A novel solution for author attribution problem in anonymous E-mail

Farahnaz Rezaeian Zadeh
Information Technology Department
Foolad Institute of Technology
Foolad shahr, Esfahan, Iran
fr.rezaeeian@gmail.com

Shohreh Ajoudanian
Computer Engineering Department
Islamic Azad University of Najafabad
Najafabad, Esfahan, Iran
shajoudanian@pco.iaun.ac.ir

***Abstract*—Todays by increasing misused by means of the E-mail system through cyber-word, extracting beneficial knowledge from suspected E-mail become more challenging and also draw more attraction by researcher cyber-crime. Recent studies in this area concentrate on traditional classification approaches such as Decision Tree and Support Vector Machines (SVM). These approaches employ to identify the author and the main aim of them is to increasing the accuracy of identification while the evidences quality is ignored. So, in this paper we propose a new approach based on data mining methods for improving the quality of evidences which lead to boost the accuracy of identification. We use writeprint as the evidence which extract from each mail of individuals. The last step for author identification is matching the writeprints with anonymous mails by applying Earth Mover Distance (EMD) criterion to identify the plausible author. In addition of advantages of EMD in high accuracy, it could help the cybercrime investigators to decision making on anonymous defender. Experiments on real data in both English and Persian language suggest the proposed approach can effectively identify the author and capture strong evidences to prove the identification.**

Keywords— Anonymously; cybercriminals; authorship identification; writeprint; frequent pattern

I. Introduction

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

In principal the E-mail system is prone to use as tools for illegal activities due to three attributes. First one related to anonymous server which assisted to route an E-mail, so the right information about its origin remain hidden. The Second attribute is the capability of E-mail for transporting executive files, Hyperlinks, Trojan horses and the script. The last characteristic is availability of internet and E-mail services in public places which make anonymous problem more complicated. In recent years, several approaches are used to prevent abuse of this communication channel, but these methods aren't enough and also satisfactory. The criminal analysis of the E-mail with regard to authorship attribution 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 through a collection of suspect $S=\{S_1,\ldots,S_n\}$. In the other word the problem is finding a plausible author from a group of suspicious authors and also gathering firm evidences to proof recognition of real guilty. In forensic science each person is distinct by his/her fingerprint and this statement can be extended in the cybercrime world for distinguishing anonymous guilty form his/her writing style. An investigator who works on cyber forensic records could extract the writing style of each suspect form their E-mails which called writeprint and apply it for identifying the real author. Note that there's no claim for uniqueness of a writeprint among all people as fingerprint but we proof this solitary through a group of suspicious authors.

The writeprint of an individual is the combinations of features that repeat frequently in his/her written E-mails. The attributes that usually use in this context are lexical, syntactical and structural features. The real author can be recognized by matching the writeprint with malicious E-mails. A writeprint is valuable in cyber forensic context when it can provide firm evidences to proof and support the identification result. There are several surveys which done in this context. These studies focus on the stylistic and structural features separately [1-3] while a few of them has studied the combination of feature to form a writeprint.

The most common techniques which applied in authorship attribution problem are the classification methods and in this section briefly discussed two important categories of these methods. The first one related to Decision Tree (c4.5) [4] in which each decision node is constructed by considering only the local information of a feature and using multiple feature concurrently is impossible in this method therefore the results could not be as accurate as need in cyber forensic context. The second method used Support Vector Machines (SVM) [5] to address the authorship attribution problem. Although SVM

