Research on Email Author Identification Algorithm

Kai Wang, Liangyue Ni, Jianyong Li, Changzheng Zhao and Kunpeng He

School of Shandong University of Science and Technology, Shandong 266590, China.

1838975646@qq.com

Abstract

Email has become one of the most basic and important applications on the Internet. However, the use of e-mail fraud, reactionary propaganda and other crimes are also increasingly serious. Therefore, this paper adopts the method of researching the identities of e-mail authors to identify the true identity of e-mail authors and provide evidences for computer forensics. Using the feature extraction method, that is to say, the similarity is calculated by analyzing the language features, structural features and format features of the email author, and then the list of the largest suspects is listed through the set algorithm. Finally, the algorithm Function and get the result. The validity of the proposed scheme, method and algorithm are verified through experiments.

Keywords

Email; computer forensics; similarity; feature extraction method.

1. Introduction

E-mail author identification is the basis of computer forensics. It can provide technical support for e-mail as electronic evidence and is crucial for prosecuting illegal e-mail crimes [1]. It also plays an active role in cracking down and deterring e-mail crimes and maintaining social stability and security. So it has extremely important realistic meaning and social influence. Most of the existing research is based on machine learning [2], but none of the studies attempted to identify e-mail authors' research by e-mail similarity.

2. Models

2.1 Terms, Definitions and Symbols

T_{ij}: The j document of the i suspect;

$Q_i = \left\{T_{i1}, T_{i2}, \cdots T_{in_i}\right\}$: The i suspect's file collection $j = 0,1,\cdots,n_i$

X_k: The k test file, $k = 0,1,\cdots,K$

S : Similarity

W: Feature weight

2.2 Assumptions

According to the characteristics of the e-mail author identify the problem to make the following assumptions:

1) An illegal mail author's mail type is basically the same. For example, fraudulent mail, mail samples are basically fraudulent mail. The author's sample email is large enough, even if it is not of the same type.

2) For the number of personal files is always enough both the total optimal value

2.3 The Foundation of Model

The overall system modeling, the first number of suspects entered into the system, the system will establish a collection, and then input files to the system, and then enter the test files, test files in turn with the collection of libraries for comparison, the final output of the similarity list. The total system flow chart shown in Fig. 1.

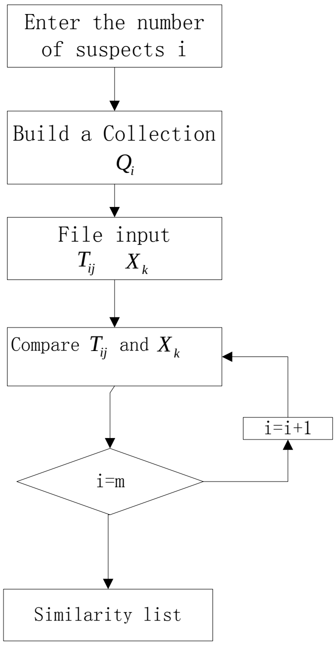
Enter the number of suspects i
Build a Collection Q_i
File input T_{ij} X_k
Compare T_{ij} and X_k
i=i+1
i=m
Similarity list

Fig 1. Total system flow chart

The similarity calculation is performed by the method of feature extraction and representation. The following three sections describe this part in detail.

2.3.1 Language Features

This study uses a feature extraction method for mail texts. Taking into account the e-mail itself shorter than the normal length of the text, the usage of some of the function words exactly reflect the author's habit, so all the words extracted e-mail documents constitute the characteristics of items. Get the weight of each feature item for each document. In order to reflect the distinctiveness of a feature word to this document, this study uses a well-known TF-IDE formula to calculate the weight of feature items.

$$W(t,d)=\frac{f(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}\left[tf(t,d)\times\log(N/n_t+0.01)\right]^2}} \tag{1}$$

$W(t,d)$ is the weight of word d in text $f(t,d)$, and d is the word frequency in text N, N is the total number of training texts, n_t is the number of texts in the training text set that appear in t, and the denominator is the normalization factor.

2.3.2 Structural Features

The mail document is an informal form of text. Its greatest feature is the freedom of text forms, short sentences and paragraphs, and more blank lines in the middle of the text. This study extracted 7 more obvious structural features of mail documents, as shown in Table 1 [3].

Table 1. Structural features

Feature number		Feature description
1		Average sentence length
2		The average paragraph length
3		Number of blank lines
4		Number of spaces
5		Word Recurrence Rate
6		The ratio of numbers
7		Punctuation ratio

For W_f, the Eigen values calculated one by one, in order to ensure the same treatment of all the features in the classification process, the introduction of scale factor $T_{i,\min}$ and T_i are respectively the minimum and maximum of the features, LB_{ri} and BY_i are defined as the lower and the upper limit, Then:

$$W_f = \left(T_i - T_{i,\min}\right) BY_i + LB_{ri} \tag{2}$$

2.3.3 Format Features

E-mail format features randomly generate many patterns, the weight of a pattern for a mail using Bayesian formula to represent the formula is as follows:

$$W_k = P(\mathrm{W}_k \left| C_j \right.) = \frac{\sum_{i=1}^{|D|} N\left(W_k, d_i\right)}{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N\left(W_k, d_i\right)} \tag{3}$$

$P(\mathrm{W}_k \left| C_j \right.)$ is the weight of the pattern W_k appearing in the mail C_j, D is the training text number of the class, $N\left(W_k, d_i\right)$ is the number of times that the pattern W appears in d_i, V is the total number of patterns, $\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N\left(W_k, d_i\right)$ is the sum of the occurrences of all modes of this class.

Reference information gain algorithm, given the similarity formula is:

$$S = \sum_{i=1}^{m} W\left(t, d\right)_i + \sum_{i=1}^{m} W_{fi} + \sum_{i=1}^{m} W_{ki} \tag{4}$$

3. Solution and Result

The research uses C language to implement the above algorithm. The main principle is to input the dataset first, and then to test and compare. Finally, the highest similarity of the final result is selected by the similarity calculation of eigenvalue.

Using text categorization to study generally accepted measures of evaluation to evaluate the performance of the email author's identity, i.e. accuracy. Fifty e-mails from five authors were used as data sets and 9 e-mails written by 5 individuals were tested. The test accuracy results are shown in Table 2. The data from the table shows that the algorithm has greater accuracy and reliability.

Table 2. Test accuracy

Personnel number	The correct number	Accuracy
1	8	0.89
2	6	0.67
3	9	1
4	7	0.78
5	7	0.78

4. Conclusion

Through the above research draw the following conclusions. Based on the analysis of the author's writing style, this paper presents a method to express the author's writing style by language features, structural features and format features. The characteristics of the identity of the e-mail signatures are high, the sample is relatively small and so on, using the feature algorithm and the similarity algorithm combined to identify the author's identity. Studied the problem of attribution verification of new e-mail authors, and set up a model to verify and verify the new e-mail.

References

[1] Chang Shuhui. Based on the writing style of Chinese e-mail author identification technology [D]. Hebei Agricultural University, 2005.

[2] Joshi P, Agarwal A, Dhavale A, et al. Handwriting Analysis for Detection of Personality Traits using Machine Learning Approach[J]. International Journal of Computer Applications, 2015, 130.

[3] Teng G F, Lai M S, Ma J B, et al. E-mail authorship mining based on SVM for computer forensic [C]// International Conference on Machine Learning and Cybernetics. IEEE, 2004:1204-1207 vol.2.