

SPAM - Different Approaches to Fighting Unsolicited Commercial Email

A survey of spam and spam countermeasures

Håvard Wik Thorkildssen

Abstract

This paper gives an overview of unsolicited email and the techniques available to defend oneself against it. It covers the most influential theories behind all major spam filtering software, including, but not limited to: simple pattern matching (pattern matching of words), statistical pattern matching (algorithms that perform heuristic tests on the header and text of the email, namely Naive Bayesian and $\frac{x}{d.f.}$), blacklists (Internet databases which are queried upon relaying an email and matched to a blacklist of known spammers), and collaborative spam-tracking databases (like Spamassassin).

1 Introduction

Unsolicited commercial email – or spam – has become a substantial problem for Internet citizens. Not only does it cause severe annoyance, it also reduces productivity, occupies valuable storage on mail-servers and clogs the Internet with junk[14]. Barry Shine, president of American ISP "The World"[16] states that 30% of staff expenses at his 20-person company is spent working on spam filters or talking to customers about spam. John Levine, cochair of the Internet Research Task Force's Anti-Spam Research Group (ASRG)[8] claims that European companies lose \$9 billion a year because of spam, primarily from lost productivity due to employees reading, manually classifying and deleting unsolicited email.

According to Keith C. Ivey[10], the word spam derived from a sketch in the British comedy show *Monty Python's Flying Circus*, where a group of vikings in a cafe sings loudly about spam, making it impossible for the other guests to chat and enjoy their meal.

Every email account holder will at some point receive a spam message, but since there has been quite a lot of media coverage about this issue, most people have learnt not to trust the contents of these emails. The spammers, who often prefer to call themselves *direct marketers*, have realized this, but since each message comes at virtually no cost, they can still make huge amounts of money on promoting questionable medical products, pornography, and get-rich-quick schemes. M. Mangalindan, from The Wall Street Journal Europe, claims that if a spammer gets as few as 100 responses for every 10,000,000 email, he can still make an attractive profit[11].

2 Spam I am?

Different types of spam

When reading literature on this field, you will find several, and diverging, definitions on what spam is. Do annoying chain-letters qualify as spam? What about an email claiming to have a miracle cure for hangovers? Ultimately, you may find that what qualifies as spam for some people, might be interesting reading for others. Therefore, one ought to take several definitions into account when designing a spam filter. Teachable spam filters, like the Naive Bayesian classifiers, have tried to do something about this problem.

In some literature you will find the term "unsolicited bulk email", which is basically a wider definition of spam, that also covers non-commercial unsolicited email. According to Keith C. Ivey[10], the problems caused by spam have nothing to do with it being commercial. Unsolicited mass mailings are harmful to the Internet regardless of their content – whether it's advertising for cheap Viagra or claims it has the answer to Life, the Universe and Everything.

Case studies

In a case study performed by Lorrie Faith Cranor and Brian A. LaMacchia[4], only 3% of unsolicited bulk email was non-commercial. Therefore, it does not contribute nearly as much to the spam problem as the the adult entertainment industry, with its 11%, and the innumerable (Nigerian) money making schemes at 35(!)%. These money making schemes are primarily pyramid-style schemes, multilevel marketing systems, and investment opportunities. Another form of mass emailing that has flooded the Internet is jokes, like the so-called "Amish Virus"-joke mentioned in Eddie Rabinovitch's paper[17]. This was a harmless, non-commercial email that spread rapidly on the Internet several years ago. It worked by the same principle as chain letters. The message contained this text: "You have just received the 'Amish Virus'. As we don't have any programming experience, this virus works on the honor system. Please delete all the files from your hard drive and manually forward this virus to everyone on your mailing list. Thanks for your cooperation." Signed: Amish Computer Engineering Department. Hopefully, most people did not fall for this joke and actually deleted files from their hard drive. It nevertheless created a burst in traffic on Internet mail servers, simply because people thought it was funny. Another example of such mass mailings, which created a similar burst in traffic on Internet mail servers, is the numerous Nostradamus-inspired conspiracy theories that flourished on the Internet some time after the Al-Quida attack on the World Trade Center. These letters claimed that Nostradamus had foreseen the events of 9-11 2001, but they later turned out to be hoaxes.¹

