Abstract

Fraudulent emails can be detected by extraction of authorship information from the contents of emails. This paper presents information extraction based on unique words from the emails. These unique words will be used as representative features to train Radial Basis function (RBF). Final weights are obtained and subsequently used for testing. The percentage of identification of email authorship depends upon number of RBF centers and the type of functional words used for training RBF. One hundred and fifty authors with over one hundred files from the sent folder of Enron email dataset are considered. A total of 300 unique words of number of characters in each word ranging from three to seven are considered. Training and testing of RBF are done by taking different lengths of words. Our simulation shows the effectiveness of the proposed RBF network for email authorship identification. The accuracy of authorship identification ranges from 95% to 97%.

References

- David Madigan, Alexander Genkin, David Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye, &quot;Author Identification on the Large Scale&quot;, Proc. of The Meeting Of The
Classification Society of North America, 2005.


- Farkhund Iqbal, Hamad Binsalleeh, Benjamin C. M. Fung, Mourad Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation, Digital Investigation,1 – 9 (2010).

Index Terms

Computer Science Security

Keywords

Email Authorship Identification Spam Word Frequency Radial Basis Function