A fusion approach to spam filtration
Ning Yu, D. Scott McCaulay, Nicolas George, and Aaron Loehrlein

Presents research by the WIDIT Lab for the Spam Track of the 2005 TREC conference. WIDIT created a filter that integrates several different methods for making a binary decision about an incoming email, i.e., whether it is spam or non-spam (referred to as a "ham"). The filter makes use of an adaptive learning process that learns the features of the test corpus in real time during the filtering process. Its methods include:

- A naïve Bayesian algorithm that continuously updates its understanding of spams and hams from each email that it analyzes

- A pattern-based filter that uses a scrolling window to identify sub-strings within emails that may identify it as a spam or a ham

- A rule-based based filter that evaluates an email according to a set of static rules devised by the WIDIT lab. For example, one of the rules states that an email is likely to be spam if its "To:" line is blank.

- Four blacklists that are initially built by querying the web for known spammers and will be updated later during the adaptive learning process. Domain name, email address and IP address appear in spam email header and URL appears in a spam email body are covered in those blacklists respectively.

- A "stop list", which contains domains, such as hotmail.com, that are known to be used by spammers and non-spammers alike, is built in the same fashion to avoid possible over tuning.

Each of these modules is governed by a fusion module, which determines how much influence each of the other modules should have in determining whether the incoming email is a spam. WIDIT found that the naïve Bayesian module was the most successful in identifying spam. However, only by taking into account information generated by each module could the filter achieve the best results, both in terms of fewest false negatives (spams misclassified as hams) and fewest false positives (hams misclassified as spams, a much more serious error).
References:


