



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A NEW APPROACH FOR COLLABORATIVE SPAM FILTERING

Dr.A.Jebaraj Ratnakumar

Principal,

Apollo Engineering College, Chennai, India

Abstract: In response to the growing volume and variety of spam, spam filters have moved away from monolithic repositories situated on central servers towards dynamic knowledge bases located on local servers. The state of the art in spam filtering sees content-based filters tending towards collaborative filters, whereby email is filtered at the mail server using content-based techniques, with users feeding information back about false positives and false negatives. This feedback enables the spam filter to track concept drift in spam and to be retrained in the case of false positives. While these filters can achieve statistically impressive accuracy rates, they remain prone to false positives, i.e., the erroneous classification of a legitimate email as spam. Underlying the current generation of spam filters are three implicit assumptions that, in our opinion, directly contribute to the presence of false positives. To address these, we propose the concept of personalised, collaborative spam filtering using Bayesian Filtering and Email Spam Filtering. In this article we discuss how the spam spread through mails like Nigerian mails, MLM Mails and Promotional Mails.

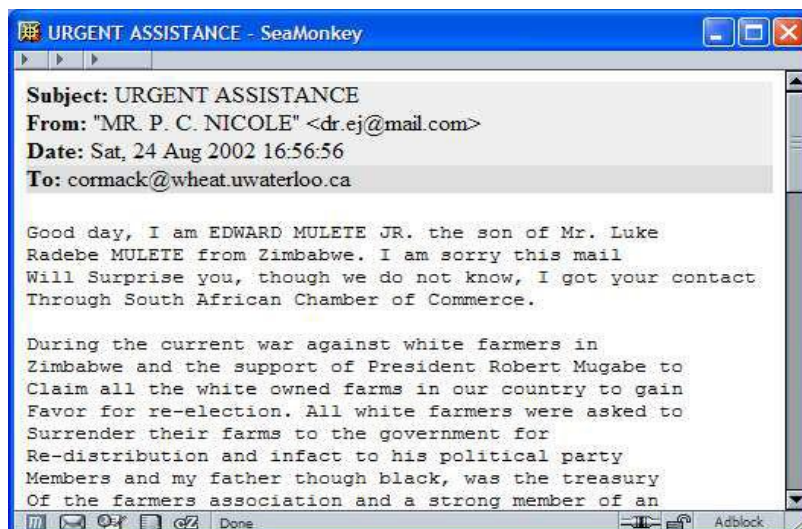
Keywords: Spam, Collaborative, Bayesian, Email, Data

