Abstract

E-mail is one of the most popular and frequently used ways of communication due to its worldwide accessibility, relatively fast message transfer, and low sending cost. Nowadays, detecting and filtering are still the most feasible ways of fighting spam emails. There are many reasonably successful spam email filters in operation. The identification of spam plays an important role in current anti-spam mechanism. For improving the accuracy of spam detection, an improved Filtering technique is presented which is based on the Improved Digest algorithm and DBSCAN clustering algorithm. Using this technique, mails are represented using improved digest algorithm and then clustered using DBSCAN clustering algorithm. All similar emails which always categorized as spam are identified and clustered together where good mails that don’t look similar like other mails are not clustered. This method greatly improves the filtering accuracy against latest proposed algorithms by 30% and improves the resistance of spam detection against increased obfuscation effort by spammers, while keeping miss-detection of good emails at a similar level of older filtering methods.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, 1996, "density-based spatial clustering of applications with noise - DBSCAN";
- Slavisa Sarafijanovic, Sabrina Perez, JeanYves Le Boudec, 2008, "Improving Digest Based Collaborative Spam Detection"; MIT Spam Conference.
- J. Han and M. Kamber, 2001, "Data Mining: Concepts and Techniques"; Morgan Kaufmann Publishers, SanFrancisco, CA,

Index Terms

Computer Science Security

Keywords

DBSCAN Digest Based Nilsimsa spam clustering