Author Identification for E-mail Forensic

Sobiya R. Khan
P.G. Dept. of Computer Science & Engineering
GHRCE, Nagpur, (M.S), India

Smita M. Nirkhi
P.G. Dept. of Computer Science & Engineering
GHRCE, Nagpur, (M.S), India

R. V. Dharaskar
M. P. G. I,
Nanded,
(M.S), India

ABSTRACT
E-mail communication has become the need of the hour, with the advent of Internet. However, it is being abused for various illegitimate purposes, such as, spamming, drug trafficking, cyber bullying, phishing, racial vilification, child pornography, and sexual harassment, etc. Several cyber crimes such as identity theft, plagiarism, internet fraud stipulate that the true identity of the e-mail's author be revealed, so that the culprits can be punished in the court of law, by gathering credible evidence against them. Forensic analysis can play a crucial role here, by letting the forensic investigator to gather evidence by examining suspected e-mail accounts. In this context, automated authorship identification can assist the forensic investigator in cyber crime investigation. In this paper we discuss how existing state-of-the-art techniques have been employed for author identification of e-mails and we propose our model for identifying most plausible author of e-mails.

General Terms
Data Mining, Machine Learning, Digital Forensic.

Keywords
Cyber Crime, E-mail forensic analysis, Author Identification, Stylometric techniques.

1. INTRODUCTION
Nowadays, e-mail has become an inseparable mechanism of communication over the Internet and Intranet. It is being used by government, industries and individual as well because of its simplicity, ease of use and expediency. However, the inherent nature of the e-mail communication is susceptible to illegitimate use. This has given rise to grave concerns with the increase in cyber crime committed via e-mails. The prime reason for this abuse is anonymity, because e-mail headers can be forged easily, and the path through which the e-mail arrived can be made anonymous. Thus, e-mail is becoming a popular and easy medium for committing various cyber crimes. The various e-mail mediated crime range from sending spam e-mails to severe crimes like child pornography. It is crucial to identify the true authors of the written e-mails in cases of forgery, identity theft, plagiarism and fraud.

However identifying the author's real identity is difficult, as the sender will try to hide his/her identity in order to shun from getting exposed. However, human beings are creators of habit and while we write something we follow certain personal traits which get reflected in our writings. This is the reason we have a consistent handwriting style for most of our life time for example, although the style may vary a bit, as we grow older. Similarly for writing, an individual has certain inherent habits, which are unconscious and deeply ingrained. This means that even if one attempt to make a conscious effort to disguise one's writing style, there will be some inherent features which will exemplify the individual's written text. Such features include the individual's familiarity to language, composition and writing style, syntactic and structural layout, and particular usage of certain taxonomy, vocabulary richness, stylistic and sub-stylistic features, to name a few. Thus, as a text categorization problem it is evident we can attempt to identify the most plausible author of a written text, if previous work of that author is available.

In order to gather sufficient and accurate evidence for courtroom, a cyber crime investigation is taken into account so that the illegitimate e-mail's real identity could be exposed and credible evidence can be collected for computer forensic professionals and law enforcement agencies. However, the large amount of cyber space activities and their anonymous nature make cyber-crime investigation extremely difficult. Conventional methods employed to deal with this problem rely on manual apprach, which is a tiring job with constantly changing e-mail ids. In this context, automatic authorship analysis of e-mail ensembles can be of high value to cybercrime investigators.

1.1 Challenges for Author Identification in E-mail
The following are the various challenges for authorship attribution in E-mail:

1. E-mails are generally short in length thus certain language based metrics may not be appropriate (e.g., vocabulary richness).
2. The composition style used in writing e-mail is often different from normal text documents even if written by the same author.
3. E-mail are generally brief, might contain lots of spelling mistakes and grammatical mistakes.
4. E-mail interaction between individuals can be frequent, similar to speech interactivity. Thus we can say that e-mail has elements from both formal writing and speech as well, hence more interactive in nature.
5. The writing style of individual can vary based on the intended recipient, for example e-mail written to boss and e-mail written to family members.
6. The vocabulary usage isn't consistent, thus giving way for forgery and imitation.
7. Usually e-mail contains few sentences/paragraphs which make it rather difficult to apply generic techniques for analysis.

However, despite all these fact it is evident that it is possible to identify certain prominent characteristics, such as syntactic, structural layout, vocabulary usage, unusual language usage, stylistic and sub-stylistic features, etc. which can be used to profile the writing style of an individual. Another new challenge is that cyber criminals can use any language to conduct crime. In fact, most big crime groups or terrorists have international characteristics. They use the Internet to formulate plans, raise funds, spread propaganda, and communicate. Applying authorship analysis in a multilingual context is becoming an important issue.

