CS545 Project Proposal: Comment Spam Identification

Eric Cheng (eric.cheng@yale.edu)
Eric Steinlauf (eric.steinlauf@yale.edu)
Yan Sui (yan.sui@yale.edu)

March 26, 2007

Traditional comment spam identification methods include blacklisting, whitelisting, and keyword filtering. These simple methods are easily evaded since spammers could obfuscate the messages by putting extra spaces or symbols, or by using dynamically generated addresses which point to the advertised sites. Statistical approaches to identifying email spam have worked well, such as DSPAM [1] and Spamassassin [2]. We would like to apply similar methods, such as Bayesian networks, to identify comment spam. The algorithm would take into account the user name and subject of a comment, in addition to its content.

References

[1] http://dspam.nuclearelephant.com/

[2] http://spamassassin.apache.org/