proved its performance in accurate recognition, but it operates like a black box and the way how to obtain the result is inexplicit and it is almost impossible to trace the evidence to prove the recognition. In the other word SVM captures the input (the malicious E-mails and E-mail of suspicious author) and generates the output (the plausible author of malicious E-mails), so the operation in how generating output from input is hidden from the user. Therefore by these descriptions 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 evidences consider as a major objective. A new approach applied to solve the authorship attribution is related to data mining techniques which have the same accuracy as SVM but the main advantage is the traceable capability of these approaches that could satisfy cyber-forensic issues. The main deficiency of the studies applied data mining approaches [6-8] to solve author attribution problem is related to pattern matching and also similarity measure. These parameters in the context of cyber-criminal research help the investigators making better and also strong decision. To our knowledge, there is lack of a good technique to cover these shortcomings. Therefore, in this paper proposed a new approach to address the deficiency of current algorithms. This approach is implemented in two main steps. At first extracting a unique writeprint for each author by means of Data Mining methods and then identifying the plausible author by comparing the malicious E-mails with the writeprint pattern through similarity measure (EMD). The proposed approach can help the cyber forensic investigators to have an accurate analysis of result and also tracing evidences (writeprint).

Frequent pattern techniques [9] assist to providing an accurate writeprint model .The combination of multiple features appear frequently in the suspect E-mail is extracted as a writeprint. Frequent pattern mining proven to be successful in finding hidden patterns in DNA sequences, customer purchasing habits, security intrusion and the other applications of pattern recognition.

The second step of proposed approaches is to compare the writeprints of each suspect with malicious E-mail pattern by helping a similarity measure called Erath' mover distances [3]. This criterion first applied as an empirical method to measure the similarity of combination and image color [10, 11] and then used for solving transportation problem which should conduct the products with cheapest cost to reach the destination [12]. Determining similarity between different documents [13] and many applications of Machine Vision are considered as an application of EMD which proof its performance. By rely on our survey; this is the first use of EMD in the Authorship attribution context which by justifying with problem structure, this criterion presents good results in determining the similarity and recognition of the anonymous author.

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 present in section 4.

II. Proposed Algorithm

In this section at first presents authorship attribution problem and stylometric attributes and in the following the proposed algorithm is described in two phases involved extracting writeprint and pattern matching in order to identifying the real author.

A. Authorship attribution problem

Let $S=\{s_1,\ldots,s_n\}$ be a set of suspected author of malicious E-mail φ. For each author s_i considers m message as $E_i=\{e_1,\ldots,e_n\}$. The objective of authorship attribution problem is to find the most plausible author S_p which has the most similarity with pattern of malicious E-mail φ. In other word a collection of E-mail E_i is matched with φ, if they share similar patterns of vocabulary usage, structural and stylometric features. The primary goal is to precisely extract stylometric patterns for each suspect, which these patterns called writeprints which use as observed evidences the in criminal courts. The writeprints are generated from the frequent features FP_i repeated in the E-mails involved in E_i and then finding plausible author is done by matching the pattern of φ with all write prints WP_i by means of EMD similarity measure.

B. Stylometric feature

Writing patterns usually include word usage feature, word arrangement, misspelling and grammatical mistakes. Chen and Abbasi have widely investigated on the structural and stylistic analysis [14]. In this study by surveying the previous works, detect the stylometric features which are proper and also most effective for E-mail forensic analysis. These features almost only applied in English text, but for the first time in this paper used these features in Persian language of course by adopting them for application in this language. The stylometric features chosen for solving authorship attribution problem, are grouped into four categories: Lexical, structural, content specific feature. The details of these features describe in following:

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

C. Writeprint extraction

In order to extracting written patterns which called writeprint of each individual, it needs to detect stylometric features in each E-mail e_i and then normalizing and discretizing in which to presenting them in vector form. After that applying frequent pattern mining algorithm and filtering common pattern to obtain an unique writeprint for each author S_i from their E-mail set (E_i). In following describe the details of Writeprint extraction process through three phases: feature discretization, frequent pattern extraction, extract unique pattern.

