Abstract

Electronic mail is one of the important means of communication. Thus, this useful tool has invaded by invaders for different purposes. One such Invasion is the posting of useless, unwanted e-mails known as spam or junk e-mails. Several methods of spam detection exist, but each has certain weaknesses. This paper address these weaknesses by implementing and describing a spam detection system in text classification mode, which uses Bayesian method vs. PCA to filter out written spam mails from the user’s mail box. In the proposed method first extract all tokens that exist in body of emails for classifying emails based on them. But sum of these tokens aren’t useful. Sum of them are repeated in two categories spam and non-spam mails equally, so they aren’t appropriate for distinguishing two types of emails. So proposed method finds best tokens as main features using feature selection methods such as genetic
algorithm (GA), forward and backward feature selection methods.

References


2. J. Dudley, “Improving the Performance of Heuristic Spam Detection using a
Multi-Objective Genetic Algorithm”, School of Computer Science and Software Engineering, The
University of Western Australia, 2007


4. R. Zitar, A. hamdan, "Genetic optimized artificial immune system in spam detection",

5. T. Liang, Y. Pi, "On Spam Detection Based on Cognitive Pattern Recognition" ,
International Conference on Computational Intelligence and Security Workshops, 2007


11. Y. Begriche, H. Labiod, “A Posterior Distribution for Anti-Spam Bayesian Statistical
Model”,Network and Information Systems Security (SAR-SSI), pp. 1-6, 2011

e-mail”, In AAAI-98 Workshop on Learning for Text Categorization,1998.

13. A. Gray, M. Haahr ,“Personalised, collaborative spam filtering Using E-Mail Networks ”,
Fourth Conference on Email and Anti-Spam, 2007

14. J. JUNG, E. SIT  “An empirical study of spam traffic and the use of DNS black lists”, In


17. G. Kunzmann, A. Binzenhoefer, “Autonomically improving the security and robustness of
structured P2P overlays”, In International Conference on Systems and Networks
Communications ,2006.

18. A. Ramachandran, D. Dagon, N. Feamster, “DNS-based blacklists keep up with bots”, In
Third Conference on Email and Anti-Spam, 2006.

19. I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and

20. J. Liu, Y. Xiao, K. Ghaboosi, H. Deng, J. Zhang, "Botnet: classification, attacks,
detection, tracing, and preventive measures," EURASIP Journal on Wireless Communications

Ensemble Based on Decision Induction”, ISDA ’06. Sixth International Conference on , PP.
Index Terms

Computer Science

Artificial Intelligence

Keywords

Spam, Electronic Emails, Genetic Algorithms, Text Classification, Forward, backward, feature selection, Naïve Bayesian