vocabulary richness, length of sentence, use of function words, layout of paragraphs, and

keywords [1]

commonly used features are lexical, syntactical, and structural and content-specific

attributes.

This approach was developed by R.Zheng et al. (2006) [3]

Write printi for intellectual property checking and plagiarism detection.

introduced by the H. Binsalleeh et al. (2010) [6]. This proposed method first clusters the given

anonymous e-mail based on the stylometric features and then extract unique writing styles from each cluster. The

writing styles in term of feature patterns provide more concrete evidence than producing some statistical numbers.

[1] J. Li, R. Zheng, and H. Chen, “From fingerprint to writeprint”, Communications of the ACM (2006); 49(4): pp.76-82.

[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 2006; 57(3):

pp.378-393.

[6] F. Iqbal, H. Binsalleeh, B.C.M. Fung, and M. Debbabi, “Mining writeprint from anonymous e-mails for forensic investigation”,

Digital Investigation, October 2010; 7: pp.56-64.

# Mining E-mail Content for Author Identification Forensics

These authorial features are

examples of stylistic evidence which is thought to be useful in establishing the authorship of a text document. It

is conjectured that a given author's style is comprised of

a number of distinctive features or attributes sucient to

uniquely identify the author. Stylometric features (\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.. However, some of these stylometric

features could be generated under the conscious control of

the author and, consequently, may be content-dependent

and are a function of the document topic, genre, epoch etc..

Rather than using content-dependent features,

e 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

di  
erent contexts.

However, as stated previously, certain characteristics such as

particular syntactic and structural layout traits, patterns of

vocabulary usage, unusual language usage, stylistic and substylistic features will remain relatively constant for a given

e-mail author

To evaluate the categorisation performance on the e-mail

document corpus, we calculate the accuracy, recall (R), precision (P) and combined F1 performance measures commonly employed in the information retrieval and text categorisation literature (for a discussion of these measures see,

for example, [40]), where:

table with features included

Style Marker Attribute Type

Number of blank lines/total number of lines

Average sentence length

Average word length (number of characters)

Vocabulary richness i.e., V=M

Total number of function words/M

Function word frequency distribution (122 features)

Total number of short words/M

Count of hapax legomena/M

Count of hapax legomena/V

Total number of characters in words/C

Total number of alphabetic characters in words/C

Total number of upper-case characters in words/C

Total number of digit characters in words/C

Total number of white-space characters/C

Total number of space characters/C

Total number of space characters/number white-space characters

Total number of tab spaces/C

Total number of tab spaces/number white-space characters

Total number of punctuations/C

Word length frequency distribution/M (30 features

# Mining E-mail Authorship

# Olivier de Vel

These author

categories have certain characteristic features which enables

better discrimination. Discriminating features for the different authors include the existence of tabs, number of lines

in the e-mail documents and the function word attributes.

Other features such as the \ratio of short words to total

number of vocabulary words" or the \ratio of words used

once to total number of vocabulary words" did not provide

for any discrimination between the author categories. We

suspect that other statistics of these features (e.g., the variance) may be more appropriate.

We also obtained values for the features for all e-mail documents for each author category to identify some of the ensemble author characteristics (see Table 5 for a subset of the

features). We observe some signicant in-between author

category di  
erences with some of the features, particularly

with some of the function word attributes. This indicates

that some features could potentially be strong discriminators of the author categories (e.g., ratio of the frequencies of

the words \a" and \all") when larger document populations

are considered