

From Fingerprint to Writeprint

Jiexun Li, Rong Zheng, Hsinchun Chen

Department of Management Information Systems
The University of Arizona
Tucson, AZ 85721
Email: {jiexun, rong, hchen}@eller.arizona.edu

Fingerprint-based identification has been the oldest biometric technique successfully used in conventional crime investigation. The unique, immutable patterns of a fingerprint, i.e., the pattern of ridges and furrows as well as the minutiae points, can help a crime investigator infer the identities of suspects.

However, circumstances have changed since the emergence and rapid proliferation of cybercrime. Generally, cybercrime includes Internet fraud, computer hacking/network intrusion, cyber piracy, spreading of malicious code, and so on. Cyber criminals post online messages over various Web-based channels to distribute illegal materials, including pirated software, child pornography materials, and stolen property. Moreover, international criminals and terrorist organizations such as Osama bin Laden and Al Qaeda use online messages as one of their major communication media. Since people are not usually required to provide their real identity in cyberspace, the anonymity makes identity tracing a critical problem in cybercrime investigation. This problem is further complicated by the sheer amount of cyber users and activities.

Unlike conventional crimes, there are no fingerprints to be found in cybercrime. Fortunately, there is another type of print, which we call “writeprint,” hidden in people’s writings. Similar to fingerprints, writeprint is composed of multiple features, such as

vocabulary richness, length of sentence, use of function words, layout of paragraphs, and key words. These writeprint features can represent an author's writing style, which is usually consistent across his or her writings, and further become the basis of authorship analysis. This study is aimed at introducing a method of identifying the key writeprint features for authors of online messages to facilitate identity tracing in cybercrime investigation.

Authorship Analysis and Writeprint Features

Authorship analysis is a process of categorizing articles by authors' writing style and is often viewed in the context of stylometric research [7]. It has its most extensive applications to historical literature [4, 8]. Some recent studies introduced this approach to online messages and showed promising results [3]. This research field can be categorized into authorship identification, authorship characterization, and similarity detection. Authorship characterization is aimed at inferring an author's background characteristics rather than identity. Similarity detection compares multiple pieces of writing without identifying the author. Authorship identification determines the likelihood of a particular author having written a piece of work by examining other works produced by that author. In this study we are particularly interested in authorship identification because it is the most relevant to cybercrime investigation.

The essence of authorship identification is to identify a set of features that remain relatively constant among a number of writings by a particular author. Given n predefined features, each piece of writing can be represented by an n -D feature vector. Supervised learning techniques such as ID3, Neural Network (NN) and Support Vector Machine (SVM) can train and generate a classifier so as to determine the category of a new vector, i.e., the

authorship of an anonymous writing. In such a process, the classification technique is very important to the performance of authorship identification. Support Vector Machine has been frequently used in previous authorship identification studies [3, 12]. We have shown in our previous work [12] that SVM outperforms other classification methods such as decision tree and neural network for authorship identification of online messages. In addition, the predefined feature set is another crucial factor. The writeprint features proposed in previous literatures can be divided into four types as follows.

Lexical features, the earliest features used in authorship analysis, represent an author's lexicon-related writing styles. Most of them are character-based and word-based features. For example, Elliot and Valenza conducted modal testing based on word usage to compare the poems of Shakespeare with those of Edward de Vere, the leading candidate as the true author of the works credited to Shakespeare [4]. In Yule's early work some more generic features were employed, such as sentence length and vocabulary richness [11]. Later Burrows developed a set of more than 50 high-frequency words which were tested on the Federalist Papers [2]. Holmes analyzed the use of "shorter" words (2 or 3 letter words) and "vowel words" (words beginning with a vowel) [5].

