Email Spam Filtering: A Systematic Review

Email Spam Filtering: A Systematic Review

Gordon V. Cormack

David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario
N2L 3G1
Canada

gvcormac@uwaterloo.ca

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is G. V. Cormack, Email Spam Filtering: A Systematic Review, Foundation and Trends® in Information Retrieval, vol 1, no 4, pp 335–455, 2006

ISBN: 978-1-60198-146-2

Foundations and Trends® in Information Retrieval

Volume 1 Issue 4, 2006

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2006, Volume 1, 4 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 1, No. 4 (2006) 335–455
© 2008 G. V. Cormack
DOI: 10.1561/1500000006

the essence of knowledge

Email Spam Filtering: A Systematic Review

Gordon V. Cormack

David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada, gvcormac@uwaterloo.ca

Abstract

Spam is information crafted to be delivered to a large number of recipients, in spite of their wishes. A spam filter is an automated tool to recognize spam so as to prevent its delivery. The purposes of spam and spam filters are diametrically opposed: spam is effective if it evades filters, while a filter is effective if it recognizes spam. The circular nature of these definitions, along with their appeal to the intent of sender and recipient make them difficult to formalize. A typical email user has a working definition no more formal than "I know it when I see it." Yet, current spam filters are remarkably effective, more effective than might be expected given the level of uncertainty and debate over a formal definition of spam, more effective than might be expected given the state-of-the-art information retrieval and machine learning methods for seemingly similar problems. But are they effective enough? Which are better? How might they be improved? Will their effectiveness be compromised by more cleverly crafted spam?

We survey current and proposed spam filtering techniques with particular emphasis on how well they work. Our primary focus is spam filtering in email; Similarities and differences with spam filtering in other communication and storage media — such as instant messaging and the Web — are addressed peripherally. In doing so we examine the definition of spam, the user's information requirements and the role of the spam filter as one component of a large and complex information universe. Well-known methods are detailed sufficiently to make the exposition self-contained, however, the focus is on considerations unique to spam. Comparisons, wherever possible, use common evaluation measures, and control for differences in experimental setup. Such comparisons are not easy, as benchmarks, measures, and methods for evaluating spam filters are still evolving. We survey these efforts, their results and their limitations. In spite of recent advances in evaluation methodology, many uncertainties (including widely held but unsubstantiated beliefs) remain as to the effectiveness of spam filtering techniques and as to the validity of spam filter evaluation methods. We outline several uncertainties and propose experimental methods to address them.

Contents

1

Introduction

The Spam Track at the Text Retrieval Conference (TREC) defines email spam as

> "Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient." [40]

Although much of the history of spam is folklore, it is apparent that spam was prevalent in instant messaging (Internet Relay Chat, or IRC) and bulletin boards (Usenet, commonly dubbed *newsgroups*) prior to the widespread use of email. Spam countermeasures are as old as spam, having progressed from *ad hoc* intervention by administrators through simple hand-crafted rules through automatic methods based on techniques from information retrieval and machine learning, as well as new methods specific to spam. Spam has evolved so as to defeat countermeasures; countermeasures have evolved so as to thwart evasion.

We generalize the TREC definition of spam to capture the essential adversarial nature of spam and spam abatement.

> **Spam**: *unwanted communication intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery.*
> **Spam filter**: *an automated technique to identify spam for the purpose of preventing its delivery.*

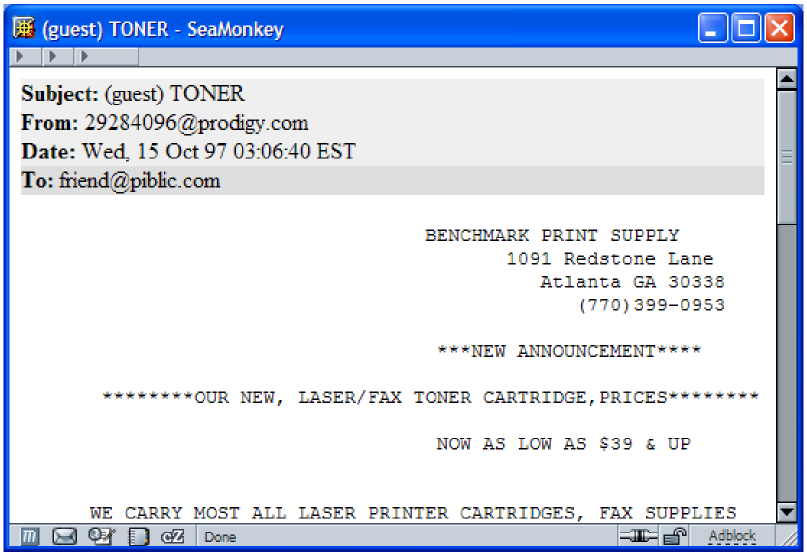
Applying these definitions requires the adjudication of subjective terms like *intent* and *purpose.* Furthermore, any evaluation of spam filtering techniques must consider their performance within the context of how well they fulfill their intended purpose while avoiding undesirable consequences. It is tempting to conclude that scientific spam filter evaluation is therefore impossible, and that the definition of spam or the choice of one filter over another is merely a matter of taste. Or to conclude that the subjective aspects can be "defined away" thus reducing spam filter evaluation to a simple mechanical process. We believe that both conclusions are specious, and that sound quantitative evaluation can and must be applied to the problem of spam filtering.

While this survey confines itself to email spam, we note that the definitions above apply to any number of communication media, including text and voice messages [31, 45, 84], social networks [206], and blog comments [37, 123]. It applies also to web spam, which uses a search engine as its delivery mechanism [187, 188].

1.1 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 (likely worthless, illegal, or non-existent) 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 [117].

A decade ago (circa 1997), the mechanism, payload, and purpose of spam were quite transparent. The majority of spam was sent by "cottage industry" spammers who merely abused social norms to promote



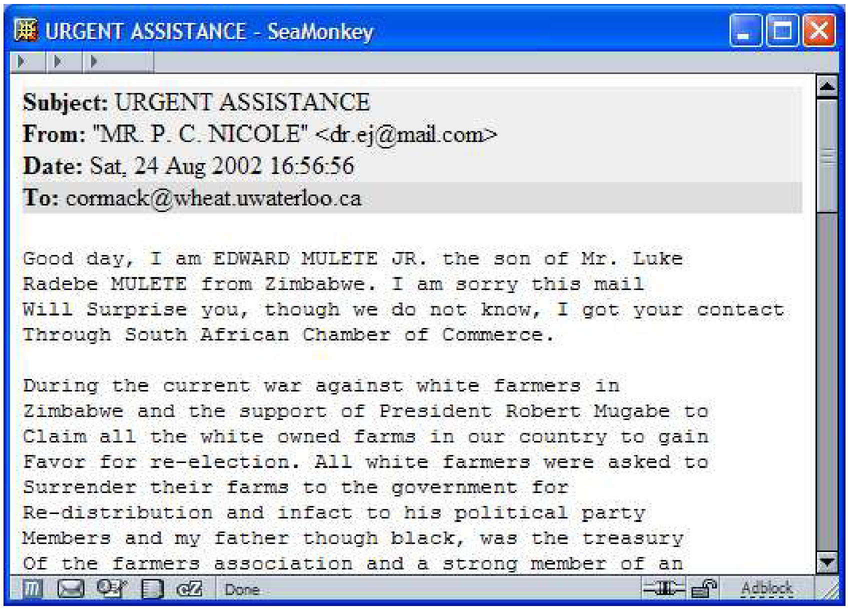
Subject: (guest) TONER
From: 29284096@prodigy.com
Date: Wed, 15 Oct 97 03:06:40 EST
To: friend@piblic.com

```
                                   BENCHMARK PRINT SUPPLY
                                          1091 Redstone Lane
                                             Atlanta GA 30338
                                                (770)399-0953

                                    ***NEW ANNOUNCEMENT****

            ********OUR NEW, LASER/FAX TONER CARTRIDGE,PRICES********

                                    NOW AS LOW AS $39 & UP


           WE CARRY MOST ALL LASER PRINTER CARTRIDGES, FAX SUPPLIES
```

Fig. 1.1 Marketing spam.

their wares (Figure 1.1). Fraud bait consisted of clumsily written "Nigerian scams" (Figure 1.2) imploring one to send bank transit information so as to receive several MILLION DOLLARS from an aide to some recently deposed leader. Cause promotion took the form of obvious chain letters (Figure 1.3), while computer viruses were transmitted as attached executable files (Figure 1.4). Yet enough people received and responded to these messages to make them lucrative, while their volume expanded to become a substantial inconvenience even to those not gullible enough to respond.