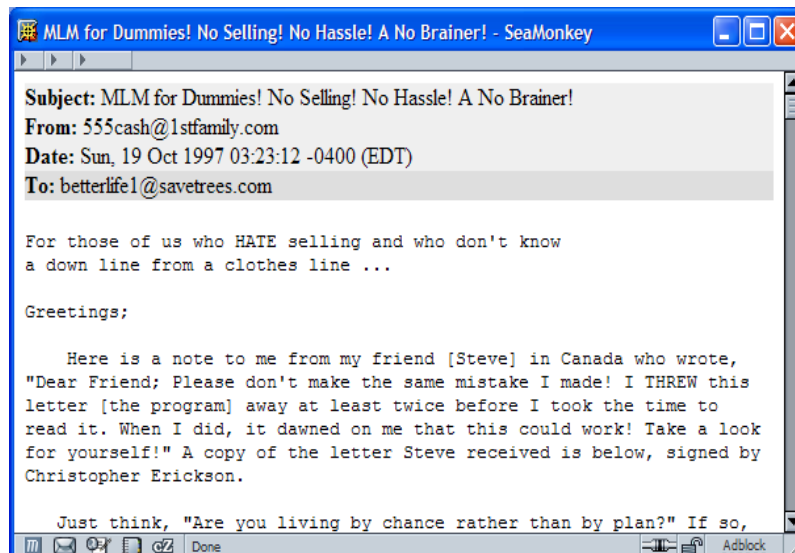
I. Introduction

Spam is usually sent in bulk. A bulk mailing consists of many copies of the same original spam message, each sent to a different recipient. The copies are usually obfuscated, i.e. modified a bit in order to look different from each other. In collaborative spam filtering it is important to determine which emails belong to the same bulk. This allows, after observing an initial portion of a bulk, for the bulkiness scores to be assigned to the remaining emails from the same bulk. This also allows the individual evidence of spamminess to be joined, if such evidence is generated by collaborating filters or users for some of the emails from an initial portion of the bulk. Then, the observed bulkiness and the estimated spamminess of a bulk can be used to better filter the remaining emails from the same bulk.

II. The Purpose of Spam

The motivation behind spam is to have information delivered to the recipient that contains a payload such as advertising for a product, bait for a fraud scheme, promotion of a cause, or computer malware designed to hijack the recipient's computer. Because it is so cheap to send information, only a very small fraction of targeted recipients — perhaps one in ten thousand or fewer — need to receive and respond to the payload for spam to be profitable to its sender. At the same time, spamming has become more specialized and sophisticated, with better hidden payloads and more nefarious purposes. Today, cottage industry spam has been overwhelmed by spam sent in support of organized criminal activity, ranging from traffic in illegal goods and services through stock market fraud, wire fraud, identity theft, and computer hijacking. Computer viruses are no longer the work of simple vandals, they are crafted to hijack computers so as to aid in identity theft and, of course, the perpetration of more spam.





Dear Lucky Winner,

This is to inform you that you have won a prize money of GBP £1,000,000.00 (**One Million Great British Pounds**) for the Yearly **Coca-Cola Lottery promotion United Kingdom** which is organized by Coca-Cola in conjunction with Chevron Texaco Oil and Gas Company, "Your Email Address" were selected through a computer ballot system drawn from Nine hundred thousand email from Canada, Australia, United States, Asia, Europe, Middle East, Africa and Oceanic as part of our international promotions program which is conducted annually, we have only selected 20 people as our winners, through electronic ballot System without the winner applying.



III. Spam Types

Spam techniques can be categorized in several ways. Based on the purpose, spam techniques can be categorized into two categories. Boosting is the technique that is applied in order to increase the relevance of pages undeservedly which is the main purpose of spamming. Hiding does not influence the machine calculated relevance of the page but hides the boosting techniques from human users. Spam techniques can be categorized by their target ranking signal as well, which can be both based on the content and the link structure of the Web

pages. Correspondingly there are spam techniques that target the content and the link based components of ranking algorithms.

Content spam usually manipulates the textual content of the page, while link spam is the creation of artificial link structure that misleads the search engines and increases the relevance of target pages.

A. Content spam

Most content based ranking methods rely on the assumption that a page is relevant to a query if the query terms appear on the page. The more times they appear, the more relevant the page is. This allows spammers to create pages that are relevant for many queries by adding a large number of terms. On the other hand, spammers can increase the relevance of a page for a few queries by the repetition of query terms. The search engine relevance score also depends on page-independent properties of the query terms (e.g. global frequency in corpus) but these properties are practically inaccessible for spammers except for possibly the very rare words.

According to Gyöngyi et al. the most popular content spamming techniques are:

- Repetition of a few terms. By adding the same word multiple times, the TF score of the given term increases and the document may get ranked higher for queries containing the repeated term.
- Dumping of a large number of distinct terms. Spammers can copy whole dictionaries and they can add misspelled versions of popular query terms. This increases the number of queries matched by the document.
- Weaving of spam terms into copied content. Its effect is similar to dumping: the document may seem to be relevant to more queries. It is a hiding technique as well, because the copied content may be of high quality that makes the document structure look reasonable.