3 Technical and practical challenges

Identification

Jim Dennis[19] states in the paper "Stop, in the Name of Spam" that the most important technical problem to overcome is that of identification. Most bulk email is sent from fictitious addresses and domains, so banning mail that

¹<http://urbanlegends.about.com/cs/historical/a/nostradamus.htm>

seemingly comes from a particular domain is therefore in vain. In addition, most of them also have forged headers, so tracing it back to its source is virtually impossible, further complicating the problem of identification. This also raises an interesting question regarding anonymity, issues and concerns that are analogous with those of caller ID in telephone networks. Should one really be able to trace all email to its sender, or should it, like the rest of the services on the Internet, simply be built on trust?

Erroneous classification

Since there currently is no obvious solution to the problem of identification, sophisticated algorithms that distinguish between legitimate mail and spam have been developed. The most significant challenge in effective spam filtering is erroneous classification of spam and legitimate email. Classification of spam can be divided into four categories: true positives, true negatives, false positives and false negatives. The true ones are, as the name implies, correctly classified email, i.e. the spam filter has correctly identified the nature of the email, whereas the false ones are erroneously classified email. Although both of the false classifications are erroneous, the misclassification is not equal in terms of *cost*. This term will be elaborated later in this paper.

False negatives occur when a spam email is not classified correctly. If the number of false negatives are at a bare minimum, they are only slightly annoying. False positives occur when desired email is classified as spam because the email has certain spam-like characteristics, e.g. if a message is sent to many recipients at the same time. They are at best only annoying. However, if you happen to be really unlucky and the spam filter is configured to delete spam immediately, you risk losing important information.

Drifting class distribution

In Naive Bayesian classifiers, the drifting in class distribution, i.e. the drifting in distribution between spam and legitimate email, will decrease the efficiency of the algorithms. The variation was, in data-sets reviewed by Tom Fawcett[6], between 16.6 and 88.2 percent, and these classifiers will need to be optimized to handle such driftings in the future.

There is also drifting in the amount of spam email account holders receive. The fact that some mail accounts receive more spam than others, depends on several factors, like the exposure of the address and the predictability of the address. The spammers have developed sophisticated address harvesting software that scans web pages, newsgroups and forums for addresses. An email account used on newsgroups are, therefore, more likely to be attacked by spam than one that is only spread to friends and relatives for private use[15]. As a countermeasure, users have started scrambling their email accounts², making it unreadable to harvesting agents. This process is often referred to as "munging", originally an acronym for "Mash Until No Good". Some have even started using fictitious email addresses in these arenas.

The total amount of spam also varies over time, since spam tends to come in waves. Fawcett points out that in 2002, the increased number of open relays

²For example, user@REMOVEME.domain.com

and proxies in some Asian countries (primarily China and Korea) created a such a surge in the amount of spam received that some ISPs were forced to block all email coming from these countries for a brief amount of time[6].

Some sites, like SpamCop³ and Spam.abuse.net, monitor these driftings.

Unequal and uncertain error costs

The concept of *cost* was mentioned briefly above. By error cost we mean the importance of a misclassification. Paul Graham, one of the first to seriously experiment with Bayesian spam-filtering[20], states that: *False positives seem to me a different kind of error from false negatives. Filtering rate is a measure of performance. False positives I consider more like bugs. I approach improving the filtering rate as optimization, and decreasing false positives as debugging*[6]. By forcing the user to go through their spam inbox several times a day, he or she might reconsider the value of spam filtering. Subsequently, the error cost of a false positive must be assigned a higher value than a false negative.

Disjunctive and changing target concept

The themes of spam email are changing over time. Some topics, like pornography, are of course perpetual. However, by analyzing spam over time, researchers have found out that the themes follow trends in society as a whole. For instance, right after the outbreak of the second Gulf war, spammers started marketing the notorious "Iraq's Most Wanted" playing cards via spam. This campaign was in fact such a huge financial success, that the New York Times wrote a story about it. Tom Fawcett[6] uses the Nigerian money scams as an example. These emails, which are basically get-rich-quick schemes, where the senders claim to be responsible for huge amounts of money held in Nigerian bank accounts, have had several bursts in the last few years. This can probably be explained by many factors. However, Fawcett believes that these bursts were primarily caused by scammers entering in and out of prison.