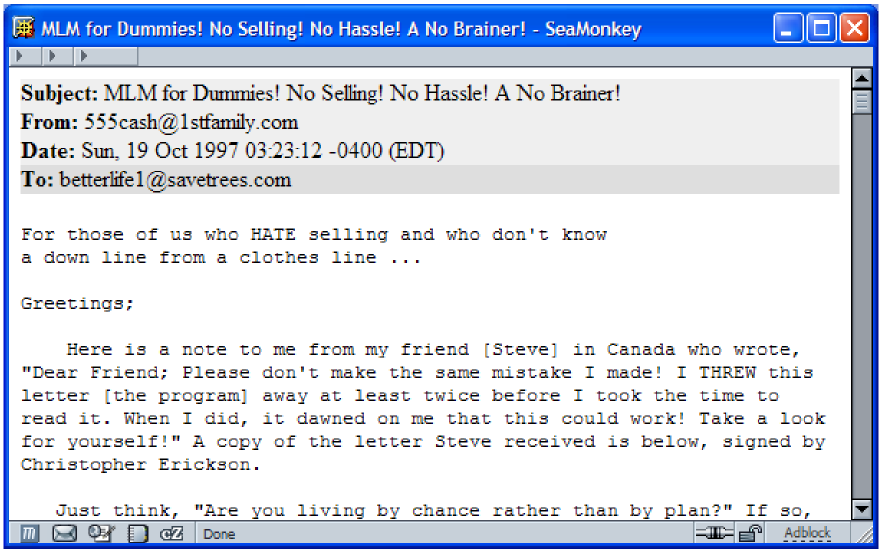
At the same time, spamming has become more specialized and sophisticated, with better hidden payloads and more nefarious purposes. Today, cottage industry spam has been overwhelmed by spam sent in support of organized criminal activity, ranging from traffic in illegal goods and services through stock market fraud, wire fraud, identity theft, and computer hijacking [140, 178]. 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!



Subject: URGENT ASSISTANCE
From: "MR. P. C. NICOLE" <dr.ej@mail.com>
Date: Sat, 24 Aug 2002 16:56:56
To: cormack@wheat.uwaterloo.ca

```
Good day, I am EDWARD MULETE JR. the son of Mr. Luke
Radebe MULETE from Zimbabwe. I am sorry this mail
Will Surprise you, though we do not know, I got your contact
Through South African Chamber of Commerce.

During the current war against white farmers in
Zimbabwe and the support of President Robert Mugabe to
Claim all the white owned farms in our country to gain
Favor for re-election. All white farmers were asked to
Surrender their farms to the government for
Re-distribution and infact to his political party
Members and my father though black, was the treasury
Of the farmers association and a strong member of an
```

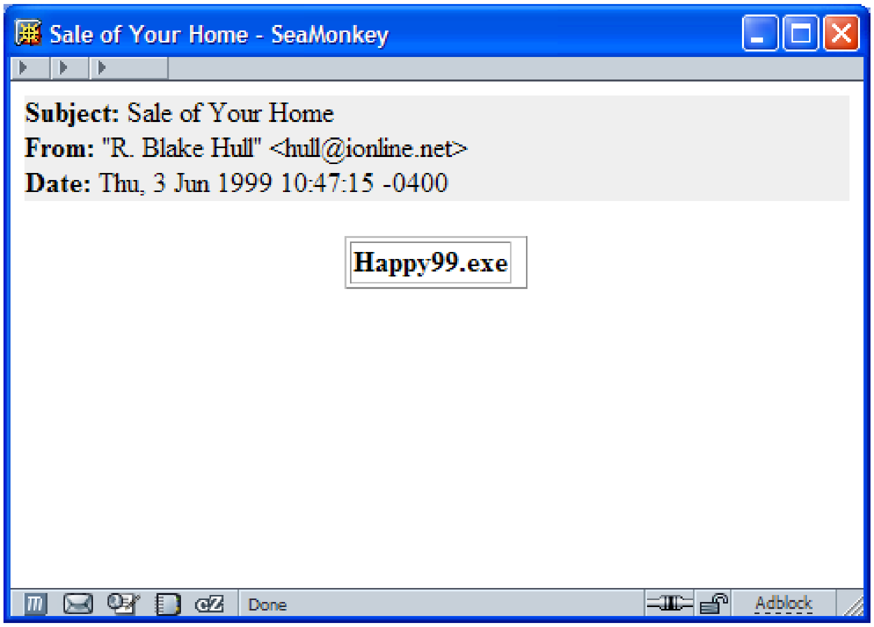
Fig. 1.2 Nigerian spam.


MLM for Dummies! No Selling! No Hassle! A No Brainer! - SeaMonkey
Done
Adblock

Subject: MLM for Dummies! No Selling! No Hassle! A No Brainer!
From: 555cash@1stfamily.com
Date: Sun, 19 Oct 1997 03:23:12 -0400 (EDT)
To: betterlife1@savetrees.com

```
For those of us who HATE selling and who don't know
a down line from a clothes line ...

Greetings;

    Here is a note to me from my friend [Steve] in Canada who wrote,
"Dear Friend; Please don't make the same mistake I made! I THREW this
letter [the program] away at least twice before I took the time to
read it. When I did, it dawned on me that this could work! Take a look
for yourself!" A copy of the letter Steve received is below, signed by
Christopher Erickson.

   Just think, "Are you living by chance rather than by plan?" If so,
```

Fig. 1.3 Chain letter spam.



Subject: Sale of Your Home
From: "R. Blake Hull" <hull@ionline.net>
Date: Thu, 3 Jun 1999 10:47:15 -0400

Happy99.exe

Fig. 1.4 Virus spam.

Spam, to meet its purpose, must necessarily have a payload which is delivered and acted upon[1] in the intended manner. Spam abatement techniques are effective to the extent that they prevent delivery, prevent action, or substitute some other action that acts as a disincentive.[2] Spam filters, by identifying spam, may be used in support of any of these techniques. At the same time, the necessary existence of a payload may aid the filter in its purpose of identifying spam.

1.2 Spam Characteristics

Spam in all media commonly share a number of characteristics that derive from our definition and discussion of the purpose of spam.

[1] The target need not be a person; a computer may receive and act upon the spam, serving its purpose just as well.

[2] Such as arresting the spammer.

1.2.1 Unwanted

It seems obvious that spam messages are unwanted, at least by the vast majority of recipients. Yet some people respond positively to spam, as evidenced by the fact that spam campaigns work [71]. Some of these individuals no doubt come to regret having responded, thus calling into question whether they indeed wanted to receive the spam in the first place. Some messages — such as those trafficking in illegal goods and services — may be wanted by specific individuals, but classed as unwanted by society at large. For most messages there is broad consensus as to whether or not the message is wanted, for a substantial minority (perhaps as high as 3% [168, 199]) there is significant disagreement and therefore some doubt as to whether the message is spam or not.

1.2.2 Indiscriminate

Spam is transmitted outside of any reasonable relationship[3] or prospective relationship between the sender and the receiver. In general, it is more cost effective for the spammer to send more spam than to be selective as to its target. An unwanted message targeting a specific individual, even if it promotes dubious products or causes or contains fraud bait or a virus, does not meet our definition of spam.

A message that is automatically or semi-automatically tailored to its target is nonetheless indiscriminate. For example, a spammer may harvest the name of the person owning a particular email address and include that name in the salutation of the message. Or a spammer may do more sophisticated data mining and sign the message with the name and email address of a colleague or collaborator, and may include in the text subjects of interest to the target. The purpose of such tailoring is, of course, to disguise the indiscriminate targeting of the message.