B. Link spam

Link spammers target the link-based ranking algorithms of search engines such as Page Rank and HITS. They create special link structures, the so-called link farms that mislead these algorithms. Assume that a search engine uses the number of different pages pointing to a given page as relevance signal. This relevance score is very easily spammable. The spammer can create a large number of pages linking to the target page at very low cost. The pages of this link farm should be accessible for the search engine crawler but this can be achieved by adding links on a page which is already accessible. The score computed for the target page will be the number of pages created. Fortunately, search engines use more robust link-based algorithms but these algorithms can also be manipulated by large link structures. For constructing complex link structures, spammers need web pages as supporting nodes.

Spammers can use two kinds of pages:

1. *Own* pages are under full control of the spammers. These pages are created by spammers. Their content, URL and link structure is authored and designed by them.
2. Accessible pages are not owned by spammers but anybody is allowed to add content.

These can be unmoderated forums and blogposts. Spammers may add links as comments to propagate the “reputation” of the accessible page towards the target page.

C. Combat Spam

Content hiding may make text or links invisible when the page is rendered. There are several techniques to make different parts of the HTML invisible:

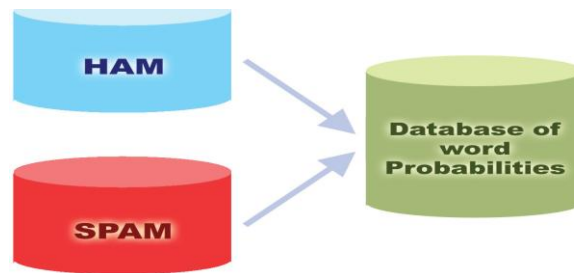
- Color scheme: Text can be made invisible using the same color for the text and background.
- Anchor images: Links can be made invisible by using small 1x1-pixel transparent anchor images.
- Scripts: There are various ways to make HTML elements invisible using scripts, e.g. by setting the “visible” attribute of HTML elements to false.

IV. Current spam detection techniques

Spam is an ever-increasing problem. The number of spam mails is increasing daily – studies show that over 50% of all current email is spam; the Gartner Group predicts this will reach 90% by 2015. Added to this, spammers are becoming more sophisticated and are constantly managing to outsmart ‘static’ methods of fighting spam. The techniques currently used by most anti-spam software are static, meaning that it is fairly easy to evade by tweaking the message a little. To do this, spammers simply examine the latest anti-spam techniques and find ways how to dodge them. To effectively combat spam, an adaptive new technique is needed. This method must be familiar with spammers’ tactics as they change over time. It must also be able to adapt to the particular organization that it is protecting from spam. The answer lies in Bayesian spam filter.

V. How the Bayesian spam filter works

Bayesian filtering is based on the principle that most events are dependent and that the probability of an event occurring in the future can be inferred from the previous occurrences of that event. This same technique can be used to classify spam. If some piece of text occurs often in spam but not in legitimate mail, then it would be reasonable to assume that this email is probably spam. Creating a tailor-made Bayesian word database before mail can be filtered using this method, the user needs to generate a database with words and tokens, collected from a sample of spam mail and valid mail (referred to as ‘ham’).