Syntactic features, including punctuation and function words, can capture an author's writing style at the sentence level. They are often "content-free" features derived from people's personal habits of organizing sentences. In the seminal work conducted by Mosteller and Wallace [8], they first used the frequency of occurrence of thirty function words (e.g. "while" and "upon") to clarify the disputed work, Federalist Papers. Subsequently different function words were examined and showed good discriminating capability [5]. Baayen et al.

concluded that incorporating punctuation frequency as a feature can improve the performance of authorship identification [1]. Stamatatos et al. introduced more complex syntactic features such as passive count and part-of-speech tags [10]. These studies demonstrated that syntactic features might be more reliable than lexical features in authorship identification.

Structural features, in general, represent the author's habits when organizing a piece of writing. Habits such as paragraph length, use of indentation, and use of signature can be strong authorial evidence of personal writing style. Structural layout traits and other features have been introduced by de Vel et al. for email author identification and achieved high identification performance [3].

Content-specific features refer to keywords in a specific topic. Although seldom used in previous studies, these features could complement "content-free" features to improve the performance of authorship identification for particular applications. In the cybercrime context, a cyber criminal often posts illegal messages involving a relatively small range of topics, e.g., pirated software and child pornography. Hence, special words or phrases closely related to specific topics may provide some clues about the author. For example, a criminal selling pirated software may use such words as "obo" or "for sale;" one distributing child pornography materials is likely to use words such as "sexy."

From a multilingual perspective, different languages may share similar writeprint features, such as structural features. However, due to the uniqueness of language, some features are not generic. For example, while most Western languages have boundaries between words, most Oriental languages do not. In addition, different languages can have different function words and word-based features. Rudman summarized almost 1,000

writeprint features for English used in authorship analysis applications [9]. Based on previous literature, Zheng et al. created a taxonomy of writeprint features for online messages, including 270 features for English (87 lexical features, 158 syntactic features, 14 structural features, and 11 content-specific features) and 114 for Chinese (16 lexical features, 77 syntactic features, 11 structural features, and 10 content-specific features) [12].

A number of studies have shown the discriminating power of different types of features. Furthermore, researchers attempt to identify an optimal set of features for authorship identification. Most previous studies of feature choice compare different types of features. Even if a type of feature is effective for authorship identification, some features in this type may be irrelevant or redundant, hence reducing the prediction accuracy. For instance, de Vel et al. observed a reduction in performance when the number of function word features was increased from 122 to 320 [3]. Feature selection should be undertaken to remove features that do not contribute to prediction [3, 5]. To our best knowledge, however, few studies have been conducted to select key features for authorship identification at the individual feature level. In addition, extracting a large number of features from online messages is time-consuming and may induce errors. Therefore, it is important to identify the key writeprint features for authorship identification of online messages. Due to the multilingual characteristic of online messages, in this paper we study writeprint features of different languages, i.e., English and Chinese.

Feature Selection

Since features are regarded as an abstract representation of writeprint, the quality of the feature selection directly influences this representation. Feature selection techniques aim to

select a subset of features that are relevant to the target concept, i.e., writeprint in this study. There are a variety of well developed methods in the pattern recognition and data mining domains to identify important features. Liu and Motoda summarized past studies of feature selection into a general framework [6]. The process of feature selection can be viewed as a search problem in feature space. Exhaustive search and heuristic search are two major search strategies. Exhaustive search tries every feature combination to achieve the optima but is computationally infeasible for large feature sets. Heuristic search uses certain rules to guide the direction of the search. This search strategy reduces the size of the search space and therefore speeds up the process significantly. The major heuristic search algorithms include hill climbing, best first search and, generic algorithm (GA). GA behaves like a metaphor of the processes of evolution in nature. The optimal solution, i.e., the chromosome with the highest fitness value, can be achieved via a number of generations by applying genetic operators such as selection, crossover, and mutation. GA can avoid local optima and provide multi-criteria optimization functions.