The rest of the paper is divided into three sections. Section II gives a brief overview of the related work. Section III gives an overview of our proposed model for e-mail author identification. Section IV presents the conclusion drawn.

2. RELATED WORK

Authorship Analysis. Authorship analysis is a process of examining the characteristics of a piece of writing to draw conclusions on its authorship. Its roots are from linguistic research area called stylometry, which refers to statistical analysis of literary style. As more sophisticated techniques, such as machine learning techniques, have been applied to this domain, this area of research has been generally recognized as authorship analysis. Authorship analysis is categorized into three major categories:

Authorship identification. This determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author. It also is called "authorship attribution" in some literature, especially by linguistic researchers. The origins of this field date back to the 18th century when English logician Augustus de Morgan suggested that authorship might be settled by determining if one text contains more long words than another. His hypothesis was investigated in [14], which subsequently published his results of authorship attribution among Bacon, Marlowe, and Shakespeare. The most thorough and convincing study in this field was conducted in [12, 13]. In their study on the mystery of the authorship of the Federalist Papers, they attributed all 12 disputed papers to Madison. Their conclusion was generally accepted by historical scholars and became a milestone in this research field.

Authorship characterization. It summarizes the characteristics of an author and generates the author profile based on his or her writings. Some of these characteristics include gender, educational and cultural background, and language familiarity. This relatively new research direction grew out of the authorship identification studies. The authors in [16] firstly made the nexus between authorship identification and characterization by analyzing the plays written by Middleton Thomas and others. He used salient common words which could best discriminate Middleton from others to define the descriptive writing habit of Middleton. More recently, other implicit characteristics of the authors have been investigated in [17-19].

Similarity detection. It compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author. Most studies in this category are related to plagiarism detection. Plagiarism involves the complete or partial replication of a piece of work without permission of the original author. Plagiarism detection attempts to detect the plagiarism activity through examining the similarity between two pieces of writings. Since similarity detection differs much from author identification in various aspects, it is beyond the scope of this article.

Authorship analysis has been applied to online messages in recent years. Authorship analysis has been used in a small but diverse number of application areas and examples include identifying authors in written literature, in program source code [11], and in forensic analysis for criminal cases.

2.1 Feature Selection

The essence of authorship analysis is the recognition of a set of features, or metrics, that remain relatively constant for a large number of writings created by the same individual. We can say that a set of writings from one author would have greater similarity in terms of such extracted features with respect to the same person's writing, than a set of writings from different individuals.

The following is a brief outline of the approaches undertaken for feature selection in authorship identification:

2.1.1 Word based approach

Initially researchers identified authors by categorizing different sets of words used by different authors. The most extensive and comprehensive application of authorship analysis is in literature and in published articles e.g., the disputed Federalist papers and Shakespeare's works. In these studies, specific author features such as unusual diction, frequency of certain words, choice of rhymes, and habits of hyphenation have been used as tests for author attribution. Modal testing based on keyword usage was conducted. However, the effectiveness of this approach is limited by the fact that word usage is highly dependent on the text topic. These features are examples of stylistic evidence which can be useful in establishing the authorship of a text document.

2.1.2 Stylometric based approach

But, for discrimination purposes we need "content-free" features. Stylometric features also termed as "style markers", used in early authorship attribution studies, were *character* or *word based*, such as vocabulary richness metrics (e.g., Zipf's word frequency distribution and its variants), word length etc. A given author's style is comprised of a number of distinctive features or attributes sufficient to uniquely identify the author. However, Stylometric features could be generated under the conscious control of the author and, hence, may be content-dependent and are a function of the document topic, genre, epoch etc.

2.1.3 Syntax based approach

Syntax-based features can be more reliable in authorship identification problems than word-based features. It is better to employ features derived from words and/or syntactic patterns, since such features are more likely to be content-independent and thus potentially more useful in discriminating authors in different contexts. The syntactic structure is usually generated dynamically and sub-consciously when language is created, similar to the case of the generation of utterances during speech composition and production. That is, language patterns or syntactic features are generated beyond an author's conscious control. e.g., the short all-purpose words (referred to as function words) such as "the", "if", "to" etc. whose frequency or relative frequency of usage is unaffected by the subject matter. Another example of syntactic feature is punctuation which is thought to be the graphical correlate of intonation which is the phonetic correlate of syntactic structure. Punctuation will vary from author to author. In [20] authors have shown that punctuation can be useful in discriminating authors. Therefore, a combination of syntactic features may be sufficient to uniquely identify an author.

