Abstract

This paper attempts to develop an algorithm to recognize spam domains using data mining techniques with the focus on law enforcement forensic analysis. Spam filtering has been the major weapon against spam, but failed to reduce the number of spam emails sent to an indiscriminate set of recipients. The proposed algorithm accepts as input, spam mails of personal account and extracts features such as stylistic, semantic, related email subjects and URLs present in the emails. The individual features are then clustered and evaluated. Further, these clusters are mapped with their respective domains. These spam domains are the URL of the webpage that spammer is trying to promote. The WHOIS information of the domain helps to get information about the source of that domain. Parameters like overall purity and the number of emails present in the cluster with highest purity is used to measure result of the individual features. An Experimental result shows that clustering of spam mails by stylistic and semantic parameter 20% less pure than other two features of spam mails.

References

- C. Wei, A. P. Sprague, G. Warner, and A. Skjellum. “Clustering spam domains
- All Spammed up, Anti-spam in a business environment. http://www. allspammedup.com/2012/05/the-cost-of-spam-is-rising/
- Anirudh Ramachandran and Nick Feamster Understanding the Network Level Behavior of Spammers &quot;, 2006, Georgia Tech.
- Marios Kokkodis and Ting-Kai Huang, An empirical study of spam and spammers behaviour; 2006, University of California, Riverside.
- G. Warner A. P. Sprague and C. Wei; Clustering malware-generated spam emails with a novel fuzzy string matching algorithm;&quot;, In Proc. of SAC &apos;09. Honolulu, Hawaii, U. S. A.
- C. Wei, A. P. Sprague, G. Warner and Anthony Skjellum; Mining Spam Email to Identify Common Origins for Forensic Application;&quot;, SAC &apos;08, March 16-20, 2008, Fortaleza, Ceará, Brazil. Copyright 2008 ACM 978-1-59593-753-7/08/0003
- C. Wei, A. P. Sprague, G. Warner and Anthony Skjellum; Identifying New Spam Domains by Hosting IPs: Improving Domain Blacklisting;&quot;, Copyright 2006 ACM 238-7-59463-783-7/08/0007
- Spamhaus DBL. http://www. spamhaus. org/dbl/
- SURBL. http://www. surbl. org
- URIBL. http://www. uribl. com

Index Terms

Computer Science
Security
Keywords
Spam  Semantics  Stylistics  Data Mining  Clustering