1.2.3 Disingenuous

Because spam is unwanted and indiscriminate, it must disguise itself to optimize the chance that its payload will be delivered and acted

[3] We have dropped the term *unsolicited* used in TREC and earlier definitions of spam, because not all unsolicited email is spam, and that which is captured by our notion of *indiscriminate*. Solicited email, on the other hand, is clearly not indiscriminate.

upon. The possible methods of disguise are practically unlimited and cannot be enumerated in this introduction (cf. [27, 67, 75]). Some of the most straightforward approaches are to use plausible subject and sender data, as well as subject material that appears to be legitimate. It is common, for example, to receive a message that appears to be a comment from a colleague pertaining to a recent news headline. Even messages with random names; for example a wire transfer from John to Judy, will appear legitimate to some fraction of its recipients. Messages purporting to contain the latest security patch from Microsoft will similarly be mistaken for legitimate by some fraction of recipients.

Spam must also disguise itself to appear legitimate to spam filters. Word misspelling or obfuscation, embedding messages in noisy images, and sending messages from newly hijacked computers, are spam characteristics designed to fool spam filters. Yet humans — or filters employing different techniques — can often spot these characteristics as unique to spam.

1.2.4 Payload Bearing

The payload of a spam message may be obvious or hidden; in either case spam abatement may be enhanced by identifying the payload and the mechanism by which actions triggered by it profit the spammer. Obvious payloads include product names, political mantras, web links, telephone numbers, and the like. These may be in plain text, or they may be obfuscated so as to be readable by the human but appear benign to the computer. Or they may be obfuscated so as to appear benign to the human but trigger some malicious computer action.

The payload might consist of an obscure word or phrase like "gouranga" or "platypus race" in the hope that the recipient will be curious and perform a web search and be delivered to the spammer's web page or, more likely, a paid advertisement for the spammer's web page. Another form of indirect payload delivery is *backscatter*: The spam message is sent to a non-existent user on a real mail server, with the (forged) return address of a real user. The mail server sends an "unable to deliver" message to the (forged) return address, attaching and thus delivering the spam payload. In this scenario we consider both the original message (to the non-existent user) and the "unable to deliver" message to be spam, even though the latter is transmitted by a legitimate sender.

The payload might be the message itself. The mere fact that the message is not rejected by the mail server may provide information to the spammer as to the validity of the recipient's address and the nature of any deployed spam filter. Or if the filter employs a machine learning technique, the message may be designed to *poison* the filter [70, 72, 191], compromising its ability to detect future spam messages.

1.3 Spam Consequences

The transmission of spam — whether or not its payload is delivered and acted upon — has several negative consequences.

1.3.1 Direct Consequences

Spam provides an unregulated communication channel which can be used to defraud targets outright, to sell shoddy goods, to install viruses, and so on. These consequences are largely, but not exclusively, borne by the victims. For example, the victim's computer may be used in further spamming or to launch a cyber attack. Similarly, the victim's identity may be stolen and used in criminal activity against other targets.

1.3.2 Network Resource Consumption

The vast majority of email traffic today is spam. This traffic consumes bandwidth and storage, increasing the risk of untimely delivery or outright loss of messages. For example, during the Sobig virus outbreak of 2003, the author's spam filter correctly identified the infected messages as spam and placed them in a quarantine folder. However, the total volume of such messages exceeded 5 GB per day, quickly exhausting all available disk space resulting in non-delivery of legitimate messages.

1.3.3 Human Resource Consumption

It is an unpleasant experience and a waste of time to sort through an inbox full of spam. This process necessarily interferes with the timeliness of email because the recipient is otherwise occupied sorting through spam. Furthermore, the frequent arrival of spam may preclude the use of email arrival alerts, imposing a regimen of batch rather than on-arrival email reading, further compromising timeliness.

Over and above the wasted time of routinely sifting through spam, some spam messages may consume extraordinary time and resources if they appear legitimate and cannot be dismissed based on the summary information presented by the mail reader's user interface. More importantly, legitimate email messages may be overlooked or dismissed as spam, with the consequence that the message is missed.

A spam filter may mitigate any or all of the problems associated with human resource consumption, potentially reducing effort while also enhancing timeliness and diminishing the chance of failing to read a legitimate message.

1.3.4 Lost Email

Sections 1.3.2 and 1.3.3 illustrate situations in which spam may cause legitimate email to be lost or overlooked. Spam abatement techniques may, of course, also cause legitimate email to be lost. More generally, spam brings the use of email into disrepute and therefore discourages its use. Users may refuse to divulge their email addresses or may obfuscate them in ways that inhibit the use of email as a medium to contact them.

In evaluating the consequences of email loss (or potential loss), one must consider the probability of loss, the importance and time criticality of the information, and the likelihood of receiving the information, or noticing its absence, via another medium. These consequences vary from message to message, and must be considered carefully in evaluating the effectiveness of any approach to spam abatement, including human sorting.

1.4 The Spam Ecosystem

Spam and spam filters are components of a complex interdependent system of social and technical structures. Many spam abatement proposals seek to alter the balance within the system so as to render spam impractical or unprofitable. Two anonymous whimsical articles [61, 1] illustrate the difficulties that arise with naive efforts to find the Final Ultimate Solution to the Spam Problem (FUSSP). Crocker [43] details the social issues and challenges in effecting infrastructure-based solutions such as protocol changes and sender authentication. Legislation, prosecution and civil suits have been directed at spammers [101, 124], however, the international and underground nature of many spam operations makes them difficult to target. Spammers and legitimate advertisers have taken action against spam abatement outfits [119]. Vigilante actions have been initiated against spammers, and spammers have reacted in kind with sabotage and extortion [103]. Economic and technical measures have been proposed to undermine the profitability of spam [89, 138].

A detailed critique of system-wide approaches to spam abatement is beyond the scope of this survey, however, it is apparent that no FUSSP has yet been found nor, we daresay, is likely to be found in the near future. And even if the email spam problem were to be solved, it is not obvious that the solution would apply to spam in other media. The general problem of adversarial information filtering [44] — of which spam filtering is the prime example — is likely to be of interest for some time to come.

We confine our attention to this particular problem — identifying spam — while taking note of the fact that the deployment of spam filters will affect the spam ecosystem, depending on the nature of their deployment. The most obvious impact of spam filtering is the emergence of technical countermeasures in spam; it is commonly held that filtering methods become obsolete as quickly as they are invented. Legal retaliation is also a possibility: Spammers or advertisers or recipients may sue for damages due to the non-delivery of messages. Spam filtering is itself a big business, a tremendous amount of money rests on our perception of which spam methods work best, so the self-interest of vendors may be at odds with objective evaluation. And filter market share will itself influence the design of spam.

In general, we shall consider the marginal or incremental effects of spam filter deployment, and mention in passing its potential role in revolutionary change.

1.5 Spam Filter Inputs and Outputs

We have defined a spam filter to be an automated technique to identify spam. A spam filter with perfect knowledge might base its decision on the content of the message, characteristics of the sender and the target, knowledge as to whether the target or others consider similar messages to be spam, or the sender to be a spammer, and so on. But perfect knowledge does not exist and it is therefore necessary to constrain the filter to use well defined information sources such as the content of the message itself, hand-crafted rules either embedded in the filter or acquired from an external source, or statistical information derived from feedback to the filter or from external repositories compiled by third parties.

The desired result from a spam filter is some indication of whether or not a message is spam. The simplest result is a binary categorization — spam or non-spam — which may be acted upon in various ways by the user or by the system. We call a filter that returns such a binary categorization a *hard classifier.* More commonly, the filter is required to give some indication of how likely it considers the message to be spam, either on a continuous scale (e.g., 1 = *sure spam*; 0 = *sure non-spam*) or on an ordinal categorical scale (e.g., *sure spam*, *likely spam*, *unsure*, *likely non-spam*, *sure non-spam*). We call such a filter a *soft classifier.* Many filters are internally soft classifiers, but compare the soft classification result to a sensitivity threshold t yielding a hard classifier. Users may be able to adjust this sensitivity threshold according to the relative importance they ascribe to correctly classifying spam vs. correctly classifying non-spam (see Section 1.7).

A filter may also be called upon to justify its decision; for example, by highlighting the features upon which it bases is classification. The filter may also classify messages into different *genres* of spam and good mail (cf. [42]). For example, spam might be advertising, phishing or a Nigerian scam, while good email might be a personal correspondence, a news digest or advertising. These genres may be important in justifying the spam/non-spam classification of a message, as well in assessing its impact (e.g., does the user really care much about the distinction between spam and non-spam advertising?).

