Jordan University of Science and Technology

Web Spam Detection Using Machine Learning in Specific Domain Features

Authors: Hassan Najadat, Ismail Hmeidi

Abstract: Abstract: In the last few years, as Internet usage becomes the main artery of the life's daily activities, the problem of spam becomes very serious for internet community. Spam pages form a real threat for all types of users. This threat proved to evolve continuously without any clue to abate. Different forms of spam witnessed a dramatic increase in both size and negative impact. A large amount of E-mails and web pages are considered spam either in Simple Mail Transfer Protocol (SMTP) or search engines. Many technical methods were proposed to approach the problem of spam. In E-mails spam detection, Bayesian Filters are widely and successfully applied for the sake of detecting and eliminating spam. The assumption that each term in the document contributes to the filtering task equally to other terms and the avoidance of user's feed back are major shortcomings that we attempt to overcome in this work. We propose an improved Na?ve Bayes Classifier that gives weight to the information fed by users and takes into consideration the existence of some domain specific features. Our results show that the improved Na?ve Bayes classifier outperforms the traditional one in terms of reducing the false positives and the false negatives and increasing the overall accuracy. Keywords: Web Spam, Na?ve Bayes, Term Frequency Matrix (TFM), Confusion Matrix (CM)