In [15], over 1,000 stylometric features have been proposed and in [21] also list a variety of different stylometric features. However, no set of significant style markers have been identified as uniquely discriminatory.

2.2 Techniques for Authorship Analysis

The following is a list of the techniques used for authorship identification in the literature described below:

1. Statistical approaches (e.g., CUSUM, Thisted and Efron test),

2. Neural networks approaches (e.g., radial basis functions, feed-forward neural networks, back-propagation network)
3. Cascade correlation
4. Genetic algorithms
5. Markov chains, etc.

As the size of feature set became larger, conventional methods gave way to some more powerful analytical methods such as machine learning methods. Statistical and machine learning techniques constitute the two most common analytical approaches to authorship attribution. Many multivariate statistical approaches such as principal component analysis have shown a high level of accuracy [8]. However, these approaches also have do have their own lacunas. Machine learning techniques emerged from the drastic increases in computational power over the past several years. These techniques include support vector machines (SVMs), neural networks, and decision trees. They have gained wider acceptance in authorship analysis studies in recent years because they provide greater scalability than statistical techniques for handling more features, and they're less susceptible to noisy data [3-5][7][9][22]. Recently, data mining technique of frequent pattern mining has also been applied successfully which has shown promising results [1][2][6]. These benefits are important for working with online messages, which involve classification of many authors and a large feature set.

2.2.1 Stylometric features.

Writing styles are defined in terms of stylometric features. Writing patterns are usually the characteristics of words usage, words sequence, composition and layouts, common spelling and grammatical mistakes, vocabulary richness, hyphenation, and punctuation. However, there is no such feature set that is optimized and is applicable equally in all domains. Stylometric analysis techniques can be broadly classified into supervised and unsupervised methods. Supervised techniques are those that require author-class labels for categorization, while unsupervised techniques make categorizations with no prior knowledge of author classes.

2.2.1.1 Supervised techniques

Supervised techniques used for authorship analysis include support vector machines (SVMs), neural networks, decision trees and linear discriminant analysis. SVM is a highly robust technique that has provided powerful categorization capabilities for online authorship analysis. It has outperformed other supervised methods

2.2.1.2 Unsupervised techniques

This category includes principal component analysis (PCA) and cluster analysis. PCA's ability to capture essential variance across large number of features in a reduced dimensionality makes it attractive for text analysis problems, which typically involve large feature sets. PCA has been used in numerous previous studies for authorship attribution and has been shown effective for online stylometric analysis.

The commonly used stylometric features used in various authorship analyses [3-4] [5] [7] are described below:

1. Token-based features are collected either in terms of characters or words. In terms of characters, for instance, frequency of letters, frequency of capital letters, total number of characters per token and character count per sentence are the most relevant metrics. These indicate the preference of an individual for certain special characters or symbols or the preferred representation of certain units.

2. Word-based lexical features may include word length distribution, words per sentence, and vocabulary richness.

3. Syntactic features were the first who discovered that punctuation and function words are context-independent and thus can be applied to identify writers based on their written works.

4. Structural features are used to measure the overall appearance and layout of a document. For instance, average paragraph length, number of paragraphs per document, presence of greetings and their position within an e-mail, are common structural features.

5. Content-specific features are collection of certain keywords commonly found in a specific domain and may vary from context to context even for the same author.

6. Idiosyncratic Features represent 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.

The recent approach of frequent pattern mining implemented in [1][2][6] have shown a promising direction for authorship identification of e-mails. Here, instead of treating the author identification as classification problem, the authors have tried to discover the write-print of the individual by identifying unique stylistic and structural features from his/her writing. There study shows that clustering techniques can be combined with classification techniques to generate better promising results.

3. PROPOSED SYSTEM

In this paper, we are extending our model of proposed system for E-mail Author Identification as briefly outlined in [10]. We are proposing to extend the approach followed in [6], and will be using the concept of frequent pattern mining along with the various writing stylometric features to identify the most plausible author of an anonymous e-mail. The frequent pattern mining algorithm FP-Growth will be used instead of Apriori as used in [6] which is more robust.

The authorship-identification process will be divided into four steps:

Step 1. Collection of E-mail Messages

This employs the collection of set of messages written by potential authors to identify the writing styles of each author.