1.5.1 Typical Email Spam Filter Deployment

Figure 1.5 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

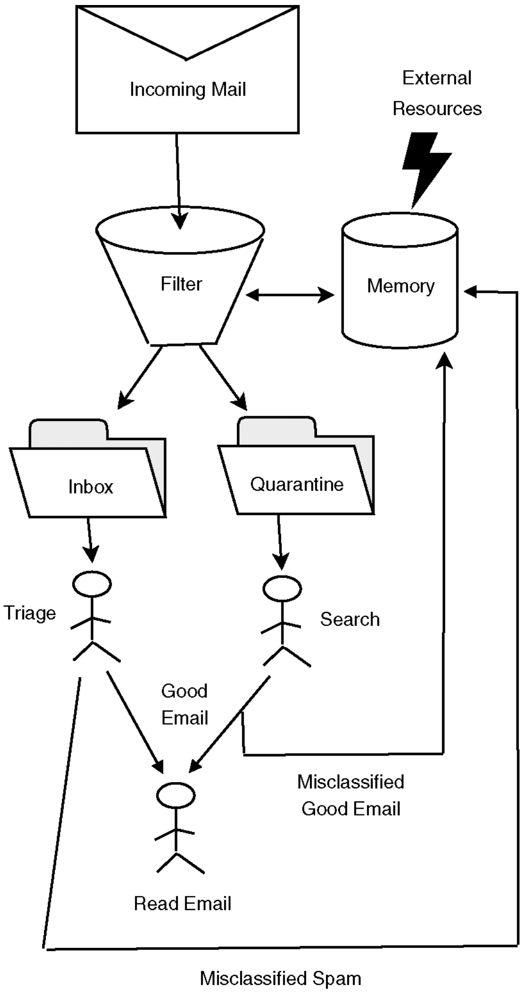


Fig. 1.5 Spam filter usage.

never) read but may be searched in an attempt to find ham messages which the filter has misclassified. 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 [133] 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.

1.5.2 Alternative Deployment Scenarios

The filter diagrammed in Figure 1.5 is on-line in that it processes one message at a time, classifying each in turn before examining the next. Furthermore, it is passive in that it makes use only of information at hand when the message is examined. Variants of this deployment are possible, only some of which have been systematically investigated:

- Batch filtering, in which several messages are presented to the filter at once for classification. This method of deployment is atypical in that delivery of messages must necessarily be delayed to form a batch. Nevertheless, it is conceivable that filters could make use of information contained in the batch to classify its members more accurately than on-line.
- Batch training, in which messages may be classified on-line, but the classifier's memory is updated only periodically. Batch training is common for classifiers that involve much computation, or human intervention, in harnessing new information about spam.
- Just-in-time filtering, in which the classification of messages is driven by client demand. In this deployment a filter would defer classification until the client opened his or her mail client, sorting the messages in real-time into inbox and quarantine.

- Deferred or tentative classification, in which the classification of messages by the filter is uncertain, and either delivery of the message is withheld or the message is tentatively classified as ham or spam. As new information is gleaned, the classification of the message may be revised and, if so, it is delivered or moved to the appropriate file.
- Receiver engagement, in which the filter probes the recipient (or an administrator representing the recipient) to glean more information as a basis for classification. Active learning may occur in real-time (i.e., the information is gathered during classification) or in conjunction with deferred or tentative classification. An example of real-time active learning might be a user interface that solicits human adjudication from the user as part of the mail reading process. A more passive example is the use of an "unsure" folder into which messages are placed with the expectation that the user will adjudicate the messages and communicate the result to the filter.
- Sender engagement, in which the filter probes the sender or the sender's machine for more information. Examples are challenge–response systems and greylisting. These filters may have a profound effect on the ecosystem as they, through their probes, transmit information back to the sender. Furthermore, they introduce delays and risks of non-delivery that are difficult to assess [106]. It may be argued that these techniques which engage the sender do not fit our notion of "filter." Nevertheless, they are commonly deployed in place of, or in conjunction with, filters and so their effects must be considered.
- Collaborative filtering, in which the filter's result is used not only to classify messages on behalf of the user, but to provide information to other filters operating on behalf of other users. The motivation for collaborative filtering is that spam is sent in bulk, as is much hard-to-classify good email, so many other users are likely to receive the same or similar messages. Shared knowledge among the filters promises to

make such spam easier to detect. Potential pitfalls include risks to privacy and susceptibility to manipulation by malicious participants.

- Social network filtering, in which the sender and recipient's communication behavior are examined for evidence that particular messages might be spam.

1.6 Spam Filter Evaluation

Scientific evaluation, critical to any investigation of spam filters, addresses fundamental questions:

- Is spam filtering a viable tool for spam abatement?
- What are the risks, costs, and benefits of filter use?
- Which filtering techniques work best?
- How well do they work?
- Why do they work?
- How may they be improved?

The vast breadth of the spam ecosystem and possible abatement techniques render impossible the direct measurement of these quantities; there are simply too many parameters for any single evaluation or experiment to measure all their effects at once. Instead, we make various simplifying assumptions which hold many of the parameters constant, and conduct an experiment to measure a quantity of interest subject to those assumptions. Such experiments yield valuable insight, particularly if the assumptions are reasonable and the quantities measured truly illuminate the question under investigation. The validity of an experiment may be considered to have two aspects: *internal validity* and *external validity* or *generalizability*. Internal validity concerns the veracity of the experimental results under the test conditions and stated assumptions; external validity concerns the generalizability of these results to other situations where the stated assumptions, or hidden assumptions, may or may not hold. Establishing internal validity is largely a matter of good experimental design; establishing external validity involves analysis and repeated experiments using different assumptions and designs.

It is all too easy to fix on one experimental design and set of test conditions and to lose sight of the overall question being posed. It is similarly all too easy to dismiss the results a particular experiment due to the limitations inherent in its assumptions. For example, filters are commonly evaluated using tenfold cross validation [95], which assumes that the characteristics of spam are invariant over time. It would be wrong to conclude, without evidence, that the results of tenfold cross validation would be the same under a more realistic assumption. It would be equally wrong to dismiss out of hand the results of experiments using this method, to do so would entail dismissal of all scientific evidence, as there is no experiment without limiting assumptions. We would be left with only testimonials, or our own uncontrolled and unrepeatable observations, to judge the efficacy of various techniques. Instead, it is appropriate to identify assumptions that limit the generalizability of current results, and to conduct experiments to measure their effect.

The key to evaluation is to conduct experiments that glean the most informative results possible with reasonable effort, at reasonable cost, in a reasonable time frame. Simple assumptions — such as the assumption that the characteristics of spam are time-invariant — yield simple experiments whose internal validity is easy to establish. Many such experiments may reasonably be conducted to explore the breadth of solutions and deployment scenarios. Further experiments, with different simple assumptions, help to establish the external validity of the results. These experiments serve to identify the parameters and solutions of interest, but are inappropriate for evaluating fine differences. Experimental designs that more aptly model real filter deployment tend to be more complex and costly due to challenges in logistics, controlling confounding factors, and precisely measuring results. Such experiments are best reserved for methods and parameters established to be of interest by simpler ones.

Among the common assumptions in spam filter evaluation are:

- Batch or on-line filtering.
- Existence of training examples.
- Accurate “true” classification for training messages.

- Accurate “true” classification for test messages.
- Recipient behavior, e.g., reporting errors.
- Sender behavior, e.g., resending dropped messages.
- Availability of information, e.g., whitelists, blacklists, rule bases, community adjudication, etc.
- Language of messages to be filtered, e.g., English only.
- Format of messages to be filtered, e.g., text, html, ASCII, Unicode, etc.
- Quantifiable consequences for misclassification or delay [96].
- Time invariance of message characteristics [57].
- Effect (or non-effect) of spam filter on sender.
- Effect (or non-effect) of spam filter on recipient.

Laboratory and field experiments play complementary roles in scientific investigation. Laboratory experiments investigate the fundamental properties of filters under controlled conditions that facilitate reproducibility, precise measurement, and ongoing evaluation. Such conditions necessitate the adoption of simplifying assumptions such as those listed above. Field experiments, on the other hand, rely on different assumptions, are very difficult to control and their results very difficult to compare. Methods from scientific fields such as epidemiology [139] may be used to measure the effects of spam filters, however, such methods are considerably more expensive and less precise than laboratory experiments.

1.7 Evaluation Measures

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*.