A GA-based Feature Selection Model

In this study we proposed a GA-based feature selection model to identify writeprint features. In such a model each chromosome represents a feature subset, where its length is the total number of candidate features and each bit indicates whether a feature is selected or not. Specifically, 1 represents a selected feature while 0 represents a discarded one. For example, a chromosome representing five candidate features, "10011," means that the first, fourth and fifth features are selected, while the other two are discarded. In the first generation each bit of chromosomes is assigned to 0, meaning that none of the features is selected. Each

chromosome, i.e., a feature subset, can be employed to train a classifier. Thus, the fitness value of each chromosome is defined as the accuracy of the corresponding classifier. By applying genetic operators in the successive generations, the GA model can generate different combinations of features to achieve the highest fitness value. Therefore, the feature subset corresponding to the highest accuracy of classification along all the generations is regarded as the optimum. The selected features in this subset are the key writeprint features to discriminate the writing styles of different authors. The process of this GA-based feature selection is shown in Figure 1.

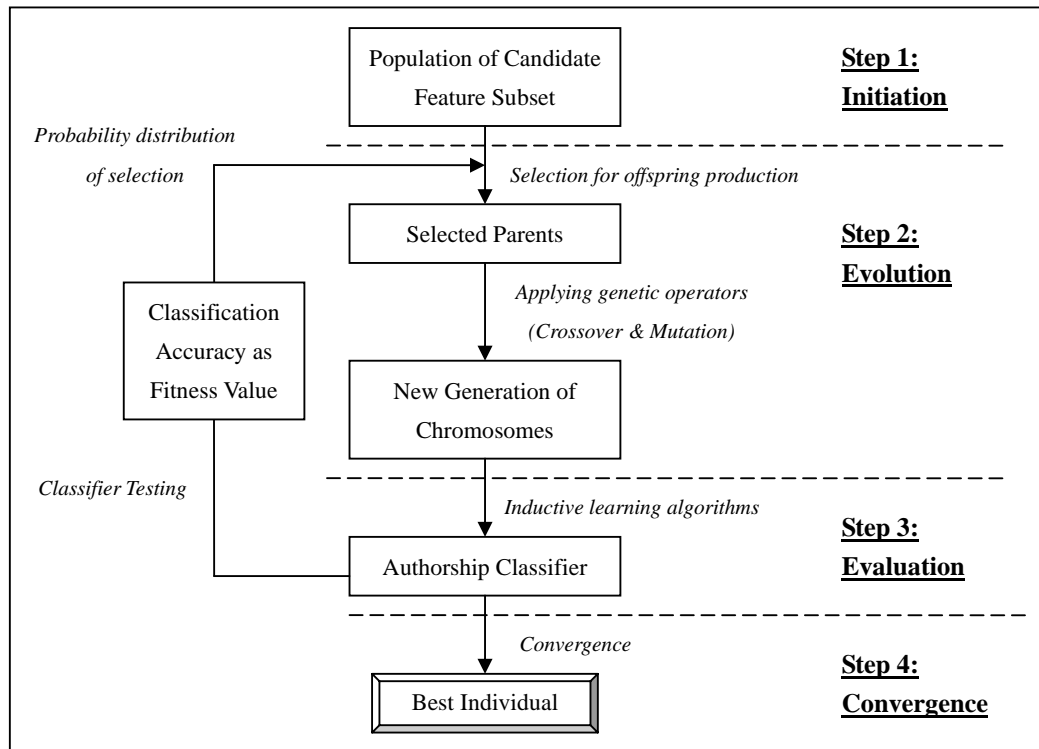


Figure 1. The Process of GA-based Feature Selection

Experimental Studies

Experiment Data Collection

To test the feasibility of the authorship identification and to identify the key writeprint features for online messages, two testbeds of online messages (English and Chinese) were

created. Since illegal online messages are of particular interest in this study, we collected messages that were involved in selling pirated software/CDs from misc.forsale.computers.* (including 27 sub-groups) in Google newsgroups to create the English testbed. The Chinese testbed is composed of online messages from the two most popular Chinese Bulletin Board Systems (smth.org and bbs.mit.net), involving seven different topics (e.g., movie, music, and novel). For each of the testbeds, we identified 10 of the most active authors with 30-40 messages collected for each of them. The average length of the messages written by each author is 169 words for the English testbed and 807 characters for the Chinese testbed. Based on our previous study [12], in total, 270 and 114 features are predefined and extracted from the English and Chinese messages, respectively. For online messages with such short length, when the full set of features are used, a sample size of about 30 messages per author is necessary to predict authorship with an accuracy of 80~90% [12].

