SpamBlazer

Best Anti Spam Software in the market

Why SpamBlazer
============

Electronic mail is an effective, fastest, most-economical and popular form of communication by the virtue of its reliability, high-speed and low-cost. However, its ubiquity is at stake because of the prevalence of massive amount of unsolicited commercial email messages referred to as spam. After many years of research and development, the computer industry has not been able to find effective means of combating spam. As late as the end of November and early December of 2006, major news media – CNN, NY Times, CNBC et al - broke the unsavory news that "9 out of 10 e-mails now spam". Below are some of the problems with spam:

1. Cluttering of user' inbox with unwanted increasing volumes of unwanted email.
2. Inundating children with adult materials – enormous cost to our future generations.
3. Crashing of mail servers.
4. Wasting of mail storage severs – this is a major problem at large sites (ISPs, Corporations, Colleges, etc) with thousands of users.
5. Wasting of network bandwidths.
6. Drowning of legitimate email in the ocean of spam messages.
7. Scamming vehicles for get-rich-quick crooks.
8. Consuming valuable user' time and energy to sort through it. Thereby precipitously reducing employees' productivity. The estimated cost as a result of wasted time caused by Spammers is in the billions of dollars.

Traditionally, there are two schools of thought in eviscerating spam messages. The first approach uses some form of rule based systems where complex rules are derived for the ever-changing format of spam messages. These methods are inherently tedious (hand-crafted), error-prone and intolerant to noisy data. These are susceptible to even the slightest change in spam messages. More recently, the second school of thought is awash with statistical learning, in particular Bayesian filters.

Bayesian filters are Bayesian networks - directed acyclic graphs - applied to classification problems, consisting of nodes that represent both classes and features of the classification task at hand. At the root nodes are the categories, intermediate and leaf nodes are denote features. Edges in the networks denote causal relationship from a parent to a child. Bayesian filters permit the assignment of degrees of beliefs to unobservable events, prior probabilities, and also the establishment of conditional probabilities of observable events, likelihoods, given unobservable events. Within a decision theoretic framework, applying Bayes' theorem, as defined below, Bayesian filters determine posterior probabilities for each of the unobservable events in the Bayesian reduced universe of the observed event and the one with the highest posterior probability represent the most plausible unobservable event in the reduced Bayesian universe. Mathematically, Bayes theorem for classification tasks is defined as follows:

$$PS\left(c_i \middle| \vec{m}_j\right) = \frac{JP\left(c_i \cap \vec{m}_j\right)}{MP\left(\ \vec{m}_j\right)} \qquad (1)$$

$$JP\left(c_i \cap \vec{m}_j\right) = PR\left(c_i\right) \times LI\left(\vec{m}_j \middle| c_i\right) \qquad (2)$$

$$MP\left(\vec{m}_j\right) = \prod_{i=1}^{i=n} \left\langle PR\left(c_i\right) \times LI\left(\vec{m}_j \middle| c_i\right)\right\rangle \qquad (3)$$

The above equations according to Bayes say that the posterior probability is the product – the joint probability - of the prior probability and the likelihood function divided by the sum of the products of prior probabilities and likelihood functions, the marginal probability. From the above equations the unobservable events are denoted by c_i that represent the category of the observed message, for a two-class problem, it is either spam or legitimate email. The electronic message is represented by a vector, $\vec{m}_j$ whose element denotes appropriate proportions of selected features.

$PS\left(c_i \middle| \vec{m}_j\right)$, is the posterior probability of an unobserved event, the category of the electronic message given the observed data, the electronic message in the reduced Bayesian universal. $JP\left(c_i \cap \vec{m}_j\right)$, is the joint probability of the unobserved event in the reduced Bayesian universal. $PR\left(c_i\right)$, is the subjective belief for each of the unobservable event. $LI\left(\vec{m}_j \middle| c_i\right)$, is the likelihood function that denotes the occurrence of the observable event given that an unobservable event has been witnessed. And finally, $MP\left(\vec{m}_j\right)$, is the marginal probability in the reduced Bayesian universal, where the observed event was noticed; it is the sum of the likelihood functions over all plausible unobservable events.

The calculation of the posterior probability is not a trivial exercise, but it is impractical to determine the likelihood function. Take the simple case of assuming that $\vec{m}_j$ is a binary vector, the computation for the likelihood function is exponential; moreover, if there were enough computer resources for the computation, there is also the problem with sparseness of data, because in real life, the observed data are significantly smaller than the permissible permutations for the likelihood function.

As a result, all the Bayesian filters in the market are coerced to assume that the elements of the vector are independent there reducing the computation complexity. However, real life has showed that data features are overlapping and interdependent, thus these filters are fundamentally flawed. They are extremely weak in terms of generalizations, they are susceptible noisy data.

SpamBlazer is a proprietary solution – an amalgam of information theory, evolutionary programming (neural networks, genetic algorithms, simulated annealing, and fuzzy logic) and security engineering – that overcomes the problems associated with the abovementioned conventional techniques. SpamBlazer is a machine learning technique and it extracts pertinent rules automatically that dictate the legitimacy of an electronic email or otherwise. SpamBlazer understands the chameleonic nature of spammers and stays ahead of them by incorporating new data into its training set and discarding dated data. SpamBlazer inherently captures the overlapping trait and interdependency of features, so it is not blighted with problems associated with Bayesian filters.