Intelligent adaptive adversaries

Spammers have, over time, become increasingly sophisticated in terms of technology. Firstly, in the way they send out huge amounts of undetected email, and secondly, in their way of dodging filters. For example, in the early days of spam filtering, one would simply filter out all email containing strings like "pornography", "viagra", and other suspicious words. As a counter-countermeasure, the spammers began obscuring the contents of their emails by using bogus HTML tags and by replacing letters in the middle of words with resembling letters, such as replacing an "a" with an "á". To the human eye, these look almost identical, and, most importantly, the word remains understandable. However, to a pattern matching spam filter, it becomes a whole new word.

³<http://www.spamcop.net/spamstats.shtml>

4 Approaches

Legal countermeasures

A number of legal countermeasures to fight spam have been attempted throughout the last decade. These have obtained varying success. USA was one of the first to introduce laws against unsolicited commercial email. The "Netizens Protection Act of 1997" made spam subject to the same regulations as junk fax. This law involved banning senders of spam, and required senders of solicited email to identify themselves. This had a limited effect since most of the spam was either sent from countries outside the US and EU, or relayed through unprotected proxies in the same countries[16]. There are currently no federal legislations on this, so spammers operating from – or in – countries that do not have such laws can safely carry on with their business for now.

Another proposed solution involves pricing each email with a small fee[4]. The sum will be small, so that the average email user will not be stricken by it. However, bulk emailers sending out thousands of unsolicited email every hour will presumably find this fee impractical – and it will certainly make their financial model less profitable.

Primitive pattern matching

Researcher Tim Bass and Lt. Col. Glenn Watt[2], at Langley Air Force Base in Virginia, USA, made an early attempt to block out huge amounts of spam that were relayed through their SMTP servers. They assumed that most of the emails were coming from a small group of individuals, since the theme was somewhat similar. They contained pornographic material and bigoted hate-mail. However, since most of the mail headers were forged, it was nearly impossible to track them down. They formed the *Tiger Team*, a group of scientists and military engineers, dedicated to fighting the spammers. The team developed a model that formed the basis of a working prototype. It consisted of a queue piped through a processing filter, where the classification was done. Their implementation is far too simple to be used for anything serious nowadays. Nevertheless, they were the first to formalize this model – a model which was later to be adapted by all spam filtering software.

Over time spammers began to adapt to these simple pattern matching algorithms, by obscuring the contents of the emails. They did this by modifying keywords like "VIAGRA" and "EXTREME" to V*I*A*G*R*A, and E}{TREME making it harder, and in large scale – impossible, for the filters to stay current[20].

Blacklists and whitelists

A spammer "blacklist" is a database of IP addresses belonging to known spammers. There several of these lists⁴, and many SMTP servers query these servers and compare the IP address of the sender to the blacklists[20]. Most of them have some sort of timeout, to make sure that dynamically assigned IP addressed are not banned forever. A whitelist is, in a sense, the complete opposite. It is a personal list of non-spammers, for example, your colleagues and friends.

⁴<http://spam.abuse.net>, <http://www.cauce.org>, <http://www.junkemail.org>

The concept of blacklists has not yet proven to be very efficient. Spammers are known for changing IP addresses and ISPs frequently[13], and a large portion of the spam is sent through open relays that are not blacklisted, and thus disguising the original sender. Whitelists, on the other hand, are very hard to maintain over time, especially if you receive email from a great number of people.

Collaborative spam-tracking databases

There have been a few attempts on creating collaborative spam-tracking databases. Though this has not been covered scientifically, it ought to be mentioned seeing as it has been implemented in the most popular piece of spam filtering software, Spamassassin. It works by piping all received email through a filter that creates a footprint of the email. The footprint is then forwarded to the spam database server, where it is compared to other footprints. Based on this, it is acknowledged or rejected. Spammers responded to this by including random characters, words, or even small poems in the body to scramble the footprint, thus making it difficult to compare them to email in the spam database[5]. The author is continuously improving the footprint algorithm to face these new challenges. This approach, particularly in combination with other classifiers like the Naive Bayesian, is known for having a high success rate. There is, however, a drawback to this method. It is very CPU- and network intensive, and it takes time to process all the email. This is especially a problem for Internet Service Providers, where you have thousands of emails coming in every hour. This type of software is, therefore, not widely used on large mail servers.