Experimental Results

We applied the GA-based feature selection model on both the English and Chinese testbeds. Due to its good performance, the Support Vector Machine was selected as the classification model. The feature selection process started with an empty feature set, i.e., no feature was used. In the successive generations, the GA model conducted a global search for the optimal feature subset by applying the crossover and mutation operators. We observed a significant increase of fitness value (i.e., classification accuracy) in a number of early generations and a relatively constant accuracy afterwards. Meanwhile, the number of selected features in the best chromosome of each generation increased from 0 to about half of the full feature set. Figure 2 shows the change of accuracy and the number of selected features along

500 generations for the two testbeds. The evolutionary process of GA converged after about 50 generations for the English dataset and 120 generations for the Chinese dataset. Among all the chromosomes in the 500 generations, the one with the highest accuracy corresponded to the optimal feature subset.

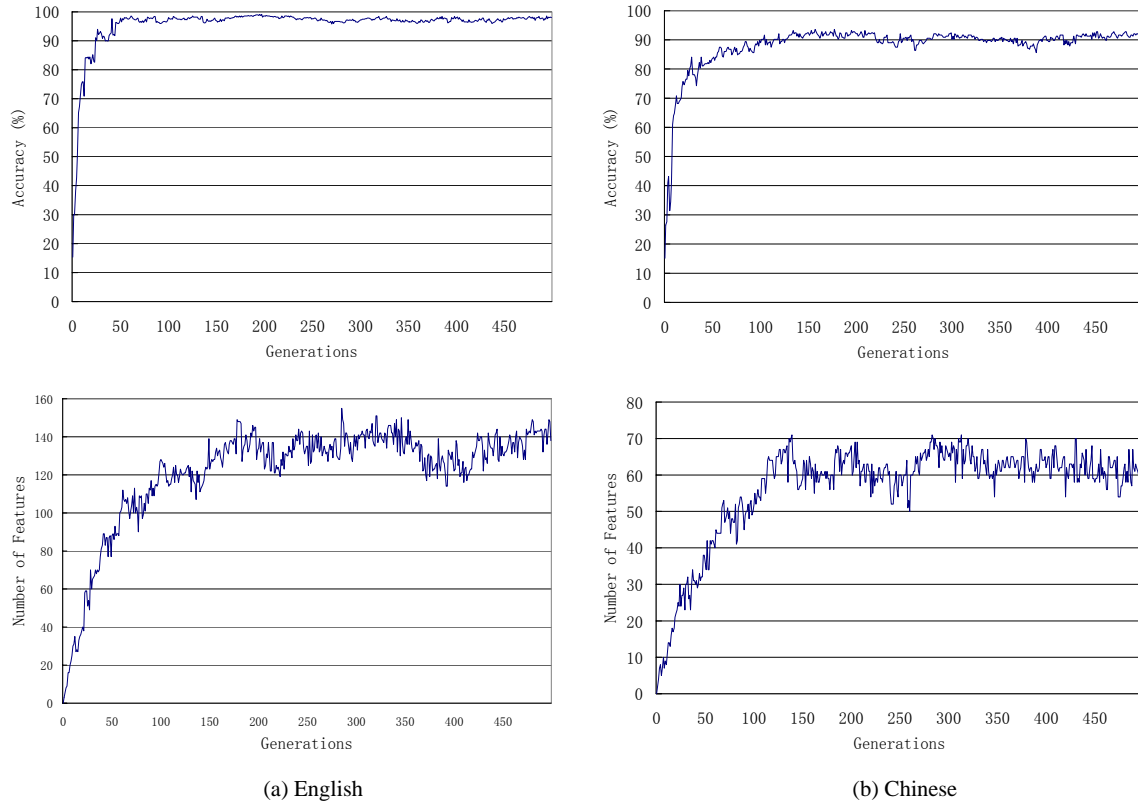


Figure 2. Experiment Results of Feature Selection on English and Chinese Testbeds

