Abstract

Spam refers to unsolicited, unwanted and inappropriate bulk email. Spam filtering has become conspicuous as they consume a lot of network bandwidth, overloads the email server and drops the productivity of global economy. Content based spam filtering is accomplished with the help of multiple pattern string matching algorithm. Traditionally Aho Corasick algorithm was used to filter spam which constructs a trie of the spam keywords. The performance degrades in the context of time as well as space as the size of trie increases with the growing spam keywords count. To counterbalance time and space loss, bit parallel multiple pattern string matching algorithm using Shift OR method is used. The method acts as filter performing approximate string matching. This implies that there are some false matches detected by the filter which requires verification. The proposed method for filtering spams has been developed using a combination of Shift AND and OR operation. The method directly works on spam keywords of equal size whereas for unequal size keywords, a new proposed equal size grouping method is developed. Both method shows improvement over the Aho Corasick algorithm in context of space complexity and also behaves as an efficient filter and reducing the number of false matches as present in Shift OR method.
References

- "The Two Percent Solution" by Jim Turley 2002
- Dylan Mors and Dermot Harnett (2009)

Index Terms

Computer Science Security

Keywords

Spam filtering String matching Bit parallelism