persian language. The total runtime for Persian language is less than 14 minutes.

4 Conclusion

In this paper, we proposed a new approach for solving authorship attribution to identify anonymous intruder in cyber-forensic context. This approach is composed of two main phases: writeprint extraction and pattern matching. The features used to create the writeprints of each individual include lexical, structural and content specific features. The extraction phase is started by normalizing and discretizing the features by means of ADCO measure for clustering. Then, the frequent patterns are extracted by Apriori algorithm. After that, the common patterns are removed to capture a unique writeprint for each suspected author. The

second phase is pattern matching by EMD measure. This measure is applied to identify the most plausible author who has a writeprint with highest similarity with the pattern of malicious e-mail. The EMD measure calculates the distance between two patterns in order to determine the similarity degree. This method provides the flexibility to detect distinction between similar patterns and reduces the errors in identifying the most plausible author.

The experimental results show the satisfactory performance in (terms of accuracy and the quality) of proposed algorithm with real life data. The proposed algorithm is evaluated various aspects such as accuracy in contrast to the other algorithms, accuracy of the different number of suspected authors and accuracy of the algorithm with various *min_sup* .The results of the experiments demonstrate the stability and robustness of this method. By conducting the experiments in English and Persian language, the results indicate the higher accuracy of algorithm in English language and also a good result for Persian language as it is the first attempt for this language. We hope to improve the current shortcoming in this language in the future. To detect effective features for generating high quality Persian writeprints, a comprehensive study regarding stylometric features is required.

The information provided by this approach help cybercrime investigators to make decision based on the strong and traceable evidence. Since these processes for finding beneficial information couldn't be done manually, the application of these methods creates a new chance for detecting intruders in cyber world and smoothes the challenges to solve anonymity problem.

References

[1] Gui-fa Teng, Mao-sheng Lai, Jian-Bin Ma, and Ying Li. E-mail authorship mining based on svm for computer forensic. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 2, pages 1204–1207. IEEE, 2004.

[2] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan 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):378–393, 2006.

[3] Olivier De Vel. Mining e-mail authorship. In *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD2000)*, 2000.

[4] Shlomo Argamon. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.

[5] John Burrows. delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

[6] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[7] Colin Campbell and Yiming Ying. Learning with support vector machines. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5 (1):1–95, 2011.

[8] Farkhund Iqbal, Rachid Hadjidj, Benjamin Fung, and Mourad Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *digital investigation*, 5:S42–S51, 2008.

[9] Farkhund Iqbal, Hamad Binsalleeh, Benjamin Fung, and Mourad Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation*, 7(1):56–64, 2010.

[10] Farkhund Iqbal, Hamad Binsalleeh, Benjamin Fung, and Mourad Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112, 2013.

[11] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[12] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[13] Yossi Rubner, Leonidas J Guibas, and Carlo Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA image understanding workshop*, pages 661–668, 1997.

[14] Frank L Hitchcock. The distribution of a product from several sources to numerous localities. *J. Math. Phys*, 20(2):224–230, 1941.

[15] Xiaojun Wan. A novel document similarity measure based on earth movers distance. *Information Sciences*, 177(18):3718–3730, 2007.

[16] Beth Logan and Ariel Salomon. Music similarity function based on signal analysis, 2001.

[17] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008.

[18] Ankit Gupta, Kishan G Mehrotra, and Chilukuri Mohan. A clustering-based discretization for supervised learning. *Statistics & probability letters*, 80(9):816–824, 2010.

[19] Qiju Kang, Ying Qian, Lijuan Sun, Hai Yu, and Jianyu Wang. Two-phase spectral clustering

based on discretization. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 5th International Conference on*, volume 1, pages 245–249. IEEE, 2013.

[20] Eric Bae, James Bailey, and Guozhu Dong. A clustering comparison measure using density profiles and its application to the discovery of alternate clusterings. *Data Mining and Knowledge Discovery*, 21(3):427–471, 2010.

[21] Christian Borgelt and Rudolf Kruse. Induction of association rules: Apriori implementation. In *Compstat*, pages 395–400. Springer, 2002.

[22] Jian Pei. *Pattern-growth methods for frequent pattern mining*. PhD thesis, Citeseer, 2002.

[23] Mehmed Kantardzic. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[24] Eibe Frank and Stefan Kramer. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the twenty-first international conference on Machine learning*, page 39. ACM, 2004.

[25] Martin Dietrich Buhmann, Martin Dietrich Buhmann, and Martin Dietrich Buhmann. *Radial basis functions: theory and implementations*, volume 5. Cambridge university press Cambridge, 2003.

[26] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3): 129–146, 1976.

[27] Judea Pearl. Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. 1985.

Farahnaz Rezaeian Zadeh received her B.Sc degree in software engineering from the Hakim Sabzevari University in 2010, with the thesis of Design and Implementation of E-Voting System. She received her M.Sc degree in Information Technology (IT) Engineering from the Foolad Institute of Technology, Foolad shahr, Esfahan, Iran in 2014 with the thesis of Multi objective Optimization Model For Deployment Configuration of Service Based Application In Cloud Platform". She has already authored more than 10 international conference papers.

Shohreh Ajoudanian is Ph.D. Candidate in the department of Computer Engineering, Islamic Azad University Science and Research Branch, Tehran, Iran. Currently, she is faulty member of Islamic Azad University of Najafabad, Isfahan, Iran. She received her M.Sc degree in 2008 from Islamic Azad University of Najafabad, Isfahan, Iran. Her research interests include formal method, software product line engineering.