To compare the discriminating power of the full feature set and the optimal set, 30-fold pair-wise t-tests were conducted respectively for the English and Chinese datasets. As shown in Table 1, the GA-based model identified a feature subset with only about half of the full set as the key features, i.e., 134 out of 270 for English, and 56 out of 114 for Chinese. For the English dataset, the optimal feature set achieved a classification accuracy of 99.01%, which is significantly higher than 97.85% achieved by the full set (p-value = 0.0417). For the Chinese dataset, the optimal feature set achieved a classification accuracy of 93.56%, which is higher

than 92.42% achieved by the full set but not significantly (p-value = 0.1270). In general, using the optimal feature subset, we can achieve a comparable (if not higher) accuracy of authorship identification.

Table 1. Comparison between Full Feature Set and Optimal Feature Subset

<i>Dataset</i>	<i>Feature set</i>	<i>No. of Features</i>	<i>Mean Accuracy</i>	<i>Variance</i>	<i>P-Value</i>
English	Full set	270	97.85%	0.002	0.0417
	Optimal subset	134	99.01%	0.001	
Chinese	Full set	114	92.42%	0.023	0.1270
	Optimal subset	56	93.56%	0.026	

The effect of feature selection is significant and promising. Furthermore, we discovered that the selected key feature subset included all four types of features. This is consistent with our previous study in [12], which showed that each type of feature contributes to the predictive power of the classification model. In particular, the relatively high proportion of selected structural and content-specific features suggests their useful discriminating power for online messages. Table 2 illustrates several key features identified from the full feature set.

Table 2. Illustration of Key Writeprint Features

<i>Feature Type</i>	<i>English</i>	<i>Chinese</i>
Lexical	Total number of upper-case letters /total number of characters; Frequency of character “@” and “\$”; Yule’s K measure (vocabulary richness); 2-letter word frequency.	Total number of English characters /total number of characters; Total number of digits /total number of characters; Honore’s R measure (vocabulary richness).
Syntactic	Frequency of punctuation “!” and “:” Frequency of function word “if” and “can”	Frequency of function word “然后(then)” and “我想(I think)”
Structural	Number of sentences per paragraph; Has separators	Number of sentences per paragraph; Has separators
Content-specific	Frequency of word “check” and “sale”	Frequency of “音乐(music)” and “小说(novel)”

The results from Table 2 have some interesting implications. Since some features in the full feature set may be irrelevant for online messages, the frequency of characters related to

online messages (e.g., “@,” “\$”) instead of other common ones (e.g., “A,” “E”) were selected. In addition, since some features may only provide redundant information, the total number of upper-case letters/ total number of characters was identified as a key feature, while the frequency of lower-case letters was discarded. Similarly, only one vocabulary richness measure, e.g., Yule’s K or Honore’s R, was selected and others were ignored. Since online messages are often short in length and flexible in style, structural layout traits such as the average length of paragraphs became more useful. In addition, content-specific features are highly related to their context. Hence features such as “sale” and “check” were identified as the key content-specific features for the English dataset based on sales of pirated software/CDs. In other contexts, different content-specific features should be identified and used accordingly.

These selected key features of writeprint can effectively represent the distinct writing style of each author and further assist us to identify the authorship of new messages. Figure 3 exemplifies a comparison of writeprints between three authors in the English dataset using five of the key features, where feature values were normalized to [0, 1]. Clearly, Mike’s distinct writeprint from the other two indicates his unique identity. The high similarity between the writeprints of Joe and Roy suggests that these two IDs might be the same person.