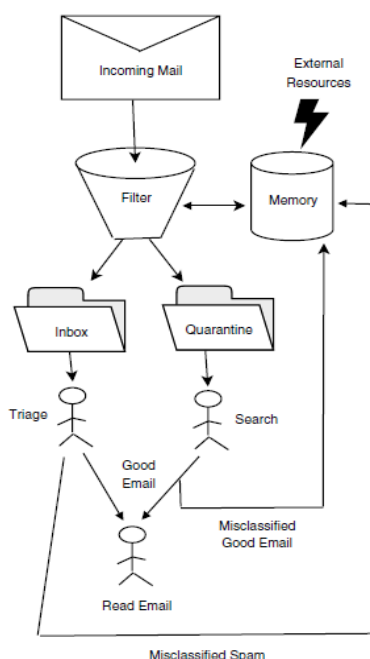
Creating a word database for the filter

VI. Advantages of Bayesian Filter

1. The Bayesian method takes the whole message into account – It recognizes keywords that identify spam, but it also recognizes words that denote valid mail. For example: not every email that contains the word “free” and “cash” is spam. The advantage of the Bayesian method is that it considers the most interesting words (as defined by their deviation from the mean) and comes up with a probability that a message is spam. The Bayesian method would find the words “cash” and “free” interesting but it would also recognize the name of the business contact who sent the message and thus classify the message as legitimate, for instance; it allows words to “balance” each other out. In other words, Bayesian filtering is a much more intelligent approach because it examines all aspects of a message, as opposed to keyword checking that classifies a mail as spam on the basis of a single word.
2. A Bayesian filter is constantly self-adapting - By learning from new spam and new valid outbound mails, the Bayesian filter evolves and adapts to new spam techniques. For example, when spammers started using “f-r-e-e” instead of “free” they succeeded in evading keyword checking until “f-r-e-e” was also included in the keyword database. On the other hand, the Bayesian filter automatically notices such tactics; in fact if the word “f-r-e-e” is found, it is an even better spam indicator, since it’s unlikely to occur in a ham mail. Another example would be using the word “5ex” instead of “Sex”. You would probably not have a word 5ex in a ham mail, and therefore the likelihood that it is spam increases.
3. The Bayesian technique is sensitive to the user. It learns the email habits of the company and understands that, for example, the word ‘mortgage’ might indicate spam if the company running the filter is, say, a car dealership, whereas it would not indicate it as spam if the company is a financial institution dealing with mortgages.
4. The Bayesian method is multi-lingual and international – A Bayesian anti-spam filter, being adaptive, can be used for any language required. Most keyword lists are available in English only and are therefore quite useless in non English-speaking regions. The Bayesian filter also takes into account certain languages deviations or the diverse usage of certain words in different areas, even if the same language is spoken. This intelligence enables such a filter to catch more spam.
5. A Bayesian filter is difficult to fool, as opposed to a keyword filter – An advanced spammer who wants to trick a Bayesian filter can either use fewer words that usually indicate spam (such as free, Viagra, etc), or more words that generally indicate valid mail (such as a valid contact name, etc). Doing the latter is impossible because the spammer would have to know the email profile of each recipient - and a spammer can never hope to gather this kind of information from every intended recipient. Using neutral words, for example the word “public”, would not work since these are disregarded in the final analysis. Breaking up words associated with spam, such as using “m-o-r-t-g-a-g-e” instead of “mortgage”, will only increase the chance of the message being spam, since a legitimate user will rarely write the word “mortgage” as “m-o-r-t-g-a-g-e”.