Step 2. Feature Extraction

E-mail messages will be pre-processed and represented in Vector Space Model whose entries will be defined by the various feature set identified for the respective individual which will represent his/her unique writing style.

Step 3. Model Generation

This is an iterative procedure in which the concept of frequent pattern mining using FP-Growth algorithm will be used to perform the task of generating unique writing features of individual authors.

Step 4. Author Identification

Once this model has been generated, it will be used to predict the authorship of unknown E-mail messages.

4. CONCLUSION

As a result of growing e-mail misuse, investigators need efficient automated methods and tools for analyzing e-mail ensembles to assist investigators gather clues and evidence. This automation should offer different functionalities ranging from e-mail storing, editing, searching, and querying to more advanced functionalities such as authorship identification, analysis and verification. Specifically, author identification is crucial in any cyber forensic investigation with the crime level going International. Since e-mail is now extremely important for inter-personal communication and professional life, this problem demands immediate attention and efficient solutions.

Data Mining and Machine Learning techniques show a promising solution in this problem domain. However other major concern with author identification is the support of various formats of e-mail & in various languages. Also, efforts should be made to overcome the limitation message-level analysis of E-mails in identifying texts shorter than 250 words. Challenging future direction is the generation of optimal feature set for a given data set along with support for multiple languages.

5. REFERENCES

[1] Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal, Adam Szporer, Djamel Benredjem, "Towards an integrated e-mail forensic analysis framework", Digital Investigation 5, pp.124–137, 2009.

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

[3] 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 ; 57(3), pp.378– 93, 2006.

[4] Zheng R, Qin Y, Huang Z, Chen H., "Authorship analysis in cybercrime investigation", In: Proc. 1st NSF/NIJ symposium. ISI Springer-Verlag; pp. 59–73, 2003.

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

[6] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi., "Mining writeprints from anonymous e-mails for forensic investigation", Digital Investigation, 2010.

[7] Olivier de Vel, "Mining E-mail Authorship", KDD-2000 Workshop on Text Mining, August 20, Boston, 2000.

[8] Abbasi A, Chen H., "Writeprints: a stylometric approach to identity level identification and similarity detection in cyberspace", ACM Transactions on Information Systems, Vol.26, No.2, Article 7, March 2008.

[9] Jiexun Li, Rong Zheng, Hsinchun Chen, "From Fingerprint to Writeprint", Communications of the ACM, 2006.

[10] Sobiya R. Khan, Smita M. Nirkhi, R. V. Dharaskar, "E-mail Mining for Cyber Crime Investigation", Proceedings of International Conference on Advances in Computer and Communication Technology, pp.138-141, February 2012.

[11] Gray, A., Sallis, P., & MacDonell, S., "Software forensics: Extending authorship analysis techniques to computer programs", Third biannual conference of the International Association of Forensic Linguists (IAFL '97), 1997.

[12] Mosteller, F., & Wallace, D.L., "Applied Bayesian and classical inference: The case of the Federalist Papers", Second edition, New York: Springer- Verlag, 1964.

[13] Mosteller, F., & Wallace, D.L., "Inference and disputed authorship: The Federalist. Reading", MA: Addison-Wesley, 1964.

[14] Mendenhall, T.C., "The characteristic curves of composition", Science, 11(11), 237–249, 1887.

[15] Rudman, J., "The state of authorship attribution studies: Some problems and solutions", Computers and the Humanities, 31, 351–365, 1998.

[16] Craig, H. , "Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?", Literary and Linguistic Computing, 14(1), 103–113, 1999.

[17] Corney, M., de Vel, O., Anderson, A., & Mohay, G. , "Gender-preferential text mining of E-mail discourse", Eighteeth annual Computer Security Applications Conference (ACSAC 2002), Las Vegas, NV, 2002.

[18] Argamon, S., S ˇ ari´c, M., & Stein, S.S., "Style mining of electronic messages for multiple authorship discrimination", Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 475–480). ACM Press, 2003.

[19] Koppel, M., Argamon, S., & Shimoni, A.R., "Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412, 2002.

[20] Chaski, C., "Empirical evaluations of language-based author identication techniques",Forensic Linguistics, 8, 2001.

[21] Tweedie, F.J., & Baayen, R.H. , " How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32, 323–352, 1998.

[22] Gui-Fa Teng'J, Mao-Sheng Lai I, Jian-Bin Ma', Ying Li, "E-mail Authorship Mining based on SVM for Computer Forensic", Proceedings of the Third International Conference on Machine Learning and Cybermetics, Shanghai, August, pp.26-29, 2004.