1) Feature discretization

Suppose that E_i is a collection of E-mail written by suspected author $S_i \in \{S_1, \dots, S_n\}$. At first, the stylometric features are extracted from each E-mail. Note that in following section when use feature word, it referred to all features describe in section 2-2. Each feature has a certain value which applies normalization process and then discretizes the values 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 and for each interval contains feature value set 1 and the others are assigned with 0. The most common techniques are used for discretization are Equal-width, Equal frequency and clustering based discretization. The proper technique for discretization values in author attribution problem based upon the feature attribute and valued distribution is cluster based methods which the effectiveness of this method has proved by comprehensive experiments doing for these kinds of values. So the technique selected for discretizing stylometric features value is based on Attribute Distribution Clustering Orthogonality (ADCO) measure which clustering the features based on information distribution and density. The main reason for choosing this measure is the power of ADOC in prepare efficient data and preventing remove or ignore essential values to construct an accurate writeprint. An instance of discretizing feature is described in following:

Suppose that A, B, C are three features which extracted from 5 E-mails. After normalizing the value of feature items in [0 1] range, by applying the discretizing methods the following results 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 contain 2 feature items $B_1 = [0.00 - 0.33], B_2 = [0.33 - 0.66], B_3 = [0.66 - 1.00]$; the feature items for C are $C_1 = [0.00 - 0.75], A_2 = [0.75 - 1.00]$. After the intervals are determined, it is the time for specifying the feature value for ε_i included in each E-mail E_i, for example these value for ε_1 correspond to $A = 0.34, B = 0.12, C = 0.50$ and it's vector is presented by $< 0,1,0,0,1,0,0,1,0 >$ (Table 1). Note that there's need to another vector which save the real value of each feature item for using in pattern matching phase.

2) Frequent pattern extraction

Writing pattern of a set of E-mails E_i (written by author S_i) is a combination of feature items which frequently occurs in E_i. In order to accurately detecting and modeling frequent pattern used approach presented by Agrawal et.al [9]. In following by means of an example describing 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 item $U = \{f_1, \dots, f_m\}$ and $\varepsilon \subseteq U$. Each E-mail ε contain feature item f_i if the value of certain feature existed in the interval of f_i. For example E-mail ε_1 is a collection of feature item which presented by $\varepsilon_1 = \{A2, B1, C1\}$ 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 $F \subseteq U$ is a pattern contains a set of feature items. Each E-mail pattern ε contain *F*, if $F \subseteq \varepsilon$. A pattern contains *k* feature item is called a *k*-pattern. Support of pattern *F* determines the percent of E-mails involved *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 problem.

Frequent pattern definition: Suppose that E_i is the collection of E-mail written by suspected author S_i and $support(F|E_i)$ represents the percent of E-mail in E_i which contains the pattern *F* ($F \subseteq U$). The pattern *F* is a frequent pattern if $support(F|E_i) \geq min\_sup$ and MST (min_sup) is a real value in range of [0, 1]. Stylistic pattern for suspected author S_i is composed of a set of frequent pattern presented by $fp_i = \{F_1, \dots, F_k\}$.

The popular Data Mining algorithms used for obtaining frequent pattern include Apriori [2], FP-growth [15] and ECLAT. In this paper the progress of frequent pattern mining is done by Apriori algorithm which prove its performance in detecting frequent pattern by comprehensive survey which is done in this context.

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

Definition of Apriori algorithm: All nonempty subsets of a frequent pattern must also be frequent.

With regard to top definition, F' doesn't consider as a frequent pattern if $support(F'|E_i) \leq min\_sup$. By this property it can conclude that f_i never be a frequent pattern if it was added to infrequent pattern F'. Therefore it's not required to generate (*k+1*)pattern from *k*-pattern F' which consider as an infrequent pattern.

3) Generate writeprint

By extracting frequent pattern, several suspected authors may have shared common pattern which violate uniqueness of pattern. So to create unique pattern which called writeprint for each individual it needs to remove common pattern.

Writeprint definition: Writeprint WP_i is a collection of pattern in which each pattern F satisfying the MST condition $(support(F|E_i) \leq min_{sup},\ support(F|E_j) \leq \min\_sup)$ and any author's writeprint doesn't share same pattern $(i \neq j)$. The model of writeprint WP_i for author S_i can be formulated as $WP_i \subseteq FP_i, WP_i \cap WP_j = \emptyset\ if\ j \leq n, 1 \leq i\ and\ i \neq j$. FP_i indicates all frequent pattern and WP_i is an unique writeprint detected in set of E-mail E_i for author S_i.