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. With this caveat, *false positive rate* (*fpr*) — the proportion of non-spam messages identified as spam (cf. Table 4.1) — is a reasonable first-order measure of failures to identify non-spam.

Failures to identify spam also vary in importance, but are generally less important than failures to identify non-spam. Viruses, worms, and phishing messages may be an exception, as they pose significant risks to the user. Other spam messages have impact in proportion to their volume; so *false negative rate* (*fnr*) — the proportion of spam identified as non-spam — is an apt measure.

The overall efficacy of a hard classifier may be characterized by the pair (*fpr*, *fnr*). A classifier with lower *fpr* and *fnr* than another is superior.[4] Whether a classifier with a lower *fpr* and higher *fnr* is superior or inferior depends on the user's sensitivity to each kind of error.

The efficacy of a soft classifier with an adjustable threshold t may be characterized by the set of all distinguishable (*fpr*, *fnr*) pairs for different values of t. This set of points defines a receiver operating characteristic (ROC) curve (cf. [58, 82, 166]). A filter whose ROC curve is strictly above that of another is superior in all deployment situations, while a filter whose ROC curve crosses that of another is superior for some threshold settings and inferior for others.

The area under the ROC curve (AUC) provides an estimate of the effectiveness of a soft classifier over all threshold settings. AUC also has a probabilistic interpretation: it is the probability that the classifier will award a random spam message a higher score than a random ham message. In the spam filtering domain, typical AUC values are of the order of 0.999 or greater, for clarity, we often report $(1 - AUC)\%$, the

[4] Under the assumption that all messages have equal misclassification cost. See Kolcz et al. [96]

area above the ROC curve, as a percentage. So $AUC = 0.999$ would be reported instead as $(1 - AUC)\% = 0.1$.

False positive rate, false negative rate, and receiver operating characteristic curves are the standard measures of (e.g., medical) diagnostic test effectiveness [66]. This review uses primarily these measures; a spam filter is a diagnostic test applied to email for which a positive result indicates spam, and a negative result indicates non-spam. In Section 4.6, we review the diverse set of measures that have been applied to spam filters, and argue that diagnostic test methods are most suitable for comparative analysis.

1.8 Systematic Review

Spam filters have evolved quickly — and somewhat separately — in several milieux with different histories, objectives, evaluation methods, and measures of success. Practitioners have been concerned primarily with keeping their heads above water, delivering spam filters as quickly as possible to combat an ever-increasing tide of spam. Academics have, in large part, studied the problem as an application of the techniques and methods of information retrieval, machine learning and computer systems. Commercial product development and product testing involve yet another set of interests, methods, and measures of success. These groups have had limited interaction; as a consequence, it is exceedingly difficult to deduce from the literature or other sources the relative performance and promise of current and proposed spam filter methods.

The literature, including the so-called gray literature (dissertations, technical reports, popular press articles, commercial reports, web publications, software documentation and cited unpublished works) was searched for articles describing a spam filter or spam filtering method and an evaluation of its effectiveness. Articles were characterized by their methods and assumptions according to the taxonomy presented here. Where sufficient information was given in the article, quantitative results were recast as $(\mathit{fpr}, \mathit{fnr})$ or summarized using $1 - AUC$ expressed as a percentage, otherwise the results were omitted from this review. Results derived using incorrect methodology, or results that are insufficiently noteworthy because their results are represented better elsewhere, were similarly omitted. Several hundred articles were considered for this review; perhaps one-third of them met our selection criteria.

Certain aspects of spam filtering are well represented in the literature, while others are hardly represented or not represented at all. This review reflects this uneven coverage, reporting some aspects in detail while leaving others as largely uncharted territory.

References

[1] "You might be an anti-spam kook if...," http://www.rhyolite.com/anti-spam/you-might-be.html.

[2] 2004 National Technology Readiness Survey: Summary report, http://www.smith.umd.edu/ntrs/NTRS_2004.pdf, 2005.

[3] 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.

[4] 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.

[5] I. Androutsopoulos, E. F. Magirou, and D. K. Vassilakis, "A game theoretic model of spam e-mailing," in *CEAS 2005 — The Second Conference on Email and Anti-Spam*, 2005.

[6] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam E-mail: A comparison of a naive bayesian and a memory-based approach," in *Proceedings of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 200)*, pp. 1–13, 2000.

[7] I. Androutsopoulos, G. Paliouras, and E. Michelakis, "Learning to filter unsolicited commercial E-Mail," Tech. Rep. 2004/2, *NCSR "Demokritos"*, October 2004.

[8] H. B. Aradhye, G. K. Myers, and J. A. Herson, "Image analysis for efficient categorization of image-based spam e-mail," in *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 2005.

[9] F. Assis, "OSBF-Lua," http://osbf-lua.luaforge.net/.

[10] F. Assis, "OSBF-Lua — A text classification module for Lua the importance of the training method," in *Fifteenth Text REtrieval Conference (TREC-2006)*, Gaithersburg, MD: NIST, 2006.

[11] A. Berg, "Creating an antispam cocktail: Best spam detection and filtering techniques," http://searchsecurity.techtarget.com/tip/1,289483,sid14_gci1116643,00.html, 2005.

[12] S. Bickel, M. Bruckner, and T. Scheffer, "Discriminative learning for differing training and test distributions," *International Conference on Machine Learning (ICML)*, 2007.

[13] S. Bickel and T. Scheffer, "Dirichlet-Enhanced spam filtering based on biased samples," *Neural Information Processing Systems (NIPS)*, 2007.

[14] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "Image spam filtering by content obscuring detection," in *CEAS 2007 — The Third Conference on Email and Anti-Spam*, 2007.

[15] Blacklists compared, http://www.sdsc.edu/ jeff/spam/Blacklists_Compared.html.

[16] A. Bratko, G. V. Cormack, B. Filipič, T. R. Lynam, and B. Zupan, "Spam filtering using statistical data compression models," *Journal of Machine Learning Research*, vol. 7, pp. 2673–2698, December 2006.

[17] A. Bratko and B. Filipič, "Spam filtering using character-level markov models: Experiments for the TREC 2005 Spam Track," in *Proceedings of 14th Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005.

[18] A. Bratko and B. Filipič, "Exploiting structural information for semi-structured document categorization," *Information Processing and Management*, vol. 42, no. 3, pp. 679–694, 2006.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[20] M. Bruckner, P. Haider, and T. Scheffer, "Highly scalable discriminative spam filtering," in *Proceedings of 15th Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, November 2006.

[21] B. Burton, *SpamProbe — A Fast Bayesian Spam Filter*. 2002. http:// spamprobe.sourceforge.net.

[22] B. Byun, C.-H. Lee, S. Webb, and C. Pu, "A discriminative classifier learning approach to image modeling and spam image identification," in *CEAS 2007 — The Third Conference on Email and Anti-Spam*, 2007.

[23] CAPTCHA: Telling humans and computers apart automatically, http://www.captcha.net/.

[24] X. Carreras and L. Márquez, "Boosting trees for anti-spam email filtering," in *Proceedings of RANLP-2001, 4th International Conference on Recent Advances in Natural Language Processing*, 2001.

[25] K. Chellapilla, K. Larson, K. Simard, and M. Czerwinski, "Designing human friendly human interactive proofs (HIPS)," in *CHI '05: SIGCHI Conference on Human Factors in Computing Systems*, pp. 711–720, 2005.

[26] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski, "Computers beat humans at single character recognition in reading based human interaction

proofs," in *CEAS 2005 — The Second Conference on Email and Anti-Spam*, 2005.

[27] S. Chhabra, *Fighting Spam, Phishing and Email Fraud.* University of California, Riverside, 2005.

[28] A. Ciltik and T. Gungor, "Time-efficient spam e-mail filtering using n-gram models," *Pattern Recognition Letters*, vol. 29, pp. 19–33, 2008.

[29] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. 32, pp. 396–402, April 1984.

[30] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the 12th International Conference on Machine Learning*, (A. Prieditis and S. Russell, eds.), pp. 115–123, Tahoe City, CA: Morgan Kaufmann, July 9–12 1995.

[31] G. Cormack, J. M. G. Hidalgo, and E. P. Sánz, "Feature engineering for mobile (SMS) spam filtering," in *30th ACM SIGIR Conference on Research and Development on Information Retrieval*, Amsterdam, 2007.