“Three-way-handshake”

Androutopoulos et al[1] suggest that rather than deleting an email classified as spam by other filters, it could be bounced back to the sender with an apology containing a “private” email-address that has not been advertised to the public (e.g. not on web pages or on Internet newsgroups, which are frequently scanned by address harvesting agents). They further improve this method by extending it to contain some sort of riddle, preferably randomly generated. Subsequently, the sender has to answer this riddle in the subject field of the new mail, before the spam filter lets it through. This riddle will prevent automated robots from simply forwarding the spam to the new mail address, since spammers would then have to manually answer the riddles. This is a time-consuming and expensive process that spammers cannot afford.

The “three-way-handshake” methods can be used in conjunction with other forms of spam filtering software, like the Naive Bayesian classifier covered later in this paper.

Sender Policy Framework

SMTP has a rather big security flaw: Any client can assert any sender address, and this flaw has been exploited by spammers to forge email[12].

Sender Policy Framework, often referred to as SPF, attempts to close this loophole. If this hole is closed, spammers can easily be blocked by certain criterias. It works by forcing the connecting client to identify himself by sending domain, and thus eliminating the problem of spammers using fictitious domains

to send unsolicited mail. In SPF mode, the Mail Transfer Agent (MTA) uses SPF to verify the sender, more precisely the SMTP MAIL FROM address, during SMTP time, i.e. when the message is received by the MTA awaiting relay.

SPF is primarily an anti-forgery effort, which also has a few positive effects: It fights domain forgery by spammers and scammers and makes it easier to identify spam, Internet worms and viruses.

SPF was originally designed to prevent joe-jobs. A "joe job" is a spam run forged to appear to come from another innocent party, and the name comes from the first recorded incident⁵.

Greylists and DCC

In 2003 Evan Harris[9] announced a term he named greylists. Unlike blacklists, greylists do not absolutely reject mail, but requires mail from unfamiliar senders to be retransmitted by their ISP's SMTP clients. Mail from familiar senders is, however, passed and relayed immediately.

The theory is based on the fact that most unsolicited bulk mail is sent via open proxies and tailored spam software that do not involve proper mail transfer agents (MTAs). The idea is to temporarily reject mail from unfamiliar senders, and as all proper MTAs will repeat a transmission if it gets rejected, this will only affect improperly configured MTAs and spam software. The MTA specification, RFC 2821⁶, states that the sending MTA should retransmit 30 minutes or later after a failure. Spam sent through an open proxy, along with some viruses and worms, are not retransmitted.

DCC is a Free Software implementation of a greylist. In DCC, the sendmail milter interface, dccm, or the general MTA interface, dccifd, sends a request to the DCC server, greylist dccd. The requests contains a simple DCC body checksum of the message, as well as an MD5 checksum of the MD5 checksums of the IP address of the SMTP client sending the mail message, the envelope sender or MAIL FROM value of the message, and the recipient RCPT TO value of the message. If the combination IP address, sender, and recipient is familiar, the DCC client tells the MTA to accept the message. Otherwise the DCC client tells the MTA to embargo or temporarily reject the message.

A difference between this implementation of greylists and other implementations is that all, or part, of the IP address of the SMTP client can be optionally ignored. This feature may be useful for legitimate mail servers that shuffle messages among SMTP clients between retransmissions.

If the sending MTA persists and retransmits the message after the embargo but within the wait time, the triple, which consists of sender, IP address, and recipient, is added to the database. Unlike other implementations of greylists, this implementation requires that the retransmitted email is nearly identical to the original copy.

If the triple is ever associated with spam, it is deleted from the greylist database. This renews the temporary embargo for subsequent mail involving the triple.

