

## Chapter 8

# **PROTECTING CRITICAL INFRASTRUCTURE AND KEY ASSETS**

### **Chapter Overview**

The Internet is a critical infrastructure and asset in the information age. However, cyber criminals have been using various web-based channels (e.g., email, web sites, Internet newsgroups, and Internet chat rooms) to distribute illegal materials. One common characteristic of these channels is anonymity. Compared with conventional crimes, cybercrime conducted through such anonymous channels imposes unique challenges for researchers and law enforcement and intelligence agencies in criminal identity tracing. Law enforcement and intelligence agencies have an urgent need for approaches that automate criminal and terrorist identity tracing in cyberspace. Three case studies in this chapter demonstrate the potential of using multilingual authorship analysis with carefully selected writing style feature sets and effective classification techniques for identity tracing in cyberspace.



## 8.1 Case Study 11: Identity Tracing in Cyberspace

With the rapid proliferation of Internet technologies and applications, cybercrime has become a major concern for the law enforcement community. Cyber criminals have been distributing messages on the Internet to conduct illegal activities. The anonymous nature of online message distribution has made criminal identity tracing a critical problem in cybercrime investigation. We developed a framework for authorship identification of online messages to address the identity tracing problem. In this framework, three types of writing style features are extracted and inductive learning algorithms are used to build feature-based classification models to identify authorship of online messages.

Data used in this study were from open sources. Three datasets, two in English and one in Chinese, were collected. One of the English datasets consisted of 153 USENET newsgroup illegal sales of pirate CDs and software messages. We manually identified the nine most active users (represented by a unique ID and email address) who frequently posted messages in these newsgroups. The Chinese dataset contained 70 Bulletin Board System (BBS) illegal CD and software for-sale messages downloaded from a popular Chinese BBS.

The two key techniques used in this study were feature selection and classification. The objective was to classify text messages into different classes with each class representing one author. Based on the review of previous studies on text and email authorship analysis, along with the specific characteristics of the messages in our datasets, we selected a large number of features that were potentially useful for identifying message authors. Three types of features were used: *style markers* (content-free features such as frequency of function word, total number of punctuations, average sentence length, etc.), *structural features* (such as use of a greeting statement, use of farewell statement, etc.), and *content-specific features* (such as frequency of keywords, special character of content, etc.). Many additional metrics were used in our study as described in Figures 8-1, 8-2, and 8-3.

For classification analysis, three popular classifiers were selected, including the C4.5 decision tree algorithm (Quinlan, 1986), backpropagation neural networks (Lippmann, 1987), and support vector machines (Cristianini & Shawe-Taylor, 2000; Hsu & Lin, 2002). Each individual classifier has been employed in previous authorship analysis research (Diederich et al., 2000). In general SVM and neural networks have better performance than decision trees (Diederich et al., 2000). However, most previous authorship studies were based on newspaper articles such as the *Federalist Papers*. Because online messages are quite different from formal articles in style and

length, we needed to test the performances of these algorithms on our datasets.

**Additional style markers in this experiment:**

Total number of words in subject  
 Total number of characters in subject (S)  
 Total number of upper-case characters in words in subject/S  
 Total number of punctuations in subject/S  
 Total number of white-space characters in subject/S  
 Total number of lines  
 Total number of characters

\* We used 122 function words and 48 markers suggested by de Vel (de Vel et al., 2001). Another 28 of the most common function words from the *Oxford English Dictionary* and seven other markers were also included.

Figure 8-1. Style markers (205 features).

**Additional structural features in this experiment:**

Types of signature (name, title, organization, email, URL, phone number).  
 Uses special characters (e.g. -----) to separate message body and signature.

\* These additional structural features were used with the email dataset only. In the newsgroup messages used in this experiment, no attachment or re-quoted text was allowed.

Figure 8-2. Structural features (9 features).

Has a price in subject  
 Position of price in message body  
 Has a contact email address in message body  
 Has a contact URL in message body  
 Has a contact phone number  
 Uses a list of products  
 Position of product list in body message  
 Indicates product categories in list  
 Format of product list

Figure 8-3. Content-specific features (9 features, for newsgroup messages only).

Three experiments were conducted on the newsgroup dataset with one classifier at a time. First, 205 style markers (67 for Chinese BBS dataset) were used; nine structural features were added in the second run; and nine

content-specific features were added in the third run. A 30-fold cross-validation testing method was used in all experiments.