[32] G. V. Cormack, "Harnessing unlabeled examples through iterative application of Dynamic Markov Modeling," in *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Berlin, 2006.

[33] G. V. Cormack, "TREC 2006 Spam Track Overview," in *Fifteenth Text REtrieval Conference (TREC-2006)*, Gaithersburg, MD: NIST, 2006.

[34] G. V. Cormack, "TREC 2007 Spam Track Overview," in *Sixteenth Text REtrieval Conference (TREC-2007)*, Gaithersburg, MD: NIST, 2007.

[35] G. V. Cormack, "University of waterloo participation in the TREC 2007 spam track," in *Sixteenth Text REtrieval Conference (TREC-2007)*, Gaithersburg, MD: NIST, 2007.

[36] G. V. Cormack and A. Bratko, "Batch and on-line spam filter evaluation," in *CEAS 2006: The Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[37] G. V. Cormack, J. M. G. Hidalgo, and E. P. Sánz, "Spam filtering for short messages," in *CIKM '07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 313–320, USA, New York, NY: ACM Press, 2007.

[38] G. V. Cormack and R. N. S. Horspool, "Data compression using dynamic Markov modelling," *The Computer Journal*, vol. 30, no. 6, pp. 541–550, 1987.

[39] G. V. Cormack and T. R. Lynam, *TREC Spam Filter Evaluation Toolkit.* http://plg.uwaterloo.ca/˜gvcormac/jig/.

[40] G. V. Cormack and T. R. Lynam, "TREC 2005 Spam Track Overview," http://plg.uwaterloo.ca/˜gvcormac/trecspamtrack05, 2005.

[41] G. V. Cormack and T. R. Lynam, "Statistical precision of information retrieval evaluation," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 533–540, USA, New York, NY: ACM Press, 2006.

[42] G. V. Cormack and T. R. Lynam, "On-line supervised spam filter evaluation," *ACM Transactions on Information Systems*, vol. 25, no. 3, 2007.

[43] D. Crocker, "Challenges in Anti-spam Efforts," *The Internet Protocol Journal*, vol. 8, no. 4, http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_8-4/anti-spam_efforts.html, 2006.

[44] N. N. Dalvi, P. M. Domingos, S. K. Sanghai, and D. Verma, "Adversarial classification," in *KDD*, (W. Klm, R. Kohavi, J. Gehrke, and W. DuMouchel, eds.), pp. 99–108, 2004.

[45] R. Dantu and P. Kolan, "Detecting spam in VoIP networks," in *SRUTI'05: Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop*, pp. 5–5, USA, Berkeley, CA: USENIX Association, 2005.

[46] J. Deguerre, "The mechanics of Vipul's Razor," *Network Security*, pp. 15–17, September 2007.

[47] S. J. Delany, P. Cunningham, and L. Coyle, "Case-based reasoning for spam filtering," *Artificial Intelligence Review*, vol. 24, no. 3–4, pp. 359–378, 2005.

[48] T. Dietterich, *Statistical Tests for Comparing Supervised Classification Learning Algorithms.* Oregon State University, 1996.

[49] V. Dimitrios and E. M. Ion Androutsopoulos, "A game-theoretic investigation of the effect of human interactive proofs on spam e-mail," in *CEAS 2007 — The Fourth Conference on Email and Anti-Spam*, 2007.

[50] N. Dimmock and I. Maddison, "Peer-to-peer Collaborative Spam Detection," *ACM Crossroads*, vol. 11, no. 2, 2004.

[51] P. Domingos and M. J. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2–3, pp. 103–130, 1997.

[52] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, "Learning fast classifiers for image spam," in *CEAS 2007 — The Third Conference on Email and Anti-Spam*, 2007.

[53] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *IEEE-NN*, vol. 10, no. 5, pp. 1048–1054, 1999.

[54] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999.

[55] C. Dwork and M. Naor, "Pricing via processing or combatting junk mail," in *CRYPTO '92*, 1992.

[56] ECML/PKDD Discovery Challenge, http://www.ecmlpkdd2006.org/challenge.htm, 2006.

[57] T. Fawcett, "'In vivo' spam filtering: A challenge problem for data mining," *KDD Explorations*, vol. 5, no. 2, December 2003.

[58] T. Fawcett, *ROC Graphs: Notes and Practical Considerations for Researchers.* HP Laboratories, 2004. http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf.

[59] D. Ferris and R. Jennings, "*Calculating the Cost of Spam for Your Organization*," http://http://www.ferris.com/?p=310061, 2005.

[60] D. Ferris, R. Jennings, and C. Williams, *The Global Economic Impact of Spam.* Ferris Research, http://www.ferris.com/?p=309942, 2005.

[61] Final ultimate solution to the spam problem, http://craphound.com/spamsolutions.txt.

[62] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded into images," *Journal of Machine Learning Research*

(special issue on Machine Learning in Computer Security), vol. 7, pp. 2699–2720, 12/2006 2006.

[63] W. Gansterer, A. Janecek, and R. Neumayer, "Spam filtering based on latent semantic indexing," in *SIAM Conference on Data Mining*, 2007.

[64] K. R. Gee, "Using latent semantic indexing to filter spam," in *SAC '03: Proceedings of the 2003 ACM symposium on Applied Computing*, pp. 460–464, USA, New York, NY: ACM Press, 2003.

[65] Z. Ghahramani, "Unsupervised learning," in *Advanced Lectures in Machine Learning*, pp. 72–112, Lecture Notes in Computer Science, vol. 3176, 2004.

[66] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: A single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.

[67] J. Goodman, D. Heckerman, and R. Rounthwaite, "Stopping spam," *Scientific American*, vol. 292, pp. 42–88, April 2005.

[68] J. Goodman and W.-T. Yih, "Online discriminative spam filter training," in *The Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[69] P. Graham, *Better Bayesian Filtering.* http://www.paulgraham.com/better.html, 2004.

[70] J. Graham-Cumming, "How to beat an adaptive spam filter," in *The Spam Conference*, 2004.

[71] J. Graham-Cumming, "People and spam," in *The Spam Conference*, 2005.

[72] J. Graham-Cumming, "Does Bayesian poisining exist?," *Virus Bulletin*, February 2006.

[73] J. Graham-Cumming, "SpamOrHam," *Virus Bulletin*, 2006-06-01.

[74] J. Graham-Cumming, "The rise and fall of image-based spam," *Virus Bulletin*, 2006-11-01.

[75] J. Graham-Cumming, "The spammer's compendium: Five yars on," *Virus Bulletin*, 2007-09-20.

[76] J. Graham-Cumming, "Why I hate challenge-response," *JGC's Anti-Spam Newsletter*, February 28, 2005.

[77] Greylisting: The next step in the spam control war, http://projects.puremagic.com/greylisting/, 2003.

[78] B. Guenter, *Spam Archive.* http://www.untroubled.org/spam/.

[79] K. Gupta, V. Chaudhary, N. Marwah, and C. Taneja, *ECML-PKDD Discovery Challenge Entry.* Inductis India Pvt Ltd, 2006.

[80] K. Gupta, V. Chaudhary, N. Marwah, and C. Taneja, "Using positive-only learning to deal with the heterogeneity of labeled and unlabeled data," in *Proceedings of ECML/PKDD Discovery Challenge Workshop*, Berlin, 2006.

[81] B. Hayes, "How many ways can you spell Viagra?," *American Scientist*, vol. 95, 2007.

[82] J. M. G. Hidalgo, "Evaluating cost-sensitive unsolicited bulk email categorization," in *Proceedings of SAC-02, 17th ACM Symposium on Applied Computing*, pp. 615–620, Madrid, ES, 2002.

[83] J. M. G. Hidalgo, "Evaluating cost-sensitive unsolicited bulk email categorization," in *SAC '02: Proceedings of the 2002 ACM Symposium on Applied Computing*, pp. 615–620, Madrid: ACM Press, March 2002.

[84] J. M. G. Hidalgo, G. C. Bringas, E. P. Sanz, and F. C. Garcia, "Content based SMS spam filtering," in *DocEng '06: Proceedings of the 2006 ACM Symposium on Document Engineering*, pp. 107–114, USA, New York, NY: ACM Press, 2006.

[85] J. M. G. Hidalgo, M. M. López, and E. P. Sanz, "Combining text and heuristics for cost-sensitive spam filtering," in *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*, pp. 99–102, USA, Morristown, NJ: Association for Computational Linguistics, 2000.