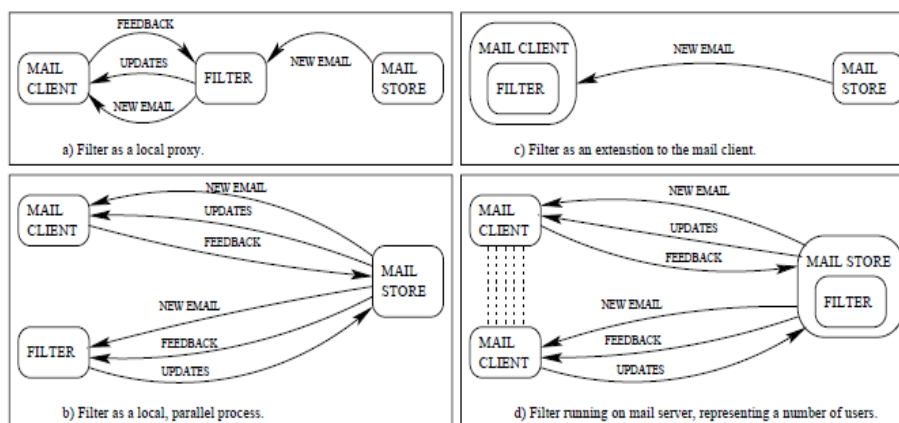
VII. Email Spam Filter

Figure outlines the typical use of an email spam filter from the perspective of a single user. Incoming messages are processed by the filter one at a time and classified as ham (a widely used colloquial term for non-spam) or spam. Ham is directed to the user’s inbox which is read regularly. Spam is directed to a quarantine file which is irregularly (or never) read but may be searched in an attempt to find ham messages which the filter has classified. If the user discovers filter errors — either spam in the inbox or ham in the quarantine — he or she may report these errors to the filter, particularly if doing so is easy and he or she feels that doing so will improve filter performance. In classifying a message, the filter employs the content of the message, its built-in knowledge and algorithms, and also, perhaps, its memory of previous messages, feedback from the user, and external resources such as blacklists or reports from other users, spam filters, or mail servers. The filter may run on the user’s computer, or may run on a server where it performs the same service for many users.



VIII. Evaluation Method

An ideal spam filter would autonomously, immediately, and perfectly identify spam as spam and non-spam as non-spam. To evaluate a spam filter, we must somehow measure how closely it approximates this ideal. Furthermore, whatever measurement we use should reflect the suitability of the filter for its intended purpose. Our ideal suggests four dimensions along which filters should be judged: autonomy, immediacy, spam identification, and non-spam identification. It is not obvious how to measure any of these dimensions separately, nor how to combine these measurements into a single one for the purpose of comparing filters. Nevertheless, reasonable standard measures are useful to facilitate comparison, provided that the goal of optimizing them does not replace that of finding the most suitable filter for the purpose of *spam filtering*.



Possible Placement of Filters

A fully autonomous spam filter would require no configuration, no updates, no training, and no feedback. Is a filter that receives nightly signature files from a central source more or less autonomous than one that requires user feedback on errors? Is the burden collecting a sample of labeled messages for training more or less onerous than delivering updates or user feedback? We cannot imagine a quantitative measure that could capture the differences between filters in this regard. They must be controlled when evaluating the other dimensions, but the relative amounts that filters employing these techniques diverge from the ideal will remain a matter of qualitative, not quantitative, evaluation. An immediate filter would introduce no CPU, disk or network overhead, and would not defer its decision pending the arrival of new information. We may measure or analyze the efficiency of the filter; modeling external delay is more difficult. Reasonable delays may not matter much, but it is difficult to quantify reasonable. A two second delay per message may be reasonable for an end user, if the filter runs continuously. If, however, the filter is launched only when the inbox is opened, a user with 100 new messages may find him or herself waiting for

several minutes. A mail server supporting 100 clients may also find a 2 second delay per message acceptable; a server supporting 100,000 clients may not.

Failures to identify non-spam and spam messages have materially different consequences. Misclassified non-spam messages are likely to be rejected, discarded or placed in quarantine. Any of these actions substantially increases the risk that the information contained in the message will be lost, or at least delayed substantially. Exactly how much risk and delay are incurred is difficult to quantify, as are the consequences, which depend on the nature of the message. Some messages are simply more important than others, while others are more likely to be missed, or delivered by separate channels, if they go astray. For example, advertising from a frequent flier program is less important than an electronic ticket receipt, but the latter is certain to be missed and retrieved, either from quarantine or from a different medium. On the other hand, failure to deliver immediately a message from one's spouse to "come home right away" could have serious consequences. For these reasons, one must be cautious about characterizing failures

to deliver non-spam in terms of a simple proportion, as such failures are rare events with causes and consequences that defeat statistical inference.