⁵http://www.wordiq.com/definition/Joe_job

⁶<http://www.ietf.org/rfc/rfc2821.txt?number=2821>

Naive Bayesian classifiers

Paul Graham, mentioned earlier in this document, was actually not the first to look to Bayes for help in curbing the spam problem. That honor can be attributed to two sets of authors: Patrick Pantel and Dekang Lin, from the University of Manitoba and a team made up of Mehran Sahami of Standford university, and three microsoft researchers, Susan Dumais, David Heckerman and Eric Horvitz[20]. Graham later took the basic idea and created a fully working open-source implementation of it.

Naive Bayesian classifiers are now used in most popular spam filtering software, like Spamassassin, Bogofilter, Spamprobe, and others. N.B. has proven to be very effective, especially in its handling of false positives. In the paper Spam Filters: Bayes vs. Chi-squared by O'Brien[16] et al, it is referred to tests performed by Hidalgo, where 91.7% of the spam was correctly classified without discarding any legitimate email. Bayesian classifiers work by representing each email as a vector $x = \langle x_1, x_2, x_3, \dots, x_n \rangle$ where x_1, \dots, x_n are the values of attributes X_1, \dots, X_n , and each attribute represents a particular word occurring or not. Therefore, the attribute is set to 1, $x_i = 1$, when a given word occurs, and set to 0, $x_i = 0$, when it does not. An email with words matching the words in a category c will, therefore, be represented by a vector $\vec{x} = \langle 1, 0, 0, 1, 1, 1, 0, 1, 0 \rangle$ when the number of words is 9, $n = 9$. According to Bayes formula, the probability that the vector x_i belongs to the category c can be described by (1).

$$P(c|\vec{x}) = \frac{P(c) \cdot P(\vec{x}|c)}{\sum_k P(K) \cdot P(\vec{x}|k)} \quad (1)$$

Naive Bayesian classifiers differs from "real" Bayesian classifiers in the assumption that X_1, \dots, X_n are conditionally independent of category c . We can, therefore, change the above equation to (2).

$$P(c|\vec{x}) = \frac{P(c) \cdot \prod P(x_i|c)}{\sum_k P(K) \cdot \prod P(x_i|k)} \quad (2)$$

This is a important improvement, since it is much easier to calculate $P(x_i|c)$ than it is to calculate $P(\vec{x}|c)$. Now you will need to set a threshold for the probability, like 0.9, and categorize all email with a $P(spam|\vec{x}) \geq 0.9$ as spam, and all email with a $P(spam|\vec{x}) < 0.9$ as legitimate mail. Based on experience, few of the probabilities end up in the middle of the range, i.e. most of the spam has a $P(spam|\vec{x}) > 0.9$ and most of the legitimate email has a $P(spam|\vec{x}) < 0.1$ [7]. A moderate threshold, like 0.9 instead of 0.99 is an effective way of combating false positives, without multiplying the amount of false negatives.

Chi by degrees of freedom

Chi by degrees of freedom, or $\frac{\chi}{d.f.}$, was originally used for identifying authorship in computational linguistic circles, but was applied to spam filtering by Cormac O'Brien and Carl Vogel in the paper Bayes vs. Chi-Squared; Letters vs. Words[16]. According to this paper, the vast majority of unsolicited emails come from only a handful of people, about 150. If you analyze the textual footprint they create, you might be able to recognize them and subsequently block email matching similar footprints in the database. Note that this will require that spammers are not consciously changing their style of writing across texts.

Experiments have been carried out by Van Gijssel[18] on trying to identify European right-wing party manifestos, and by Vogel et al[3] on identifying attributed Shakespeare poems. Where Bayesian classifiers look at words, $\frac{x}{d.f.}$ looks at the entire email. These programs are, like Naive Bayesian classifiers, trained with data-sets containing both spam and legitimate email. The training files are concordanced: each file is indexed by its n-grams, with frequency counts. The type of n-gram to use (unigrams, bigrams, et cetera) will need to be specified. The program then calculates the similarity value between the created files in terms of n-gram frequency, by carrying out a $\frac{x}{d.f.}$ -test. This test consists of dividing the chi-square test value by the number of its degrees of freedom (number of n-grams minus one). The resulting values can later be used for identifying the writer of the text.