[86] S. Holden, "Spam Filtering II," *Hakin9, 02/2004*, pp. 68–77, 2004.

[87] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression.* New York: Wiley, 2000.

[88] J. Hovold, "Naive bayes spam filtering using word-position-based attributes," in *Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS 2005)*, Palo Alto, CA, July 2005.

[89] M. Ilger, J. Strauss, W. Gansterer, and C. Proschinger, *The Economy of Spam.* Vol. FA384018-6, Instituted of Distributed and Multimedia Systems, Univeristy of Vienna, 2006.

[90] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *Proceedings of ICML-97, 14th International Conference on Machine Learning*, (D. H. Fisher, ed.), pp. 143–151, US, Nashville, San Francisco: Morgan Kaufmann Publishers, 1997.

[91] T. Joachims, *Transductive Inference for Text Classification Using Support Vector Machines.* 1999.

[92] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Integelligence*, pp. 338–345, 1995.

[93] Y. Junejo and A. Karim, "A two-pass statistical approach for automatic personalized spam filtering," in *Proceedings of ECML/PKDD Discovery Challenge Workshop*, Berlin, 2006.

[94] B. Klimt and Y. Yang, "Introducing the Enron corpus," in *CEAS 2004 — The Conference on Email and Anti-Spam*, 2004.

[95] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, pp. 1137–1145, 1995.

[96] A. Kolcz and J. Alspector, "SVM-based filtering of E-mail spam with content-specific misclassification costs," *TextDM 2001 (IEEE ICDM-2001 Workshop on Text Mining)*, 2001.

[97] A. Kolcz and A. Chowdhury, "Hardening fingerprints by context," in *CEAS 2007 — The Third Conference on Email and Anti-Spam*, 2007.

[98] A. Kolcz and A. Chowdhury, "Lexicon randomization for near-duplicate detection with I-match," *Journal of Supercomputing*, vol. DOI 10.1007/s11227-007-0171-z, 2007.

[99] A. Kolcz, A. Chowdhury, and J. Alspector, "The impact of feature selection on signature-driven spam detection," in *CEAS 2004 — The Conference on Email and Anti-Spam*, 2004.

[100] P. Komarek and A. Moore, “Fast robust logistic regression for large sparse datasets with binary outputs,” in *Artificial Intelligence and Statistics*, 2003.

[101] A. Kornblum, “Searching for John Doe: Finding spammers and phishers,” in *CEAS 2004 — The Conference on Email and Anti-Spam*, 2004.

[102] S. Kotsiantis, “Supervised learning: A review of classification techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.

[103] B. Krebs, “In the fight agains spam E-mail, Goliath wins again,” *Washington Post*, May 17 2006.

[104] H. Lee and A. Y. Ng, “Spam deobfuscation using a hidden Markov model,” in *CEAS 2005 – The Second Conference on Email and Anti-Spam*, 2005.

[105] S. Lee, I. Jeong, and S. Choi, “Dyamically weighted hidden Markov model for spam deobfuscation,” in *IJCAI 07*, pp. 2523–2529, 2007.

[106] J. R. Levine, “Experiences with greylisting,” in *CEAS 2005: Second Conference on Email and Anti-Spam*, 2005.

[107] D. D. Lewis and J. Catlett, “Heterogeneous uncertainty sampling for supervised learning,” in *Proceedings of ICML-94, 11th International Conference on Machine Learning*, (W. W. Cohen and H. Hirsh, eds.), pp. 148–156, US, New Brunswick, San Francisco: Morgan Kaufmann Publishers, 1994.

[108] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka, “Training algorithms for linear text classifiers,” in *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, (H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, eds.), (Zürich, CH), pp. 298–306, New York, US: ACM Press, 1996.

[109] K. Li, C. Pu, and M. Ahamad, “Resisting SPAM delivery by TCP damping,” in *CEAS 2004 — The Conference on Email and Anti-Spam*, 2004.

[110] B. Lieba and J. Fenton, “DomainKeys identified email (DKIM): Using digital signatures for domain verification,” in *CEAS 2007: The Third Conference on Email and Anti-Spam*, 2007.

[111] B. Lieba, J. Ossher, V. T. Rajan, R. Segal, and M. Wegman, “SMTP path analysis,” in *2nd Conference on Email and Anti-spam*, 2005.

[112] Ling-Spam, PU and Enron Corpora, http://www.iit.demokritos.gr/skel/i-config/downloads/.

[113] T. Lynam and G. Cormack, *TREC Spam Filter Evaluation Took Kit.* http://plg.uwaterloo.ca/~trlynam/spamjig.

[114] T. R. Lynam and G. V. Cormack, “On-line spam filter fusion,” in *29th ACM SIGIR Conference on Research and Development on Information Retrieval*, Seattle, 2006.

[115] J. Lyon and M. Wong, *Sender-ID: Authenticating E-mail RFC 4406.* Internet Engineering Task Force, 2006.

[116] Mail abuse prevention system, http://www.mail-abuse.com/, 2005.

[117] M. Mangalindan, “For bulk E-mailer, pestering millions offers path to profit,” *Wall Street Journal*, November 13, 2002.

[118] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” in *Proceedings of Eurospeech ’97*, pp. 1895–1898, Rhodes, Greece, 1997.

[119] R. McMillan, "US Court threatens Spamhaus with shut down," *InfoWorld*, October 09 2006.

[120] B. Medlock, "An adaptive, semi-structured language model approach to spam filtering on a new corpus," in *Proceeding of CEAS 2006 — Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[121] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Naive Bayes — Which Naive Bayes?," in *Proceedings of CEAS 2006 — Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[122] E. Michelakis, I. Androutsopoulos, G. Paliouras, G. Sakkis, and P. Stamatopoulos, "Filtron: A learning-based anti-spam filter," in *Proceedings of the 1st Conference on Email and Anti-Spam (CEAS 2004)*, Mountain View, CA, July 2004.

[123] G. Mishne and D. Carmel, *Blocking Blog Spam with Language Model Disagreement.* 2005.

[124] E. Moustakas, C. Ranganathan, and P. Duquenoy, "Chunk-Kwei: A pattern-discovery-based System for the automatic identificaton of unsolicited email messages (spam)," in *CEAS 2004 — The Conference on Email and Anti-Spam*, 2004.

[125] J. Niu, J. Xu, J. Yao, J. Zheng, and Q. Sun, "WIM at TREC 2007," in *Sixteenth Text REtrieval Conference (TREC-2007)*, Gaithersburg, MD: NIST, 2007.

[126] C. S. Oliveira, F. G. Cozman, and I. Cohen, "Splitting the unsupervised and supervised components of semi-supervised learning," in *ICML 2005 LPCTD Workshop*, 2005.

[127] R. M. Pampapathi, B. Mirkin, and M. Levene, "A suffix tree approach to email filtering," Tech. Rep., Birkbeck University of London, 2005.

[128] C. Perlich, F. Provost, and J. S. Simonoff, "Tree induction vs. logistic regression: A learning-curve analysis," *Journal of Machanic Learning and Research*, vol. 4, pp. 211–255, 2003.

[129] B. Pfahringer, "A semi-supervised spam mail detector," in *Proceedings of ECML/PKDD Discovery Challenge Workshop*, Berlin, 2006.

[130] Project Honeypot, http://www.projecthoneypot.org/.

[131] C. Pu and S. Webb, "Observed trends in spam construction techniques," in *Proceedings of CEAS 2006 — Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[132] J. R. Quinlan, *C4.5: Programs for Machine Learning.* USA, San Francisco, CA: Morgan Kaufmann Publishers Inc., 1993.

[133] A. Ramachandran, D. Dagon, and N. Feamster, "Can DNS-based blacklists keep up with bots?," in *CEAS 2006 — The Second Conference on Email and Anti-Spam*, 2006.

[134] E. S. Raymond, D. Relson, M. Andree, and G. Louis, "BogoFilter," http://bogofilter.sourceforge.net/, 2004.

[135] F. R. Rideau, *Stamps vs. Spam: Postage as a Method to Eliminate Unsolicited Email.* http://fare.tunes.org/articles/stamps_vs_spam.html, 2002.

[136] G. Robinson, "A statistical approach to the spam problem," *Linux Journal*, vol. 107, no. 3, March 2003.