We used *accuracy*, *recall*, and *precision* to evaluate the prediction performance of the three classifiers. Accuracy represents the overall prediction performance of a classifier. For each particular author, we used precision and recall to measure the effectiveness of a classifier. The three measures are defined in equations (1)-(3).

$$\text{Accuracy} = \frac{\text{Number of messages whose author was correctly identified}}{\text{Total number of messages}} \quad (1)$$

$$\text{Recall} = \frac{\text{Number of messages correctly assigned to the author}}{\text{Total number of messages written by the author}} \quad (2)$$

$$\text{Precision} = \frac{\text{Number of messages correctly assigned to the author}}{\text{Total number of messages assigned to the author}} \quad (3)$$

We summarize the results as follows:

- *SVM and neural networks achieved better performance than the C4.5 decision tree algorithm.* For example, using style markers on the email dataset, the C4.5, neural networks, and SVM achieved accuracies of 74.29%, 81.11%, and 82.86% respectively. SVM also achieved consistently higher accuracies, precision, and recall than the neural networks. However, the performance differences between SVM and neural networks were relatively small. Our results were generally consistent with previous studies, in that neural networks and SVM typically had better performance than decision tree algorithms (Diederich et al., 2000).
- *Using style markers and structural features outperformed using style markers only.* We achieved significantly higher accuracies for all three datasets ( $p$ -values were all below 0.05) by adopting the structural features. This possibly resulted from an author's consistent writing patterns present in the message's structural features.
- *Using style markers, structural features, and content-specific features did not outperform using style markers and structural features.* The results indicated that using content-specific features as additional features did not improve the authorship prediction performance significantly (with  $p$ -value of 0.3086). We thought this was because authors of illegal

messages typically delivered diverse contents in their messages and little additional information could be derived from the message contents to determine the authorship. We also observed that high accuracies were obtained using only style markers as input features for the English datasets. The accuracies ranged from 71% to 89%. The results indicated that style markers alone contain a large amount of information about people's online message writing styles and were surprisingly robust in predicting the authorship.

- *There is a significant drop in prediction performance measures for the Chinese BBS dataset compared with the English datasets.* For example, when using style markers only, C4.5 achieved average accuracies of 86.28% and 74.29% for the English newsgroup and email datasets, while for the Chinese dataset it only achieved an average accuracy of 54.83%. A possible reason was that only 67 Chinese style markers were used in our current experiments, significantly fewer than the 205 style markers used with the English dataset. We expect to achieve higher prediction performances if additional Chinese style markers are identified and included. We also observed that when structural features were added all three algorithms achieved relatively high precision, recall, and accuracies (from 71% to 83%) for the Chinese dataset. Considering the significant language differences, our proposed approach to the problem of online message identity tracing appears promising in a multilingual context.

## 8.2 Case Study 12: Feature Selection for Writeprint

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 sentences, use of function words, layout of paragraphs, and keywords. 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.

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 compared 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 (de Vel, 2001). Feature selection should be undertaken to remove features that do not contribute to prediction. 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 study we examine writeprint features for different languages, i.e., English and Chinese.

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 in a general framework (Liu & Motoda, 1998). 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. In this study we proposed a genetic algorithm-based (GA) 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. 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.

To test the feasibility of authorship identification and to identify the key writeprint features for online messages, the two online message testbeds (English and Chinese) described in the previous section were used. 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 8-1, the GA-based model identified a feature subset with 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 the 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 the 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 8-1. Comparison between full feature set and optimal feature subset.

Dataset	Feature set	No. of Features	Mean Accuracy	Variance	P-Value
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 (Zheng et al., 2003), 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 8-2 illustrates several key features identified from the full feature set.

Table 8-2. Illustration of key English and Chinese writeprint features.

Feature Type	English	Chinese
<b>Lexical</b>	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).
<b>Syntactic</b>	Frequency of punctuation “!” and “.” Frequency of function word “if” and “can”	Frequency of function word “然后 (then)” and “我想(I think)”
<b>Structural</b>	Number of sentences per paragraph; Has separators	Number of sentences per paragraph; Has separators
<b>Content-specific</b>	Frequency of word “check” and “sale”	Frequency of “音乐(music)” and “小说(novel)”

The results from Table 8-2 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.