IX. Conclusion and Future Enhancements

Bayesian filtering, if implemented the right way and tailored to your application is by far the most effective technology to combat spam. Is there a downside? Well, in a way there is one own side, but this can easily be overcome: Before you can use and judge the Bayesian filter, you have to wait at least two weeks for it to learn – that or create the ham or spam databases yourself. This task can be quite complex, so it is best to wait until the filter has had time to learn. Over time, the Bayesian filter becomes more and more effective as it learns more about your organization's email habits. It is probable that basic anti-spam software might perform better initially, but after a few weeks the Bayesian filter catches up and well outperforms the conventional anti-spam filters once and for all. While other filtering methods may be as effective for now, Bayesian filtering includes an element of changeability that makes it very difficult for spammers to circumvent. Savvy spammers with a lot of resources try to buy every anti-spam filtering product they can get their hands on, and test their spam messages against it so they can be sure the spam will get through to the recipient. Since every site with a Bayesian filter installed has a different set of tokens in their database, it makes it impossible for the spammer to deliberately craft a message that will bypass a site's Bayesian filter. Because the Bayesian filter is constantly being updated, every sneaky change in message wording made by spammers will be quickly identified and rendered ineffective. This makes Bayesian filtering the bane of every spammer, and something every site serious about stopping spam should have. By information retrieval standards, spam filters yield outstanding results. In this paper we discussed that understanding and improvement of the effectiveness of spam filters is best achieved by a collaboration of Bayesian Spam filters and Email Spam filters using common measures and statistical methods. To this end, we have reviewed existing results with respect to common criteria and have outlined designs for future studies.

X. References

- [1] A.J. Alberg, J. W. Park, B. W. Hager, M. V. Brock, and M. Diener-West, "The use of overall accuracy to evaluate the validity of screening or diagnostic tests," *Journal of General Internal Medicine*, vol. 19, no. 1, 2004.
- [2] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of Naive Bayesian anti-spam filtering," *CoRR*, vol. cs.CL/0006013, Informal Publication, 2000.
- [3] I. Androutsopoulos, E. F. Magiourou, and D. K. Vassilakis, "A game theoretic model of spam e-mailing," in *CEAS 2005 — The Second Conference on Email and Anti-Spam*, 2005.
- [4] R. Baeza-Yates and E. Davis. Web page ranking using link attributes. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, pages 328–329, 2004.
- [5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, Boston, 1999.
- [6] Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. PageRank increase under different collusion topologies. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.
- [7] Ziv Bar-Yossef, Andrei Z. Broder, Ravi Kumar, and Andrew Tomkins. Sic transit Gloria telae: Towards an understanding of the web's decay. In *Proceedings of the 13th World Wide Web Conference (WWW)*, pages 328–337. ACM Press, 2004.
- [8] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: the topology of the word-wide web. *Physica A*, 281:69–77, 2000.
- [9] S. Sarafijanovic and J.-Y. Le Boudec. Artificial immune system for collaborative spam filtering. In *Proceedings of NISCO 2007, The Second Workshop on Nature Inspired Cooperative Strategies for Optimization*, Acireale, Italy, November 8-10, 2007.
- [10] SpamAssassin. <http://spamassassin.org/>, Feb 2008.
- [11] SpamAssassin-Public-Corpus. <http://spamassassin.org/publiccorpus/>, Feb 2008.
- [12] F. Zhou, L. Zhuang, B. Zhao, L. Huang, A. Joseph, and J. Kubiawicz. Approximate object location and spam filtering on peer-to-peer systems. In *Proceedings of ACM/IFIP/USENIX Int'l Middleware Conf.*, LNCS 2672, pp. 1-20.
- [13] Rhyolite Software. Distributed Checksum Clearinghouse. <http://www.rhyolite.com/anti-spam/dcc/>, 2004.
- [14] The Spamhaus Project. The Exploits Block List. <http://www.spamhaus.org/xbl/index.lasso>, 2004.
- [15] The Spamhaus Project. The Spamhaus Block List. <http://www.spamhaus.org/sbl/>, 2004.
- [16] Feng Zhou, Li Zhuang, Ben Y. Zhao, Ling Huang, Anthony D. Joseph, and John D. Kubiawicz. Approximate object location and spam filtering on peer-to-peer systems. In *ACM/IFIP/USENIX International Middleware Conference*, June 2003.