Abstract

Email filtering is way to seperate good email from useless email. Problem of email filtering is Text categorization (TC) which have only two classes. They are spam class and legitimate/ham/nonspam class. Nonspam email is important email that not dissapointed for someone who receive that email, but spam (Stupid Pointless Annoying Messages) is unimportant email that disrupt because it uses many of space memory in computer and if children get this kind of email, they can access uneducated email such as file pornography. One of the method for handling spam email is Statictical Filtering. This filtering method is need to be trained firstly uses two email collection, first collection is spam email and the other collection is nonspam email. With this method, Statictical Filtering predicts spam probability based on words which is usually current in spam email collection or legitimate/nonspam email collection for every new email.

Markov Random Field is one of the method that used statistical filtering method, not only count words but also phrases. Relation among words is important and it will makes phrases based on its neighbourhood size. Forming words and pharses is by *Sparse Binary Polynomial Hashing* (SBPH) method. These words and phrases are called features. Each feature will be weighted using *Exponential Weighting Sequences* or *Exponential Series* (ES) and *Minimum Weighting Sequences* (MWS). We also need to look neighborhood relation among all features in an email. Features which belong to a neighborhood is called cliques.

The good size of neighborhood which can give the best accuration is found in 5 or 6 when use ES weighting which reach 86.67% accuration at 0.9090 threshold. Some parameters beside accuration had been tested are spam precision, nonspam precision, spam recall, nonspam recall, and f-measure. Refers to the result of the experiment, MRF is proved that it can make the classification result be better and give the good result in accuration.

Keywords: *email filtering, statistical filtering, feature, Markov Random Field (MRF), SBPH, ES, MWS, neighborhood, cliques.*