### **8.3 Case Study 13: Developing an Arabic Authorship Model**

The evolution of the Internet as a major international communication medium has spawned the advent of a multilingual dimension. Application of authorship identification techniques across multilingual web content is important due to increased globalization and the ensuing security issues that are created.

Arabic is one of the six official languages of the United Nations and the mother tongue of over 300 million people. The language is gaining interest due to its socio-political importance and differences from Indo-European languages. The morphological challenges pertaining to Arabic pose several critical problems for authorship identification techniques. These problems could be partially responsible for the lack of previous authorship analysis studies relating to Arabic.

In this study, we apply an existing framework for authorship identification to Arabic web forum messages. Techniques and features are incorporated to address the specific characteristics of Arabic, resulting in the creation of an Arabic language model. We also present a comparison of English and Arabic language models.

Most previous authorship studies have only focused on English, with a few studies done on Greek and Chinese. Stamamatos et al. applied authorship identification to a corpus of Greek newspaper articles (Stamamatos et al., 2001). Peng et al. conducted experiments on English documents, Chinese novels, and Greek newspapers using an n-gram model

(Peng et al., 2003). Zheng et al. performed authorship identification on English and Chinese web forum messages (Zheng et al., 2003). In all previous studies, English results were better than other languages. Applying authorship identification features across different languages is not without its difficulties. Since most writing style characteristics were designed for English, they may not always be applicable or relevant for other languages. Structural and other linguistic differences can create feature extraction nightmares.

Arabic is a Semitic language, meaning that it belongs to the group of Afro-Asian languages which also includes Hebrew. It is written from right to left with letters in the same word being joined together, similar to English cursive writing. Semitic languages have several characteristics that can cause difficulties for authorship analysis. These challenges include properties such as inflection, diacritics, word length, and elongation.

- **Inflection:**

Inflection is the derivation of stem words from a root. Although the root has a meaning, it is not a word but rather a class that contains stem instances (words). Stems are created by adding affixes (prefixes, infixes, and suffixes) to the root using specific patterns. Words with common roots are semantically related. Arabic roots are 3-5 letter consonant combinations with the majority being 3-letters. Al-Fedaghi and Al-Anzi believe that as many as 85% of Arabic words are derived from a tri-lateral root, suggesting that Arabic is highly inflectional (Al-Fedaghi and Al-Anzi, 1989). Inflection can cause feature extraction problems for lexical features because high levels of inflection increase the number of possible words, since a word can take on numerous forms.

- **Diacritics:**

Diacritics are markings above or below letters, used to indicate special phonetic values. An example of diacritics in English would be the little markings found on top of the letter “e” in the word résumé. These markings alter the pronunciation and meaning of the word. Arabic uses diacritics in every word to represent short vowels, consonant lengths, and relationships between words.

- **Word Length:**

Arabic words tend to be shorter than English words. The shorter length of Arabic words reduces the effectiveness of many lexical features. The short-word count feature, used to track words whose length is 3-letters or smaller, may have little discriminatory potential when applied to Arabic. Additionally, the word-length distribution may also be less effective since Arabic word-length distributions have a smaller range.

- Elongation:

Arabic words are sometimes stretched out or elongated. This is done for purely stylistic reasons using a special Arabic character that resembles a dash (“-”). Elongation is possible because Arabic characters are joined during writing. Table 8-3 shows an example of elongation. The word MZKR (“remind”) is elongated with the addition of four dashes between the “M” and the “Z.” Although elongation provides an important authorship identification feature, it can also create problems.

Table 8-3. An Arabic elongation example.

Elongated	English	Arabic	Word Length
No	MZKR	مذكر	4
Yes	M----ZKR	مــــذكر	8

Our testbed consisted of English and Arabic datasets. The English dataset was adapted from Zheng et al.’s study and consists of messages from USENET newsgroups (Zheng et al., 2003). The dataset identifies 20 authors engaged in potentially illegal activities relating to computer software and music sales and trading. The data consists of 20 messages per author for a total of 400 messages. The Arabic dataset was extracted from Yahoo groups and is also composed of 20 authors and 20 messages per author. These authors discuss a broader range of topics including political ideologies and social issues in the Arab world. Based on previous studies, there are numerous classification techniques that can provide adequate performance. In this research, we adopted two popular machine learning classifiers; ID3 decision trees and Support Vector Machine (SVM). The Arabic feature set was modeled after the English feature set. It includes 410 features, with the key differences highlighted in Table 8-4.