D. Identifying anonymity author

In order to accomplish the author attribution problem for finding guilty author, comparing the writeprint of each individual with the pattern of malicious E-mail φ was sent to victim. So in this phase detects the degree of similarity between capture evidences and the pattern of φ and then make decision about plausible author. Note that to apply EMD, used real value of feature items included in writeprint. In following describe EMD measure to calculate similarity of pattern.

1) Calculate similarity of different pattern

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

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

Define f_{fp_i,p_j} as 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).

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

$support(fp_i|\varphi)$ determine the support degree of pattern fp_i in malicious mail φ. The support degree of pattern p_j respective to writeprint $WP[S_Y]$ is calculated by $support(fp_i|WP[S_Y])$. EMD by obtain the distance measure between patterns and determine the cost of transformation can simply compute the similarity degree to identify the plausible author by his/her whiteprints. If distance between patterns d_{fp_i,p_j} define the neighborhood for feature values and f_{fp_i,p_j} indicate cost of transformation with concept of MST, the EMD measure can be calculate based upon these two parameters. The low degree of support value shows the higher cost of transformation which leads lower similarity between patterns. EMD measure is calculated by following equation:

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

III. Experimental Evaluation

The main objective of conducting the experiment is to evaluate the accuracy of proposed method and prove the uniqueness and power of detected evidence to support the conclusion and also analysis the efficiency of them to present in the forensic course. As the proposed framework consists two main phase in two different context (DataMinig and Pattern matching), we decide to implement the approach by means of Mweka Tool which Running Machine Learning Tool Weka from MATLAB. This tool makes facility for using Data Minig algorithm in proposed approach with pattern matching method.

Dataset is used to evaluate proposed approach origins from Enron E-mail dataset which contains 200,399 real-life E-mails from 158 employees of the Enron Corporation, but our experiments doesn't limited to English language and employed this approach for Persian Language. By our knowledge this is the first use of Persian language for this application. The data used for Persian language is come from personal E-mail (due to lack of accessibility to reputed Persian E-mail dataset). The experiment is done in 2 section related to English and Persian languages.

A. English language

In order to design the experiments, we used m E-mail for n employees which randomly selected from the Enron E-mail Dataset. These n employee consider as suspected authors in this problem which represented by $S = \{S_1, \dots, S_n\}$. For each suspected authors S_i select m E-mail $E_i = \{e_1, \dots, e_m\}$ where $1/3$ applied for testing and $2/3$ of whole m allocated for learning. Let $E = \{E_1, \dots, E_n\}$ be a set of E-mails belongs to n suspected authors.

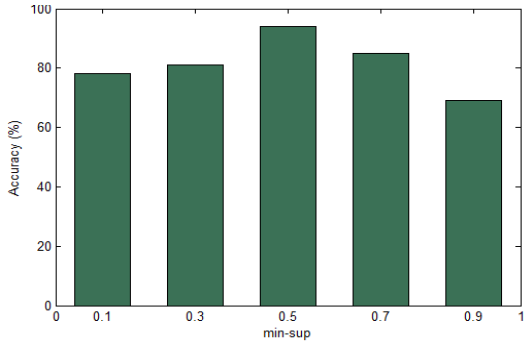

Figure 1 Accuracy of proposed method across various min_sup (m=15, n=10)

In order to analysis the influence of MST, the proposed algorithm is experimented with different *min_sup*. As depicted in figure 1 the accuracy of proposed kept robustness when applied various *min_sup*. The number of author selected for this assessment is 10 and respectively the E-mail for each author consider with 15 numbers (i.e., a total of 150 E-mails).