5 Conclusion

Spam is a continual nuisance and a large number of people are working on methods to keep it out of the user's inbox. However, the spam problem is growing exponentially and filters are having an increasingly hard time separating junk from legitimate email. Trends point toward using filters that provide identification on the SMTP layer (primarily SPF and DCC greylists), and statistical filters, like the Naive Bayesian, on the end-users email client. Most modern email software, like Outlook, Mozilla Mail, and Opera Mail, have some sort of statistical spam filter software built-in. They have a reputation for being fairly accurate, and complement nicely with server-side anti-spam software. However, this requires a shift in the behavioural pattern of the end-user: Users will have to manually look through the spam inbox every once in a while, to make sure no legitimate email is misclassified. Primitive pattern matching filters are now virtually obsolete, since spammers have found effective ways to evade these filters. Blacklists and whitelists in combination with collaborative spam-tracking databases are frequently used by multifunctional spam filters and they too have proven to be quite effective.

Despite the efforts of anti-spam researchers, spammers are working even harder at finding new ways of dodging the spam filters. Recently, viruses that function as open MTA proxies when infected have gained a lot of popularity among the spammers. These viruses spread through the Internet either via regular email, or more commonly, through security holes in the Windows operating system. If we do not find a feasible and effective way of combating spam, the email media will be rendered unusable and new technology will have to replace it. None of these papers claim to have a definite answer on how to eliminate spam entirely. However, they all offer feasible methods on how to limit the amount of spam.

References

- [1] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinou, and Constantine D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–167. ACM Press, 2000.
- [2] Tim Bass and Lt. Col. Glenn Watt. A simple framework for filtering queued smtp mail (cyberwar countermeasures).
- [3] Carl Vogel Cormac O'Brien. A forensic examination of a funerall elegy, 2003.
- [4] Lorrie Faith Cranor and Brian A. LaMacchia. Spam! *Commun. ACM*, 41(8):74–83, 1998.
- [5] Matt Bishop Daniel Faigin, Tasneem Brutch. Miracle cures and toner cartridges: Finding solutions to the spam problem, 2003.
- [6] Tom Fawcett. In vivo spam filtering: A challenge problem for kdd, 2003.
- [7] Kevin R. Gee. Using latent semantic indexing to filter spam. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 460–464. ACM Press, 2003.
- [8] David Geer. Will new standards help curb spam? *Computer (IEEE)*, 2004.
- [9] Evan Harris. The next step in the spam control war: Greylisting. *Unpublished manuscript*, 2003.
- [10] Keith C. Ivey. Information superhighway. *CAP*, 1998.
- [11] Elias Levy. Crossover: Online pests plaguing the offline world. 1(6):71–73, November/December 2003.
- [12] Wang Weinong Li Cheng. Internet mail transfer and check system based on intelligence mobile agents. *Proceedings of the 2002 Symposium on Applications and the Internet (SAINT'02)*, 2002.
- [13] Paul McFedries. Slicing the ham from spam. *IEEE Spectrum*, 2004.
- [14] Wyman Miles. A high-availability high-performance e-mail cluster. In *Proceedings of the 30th annual ACM SIGUCCS conference on User services*, pages 84–88. ACM Press, 2002.
- [15] Peter G. Neumann and Lauren Weinstein. Inside risks: spam, spam, spam! *Commun. ACM*, 40(6):112, 1997.
- [16] Cormac O'Brien and Carl Vogel. Spam filters: bayes vs. chi-squared; letters vs. words. In *Proceedings of the 1st international symposium on Information and communication technologies*, pages 291–296. Trinity College Dublin, 2003.
- [17] Eddie Rabinovitch. Securing your internet connection: A sequel. *IEEE Communications Magazine*, 2002.

- [18] Sofie Van Gijssel and Carl Vogel. Inducing a cline from corpora of political manifestos. In *Proceedings of the 1st international symposium on Information and communication technologies*, pages 297–303. Trinity College Dublin, 2003.
- [19] Various. Stop, in the name of spam. *Communications of the ACM*, 1998.
- [20] Steven J. Vaughan-Nichols. Saving private e-mail. *IEEE Spectrum*, 2003.