The results for the comparison of the different feature types and techniques are summarized in Table 8-5 and Figure 8-4. In both datasets the accuracy kept increasing with the addition of more feature types. The maximum accuracy was achieved with the use of SVM and all feature types for English and Arabic. Using all features with the SVM classifier, we were able to achieve an accuracy level of 85.43% for the Arabic dataset; a level lower than the 96.09% achieved for the English dataset.

Table 8-4. Differences between English and Arabic feature sets.

Feature Type	Feature	English	Arabic
<b>Lexical, F1</b>	Short-Word Count	Track all words 3 letters or less	Track all words 2 letters or less
	Word-Length Distribution	1-20 letter words	1-15 letter words
	Elongation	N/A	Track number of elongated words
<b>Syntactic, F2</b>	Function Words	150 words	250 words
	Word Roots	N/A	30 roots
<b>Structural, F3</b>	No Differences	-	-
<b>Content Specific, F4</b>	Number of words	11	25

Table 8-5. Accuracy for different feature sets across techniques.

Accuracy (%)	English Dataset		Arabic Dataset	
	C4.5	SVM	C4.5	SVM
<b>F1</b>	86.98%	92.84%	68.07%	74.20%
<b>F1+F2</b>	88.16%	94.00%	73.77%	77.53%
<b>F1+F2+F3</b>	88.29%	94.11%	76.23%	84.87%
<b>F1+F2+F3+F4</b>	89.31%	<b>96.09%</b>	81.03%	<b>85.43%</b>

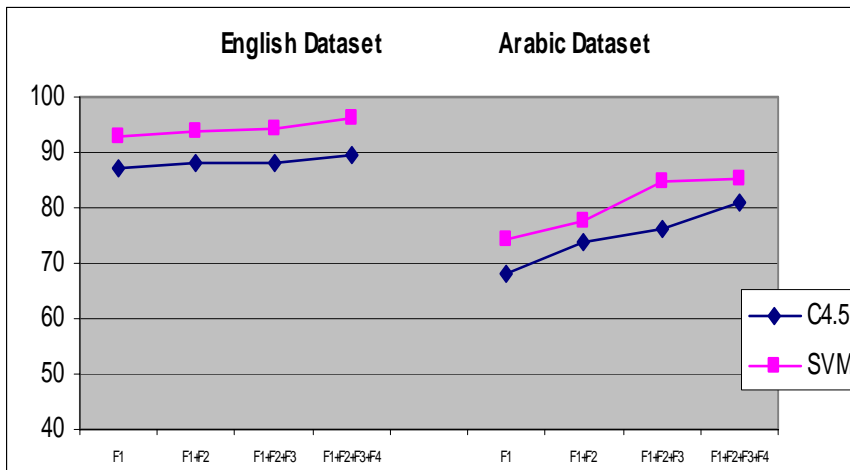


Figure 8-4. Authorship identification accuracies for different feature types and techniques.

A comparison of C4.5 and SVM revealed that SVM significantly outperformed the decision tree classifier in all cases. This is consistent with previous studies that also showed SVM to be superior. The difference

between the two classifiers was consistent across English and Arabic, with English accuracies being about 10% higher.

In the future we would like to analyze authorship differences at the group level within a specific language. Identification of unique writing style characteristics for speakers of the same languages across different geographic locations (e.g., Iraq vs. Palestine), cultures (e.g., Sunni vs. Shiite), and interest (e.g., terrorist) groups could prove to be an interesting endeavor.

## **8.4 Future Directions**

Cyber-infrastructure protection and cyber-trust are some of the most pressing ISI research topics of relevance to IT researchers. Unlike traditional critical infrastructure (e.g., highways, bridges, dams, and etc.), cyber-infrastructure can be attacked from any part of the world. The openness of the Internet protocol also invites unwanted and unforeseeable intrusions and disruptions. International terrorists and criminals have long been using the Internet for their illegal and covert activities. Much ISI research is needed in intrusion detection, computer forensics, Internet identity frauds, and grid computing and sensors in the next decade.

## **8.5 Questions for Discussion**

1. What are the research opportunities in NSF cyber-infrastructure and cyber-trust programs? What are some ways to leverage national security research opportunities in these areas?
2. How can grid computing and sensors be incorporated into ISI research?
3. How can multilingual research be applied in cyber-infrastructure protection?
4. What is computer forensics? How can it be applied to the Internet and cyberspace? What are the research opportunities?