As depict in figure 2 by increasing the number suspected authors, the reduction of accuracy isn't considerable and the influence of various number of author in accuracy can prove stableness of the proposed algorithm as the degree of reduction accuracy only be 19% by increasing the author to 16. In order to show the efficiency in contrast of the other algorithm (as this domain is in infancy age, the related algorithm aren't so much, we also choose an algorithm which is more closer to our proposed method and have a good performance in compare with other methods [6-8]) we used AuthorMiner algorithm which applied in [6] and it consider as few method which used DataMinig technique for solve author attribution problem in cyber-forensic domain. As shown in figure 2 the accuracy of proposed method is higher than the AuthorMiner by different number of suspected authors

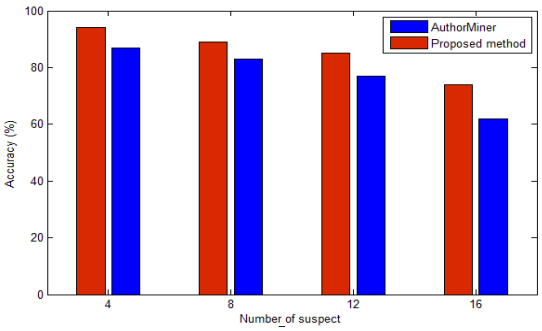

Figure 2: Comparison proposed method with AuthorMiner Algorithm [6] (m=15).

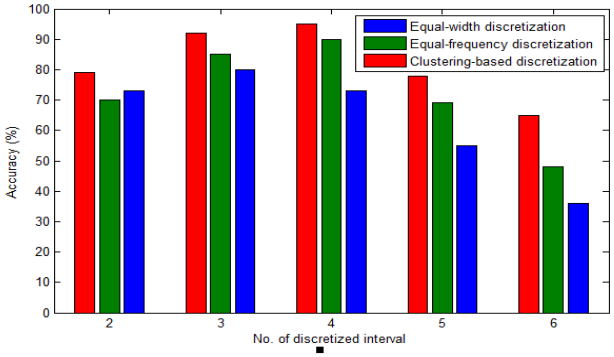

Figure 3: Performance of different discretization method in proposed algorithm

In the cyber forensic domain the efficient data affect on quality of evidence and also in accuracy of identification. One of the important phases which could prepare suitable data is discretization process and determining the disjoint point for clustering data. As figure 3 depicts three kinds of discretization methods are compared and clustering based method which using ADOC measure for clustering data presents the higher accuracy than Equal-width and Equal frequency discretization based algorithm.

The proposed algorithm applied EMD measure to identifying the anonymous author among a group of suspicious 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 of its power for detecting the plausible author based upon the most similarity degree. The experiment result show (Table 2) the short distance between writeprint of an individual and its E-mail's patterns and respectively long distance between patterns of various authors show the uniqueness of captured writeprints. As shown in Table 2 the short distance reflect lower value which means higher degree of similarity and long distance referred to lower similarity. This facility eases the decision making for investigator about detecting plausible author and presents a clear view of captured evidences 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 identifying the plausible author with strong evidences.

Table 1 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.
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**

B. *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. The framework of these experiments is designed like English (as mention in 3-1 section). By comparing the performance of proposed algorithm in Persian and English language it shows that the accuracy of identification in Persian is lower than English but it's not so considerable. The experimental results on Persian language reflect **good** efficiency of proposed algorithm on Persian language but there's need to analysis the specialized feature that related to Persian language as the feature which used in this paper is same for English and we only adopt them in Persian language.

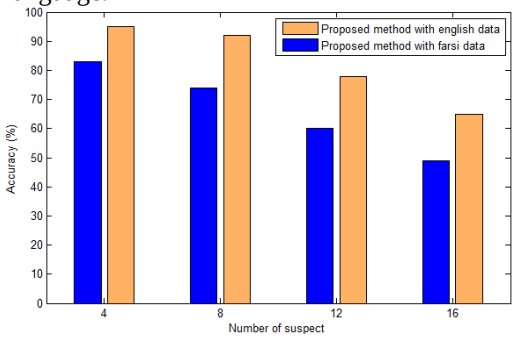

Figure 4: Accuracy of proposed method in Persian and English Language

IV. Conclusion

In this paper, we proposed a new approach for solving authorship attribution to identify anonymity defender in cyber forensic context. This approaches composed two main phase: writeprint extraction and pattern matching for detecting plausible author based upon the obtained writeprint. The feature used to create writeprints of each individual contain lexical, structural and content specific feature. Extraction phase is done by normalization and discretization feature by means ADOC measure for clustering and then extract frequent pattern with Apriori algorithm. After that removes the common patterns to capture the unique writeprints for each suspected author. The second phase is related to pattern matching by EMD measure to identify plausible author which his/her writeprint has the higher similarity with pattern of malicious E-mail. EMD calculate the distance between two patterns in order to determine the similarity degree. This method makes flexibility to detect distinction between similar patterns and reduces the errors in identifying plausible author.