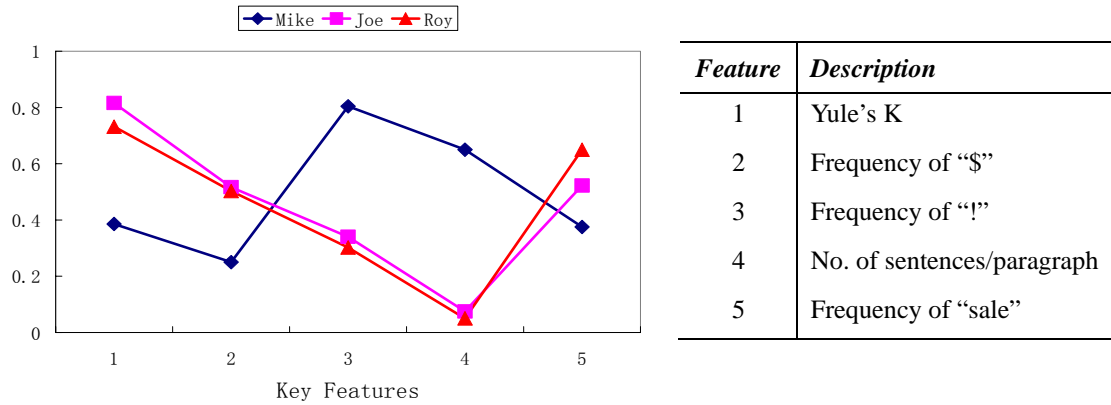


Figure 3. Comparison of Writeprints between Three Authors

Conclusions and Future Works

The absence of fingerprints in cyberspace leads law enforcement and intelligence community to seek for new approaches to trace criminal identity in cybercrime investigation. To address this problem, we propose to develop writeprint to help identify an author in cyberspace. We developed a GA-based feature selection model to identify the key features of writeprint specifically for online messages. Experimental studies on English and Chinese testbeds of online messages demonstrated the power and potential of the GA-based feature selection model. The identified key features could achieve comparable or even higher classification accuracy and effectively differentiate the writeprint of different online authors.

Currently, several interesting issues in this research domain are still open. Given the key features selected, we will continue to rank and cluster them based on their functional traits, and further provide a visual representation of an author’s writeprint. Due to the multinational nature of cybercrime, we plan to employ this feature selection model to identify the key writeprint features in other languages such as Arabic and Spanish. In addition, we are also interested in applying the writeprint identification approach to other related problems such as plagiarism detection and intellectual property checking.

References

1. Baayen, H., Halteren, H. v., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. *In Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002)*.
2. Burrows, J. F. (1992). Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing*, 2, 61-67
3. de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining E-mail content for author identification forensics. *SIGMOD Record*, 30(4), 55-64
4. Elliot, W., & Valenza, R. (1991). Was the Earl of Oxford the true Shakespeare? *Notes and Queries*, 38, 501-506.
5. Holmes, D. I. (1998). The evolution of stylometry in humanities. *Literary and Linguistic Computing*, 13(3), 111-117.
6. Liu, H., & Motoda, H. (1998). Feature selection for knowledge discovery and data mining. Kluwer Academic Publishers, Norwell, MA
7. McEnery A., & Oakes, M. (2000). Authorship studies/textual statistics. Marcel Dekker.
8. Mosteller, F., & Wallace, D. L. (1964). Inference and disputed authorship: the Federalist: Addison-Wesley
9. Rudman, J. (1998). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities*, 31, 351-365.
10. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471-495.
11. Yule, G. U. (1944). The statistical study of literary vocabulary. Cambridge University Press
12. Zheng, R., Li, J., Huang, Z. & Chen, H. (2003). A framework of authorship identification for online messages: writing style features and classification techniques, submitted to the *Journal of the American Society for Information Science and Technology (JASIST)*.