The experimental results show the satisfactory performance of proposed algorithm with real life data. The proposed algorithm is evaluated in various aspects as accuracy in contrast of the other algorithm, accuracy with various numbers of suspected authors, accuracy of algorithm with different min_sup which the result proved stableness and robustness of this method. By conducting the experiments on English and Persian language, the results indicate better performance of algorithm on English language and a good result for Persian language as it was the first attempt on this language. We hope to improve the exited shortcoming in this language for further studying. The most of these deficiencies are related to stylometric feature which need a comprehensive analysis of Persian language properties to detect effective features for generating Persian writeprints.

The information provided by this approach made cybercrime investigator make decision based upon strong and traceable evidences. As these processes for finding benefit information couldn't be done manually and application of these methods create a new chance for detecting defender in cyber world and facilitate the anonymity problem.

Reference

[1] Gui-fa Teng; Mao-sheng Lai; Jian-Bin Ma; Ying Li, *"E-mail authorship mining based on SVM for computer forensic,"* Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on , vol.2, no., pp.1204,1207 vol.2, 26-29 Aug. 2004

[2] Zheng, Rong, et al. "*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 (2006): 378-393.

[3] De Vel, Olivier. "*Mining E-mail authorship.*" Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000). 2000.

[4] Quinlan, J. Ross. "*Induction of decision trees.*" Machine learning 1.1 (1986): 81-106.

[5] Cristianini, Nello, and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[6] Iqbal, Farkhund, et al. *"A novel approach of mining write-prints for authorship attribution in E-mail forensics."* digital investigation 5 (2008): S42-S51.

[7] Iqbal, Farkhund, et al. *"Mining writeprints from anonymous E-mails for forensic investigation."* digital investigation 7.1 (2010): 56-64.

[8] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, *"A unified data mining solution for authorship analysis in anonymous textual communications,"* Information Sciences, vol. 231, pp. 98-112, 2013.

[9] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. *"Mining association rules between sets of items in large databases."* ACM SIGMOD Record. Vol. 22. No. 2. ACM, 1993.

[10] Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. *"The earth mover's distance as a metric for image retrieval."* International Journal of Computer Vision 40.2 (2000): 99-121.

[11] Rubner, Yossi, Leonidas J. Guibas, and Carlo Tomasi. *"The earth mover's distance, multi-dimensional scaling, and color-based image retrieval."* Proceedings of the ARPA image understanding workshop. 1997.

[12] Hitchcock, Frank L. *"The distribution of a product from several sources to numerous localities." J. Math.* Phys 20.2 (1941): 224-230.

[13] Wan, Xiaojun. *"A novel document similarity measure based on earth mover's distance."* Information Sciences 177.18 (2007): 3718-3730.

[14] Abbasi, Ahmed, and Hsinchun Chen. "*Writeprints: A stylometric approach to identity-level identification and*

similarity detection in cyberspace." ACM Transactions on Information Systems (TOIS) 26.2 (2008): 7.

[15] Han, Jiawei, and Jian Pei. *"Mining frequent patterns by pattern-growth: methodology and implications." ACM SIGKDD* explorations newsletter 2.2 (2000): 14-20.

[16] Zaki, Mohammed Javeed. "*Scalable algorithms for association mining.*" Knowledge and Data Engineering, IEEE Transactions on 12.3 (2000): 372-390.

[17] Logan, Beth, and Ariel Salomon. *"A music similarity function based on signal analysis."* IEEE International Conference on Multimedia and Expo. 2001.

[18] Bae, Eric, 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 (2010): 427-471.