[137] R. Roman, J. Zhou, and J. Lopez, "An anti-spam scheme using prechallenges," *Computer Communications*, vol. 29, no. 15, pp. 2739–2749, 2006.

[138] C. Rossow, "Anti-Spam measures of European ISPs/ESPs: A survey based analysis of state-of-the-art technologies, current spam trends and recommendations for future-oriented anti-spam concepts," *Institute for Internet Security*, August 2007.

[139] K. J. Rothman and S. Greenland, *Modern epidemiology.* Lippinscott Williams and Wilkins, 1998.

[140] R. Rowland, *Spam, Spam, Spam: The Cyberspace Wars.* CBC, http://www.cbc.ca/news/background/spam/, 2004.

[141] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, AAAI Technical Report WS-98-05, 1998.

[142] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "Stacking classifiers for anti-spam filtering of e-mail," in *Empirical Methods in Natural Language Processing (EMNLP 2001)*, pp. 44–50, 2001.

[143] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists," *Information Retrieval*, vol. 6, no. 1, pp. 49–73, 2003.

[144] M. Sasaki and H. Shinnou, "Spam detection using text clustering," in *CW '05: Proceedings of the 2005 International Conference on Cyberworlds*, pp. 316–319, USA, Washington, DC: IEEE Computer Society, 2005.

[145] R. E. Schapire and Y. Singer, "Improved boosting using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[146] K. M. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003.

[147] D. Sculley, "Online active learning methods for fast label-efficient spam filtering," in *Proceeding of the CEAS 2007 — Fourth Conference on Email and Anti-Spam*, Mountain View, CA, 2007.

[148] D. Sculley and C. E. Brodley, "Compression and machine learning: A new perspective on feature space vectors," in *Data Compression Conference (DCC 06)*, pp. 332–341, Snowbird, 2006.

[149] D. Sculley and G. V. Cormack, *Filtering Spam in the Presence of Noisy User Feedback.* Tufts University, 2008.

[150] D. Sculley and G. M. Wachman, "Relaxed online support vector machines for spam filtering," in *30th ACM SIGIR Conference on Research and Development on Information Retrieval*, Amsterdam, 2007.

[151] D. Sculley and G. M. Wachman, "Relaxed online SVMs in the TREC Spam filtering track," in *Sixteenth Text REtrieval Conference (TREC-2007)*, Gaithersburg, MD: NIST, 2007.

[152] D. Sculley, G. M. Wachman, and C. E. Brodley, "Spam classification with on-line linear classifiers and inexact string matching features," in *Proceedings*

of the 15th Text REtrieval Conference (TREC 2006), Gaithersburg, MD, November 2006.

[153] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[154] R. Segal, J. Crawford, J. Kephart, and B. Leiba, "SpamGuru: An enterprise anti-spam filtering system," in *First Conference on Email and Anti-Spam (CEAS)*, 2004.

[155] R. Segal, T. Markowitz, and W. Arnold, "Fast uncertainty sampling for labeling large e-mail corpora," in *Proceedings of the CEAS 2006 — Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[156] Shakhnarovish, Darrell, and Indyk, *Nearest-Neighbor Methods in Learning and Vision,* (Shakhnarovish, ed.), MIT Press, 2005.

[157] V. Sharma and A. O'Donnell, "Fighting spam with reputation systems," *ACM Queue*, November 2005.

[158] C. Siefkes, F. Assis, S. Chhabra, and W. S. Yerazunis, "Combining winnow and orthogonal sparse bigrams for incremental spam filtering," in *PKDD '04: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 410–421, USA, New York, NY: Springer-Verlag, New York, Inc., 2004.

[159] P. Y. Simard, R. Szeliski, J. Benaloh, J. Couvreur, and I. Calinov, "Using character recognition and segmentation to tell computer from humans," in *ICDAR '03: Seventh International Conference on Document Analysis and Recognition*, 2003.

[160] J. Snyder, "Spam in the wild, the sequel," *Network World 12/20/04*, 2004.

[161] E. Solan and E. Reshef, "The effects of anti-spam methods on spam mail," in *CEAS 2006 — The Third Conference on Email and Anti-Spam*, 2006.

[162] Spam testing methodology, http://www.opus1.com/www/whitepapers/spamtestmethodology.pdf, 2007.

[163] spamassassin.org, *The Spamassassin Public Mail Corpus*. http:// spamassassin.apache.org/publiccorpus, 2003.

[164] spamassassin.org, *Welcome to SpamAssassin*. http://spamassassin.apache.org, 2005.

[165] Spambase, http://mlearn.ics.uci.edu/databases/spambase/.

[166] J. A. Swets, "Effectiveness of information retrieval systems," *American Documentation*, vol. 20, pp. 72–89, 1969.

[167] T. Takemura and H. Ebara, "Spam mail reduces economic effects," in *Second International Conference on the Digital Society*, pp. 20–24, 2008.

[168] W. Tau Yih, R. McCann, and A. Kolcz, "Improving spam filtering by detecting gray mail," in *Proceedings of CEAS 2007 — Fourth Conference on Email and Anti-Spam*, Mountain View, CA, 2007.

[169] D. M. J. Tax and C. Veenman, "Tuning the hyperparameter of an AUC-optimized classifier," in *Seventeenth Belgium-Netherlands Conference on Artificial Intelligence*, pp. 224–231, 2005.

[170] The CEAS 2007 Live Spam Challenge, http://www.ceas.cc/2007/challenge/challenge.html, 2007.

[171] The penny black project, http://research.microsoft.com/research/sv/Penny Black/.

[172] TREC 2005 Spam Corpus, http://plg.uwaterloo.ca/˜gvcormac/treccorpus, 2005.

[173] TREC 2006 Spam Corpora, http://plg.uwaterloo.ca/˜gvcormac/treccorpus, 2006.

[174] TREC 2007 Spam Corpus, http://plg.uwaterloo.ca/˜gvcormac/treccorpus, 2007.

[175] K. Tretyakov, "Machine learning techniques in spam filtering," Tech. Rep., Institute of Computer Science, University of Tartu, 2004.

[176] N. Trogkanis and G. Paliouras, "Using positive-only learning to deal with the heterogeneity of labeled and unlabeled data," in *Proceedings of ECML/PKDD Discovery Challenge Workshop*, Berlin, 2006.

[177] H. Tschabitscher, *What you Need to Know about Challenge-Response Spam Filters.* http://email.about.com/cs/spamgeneral/a/challenge_resp.htm.

[178] D. Turner, M. Fossi, E. Johnson, T. Mack, J. Blackbird, S. Entwisle, M. K. Low, D. McKinney, and C. Wueest, *Symantec Global Internet Security Threat Report: Trends for July-December 07.* Symantec, http://eval.symantec.com/mktginfo/enterprise/white_papers/b-whitepaper_internet_security_threat_report_xiii_04-2008.en-us.pdf, 2007.

[179] A. Tuttle, E. Milios, and N. Kalyaniwalla, "An evaluation of machine learning techniques for enterprise spam filters," Technical Report CS-2004-03, Halifax, NS: Dalhousie University, 2004.

[180] C. J. Van Rijsbergen, *Information Retrieval.* Department of Computer Science, University of Glasgow, Second ed., 1979.

[181] Veritest Anti-Spam Benchmark Service Autumn 2005 Report http://www.tumbleweed.com/pdfs/VeriTest_Anti-Spam_Report_Vol4_all_c.pdf, 2005.

[182] E. Voorhees, *Fourteenth Text REtrieval Conference (TREC-2005).* Gaithersburg, MD: NIST, 2005.

[183] E. Voorhees, *Fifteenth Text REtrieval Conference (TREC-2005).* Gaithersburg, MD: NIST, 2006.

[184] E. Voorhees, *Sixteenth Text REtrieval Conference (TREC-2005).* Gaithersburg, MD: NIST, 2007.

[185] E. M. Voorhees and D. K. Harman, eds., *TREC — Experiment and Evaluation in Information Retrieval.* Boston: MIT Press, 2005.

[186] Z. Wang, W. Josephson, Q. LV, M. Charikar, and K. Li, "Filtering image spam with near-duplicate detection," in *CEAS 2007 — The Third Conference on Email and Anti-Spam*, 2007.

[187] Web Spam Challenge, 2008.

[188] S. Webb, J. Caverloo, and C. Pu, "Introducing the webb spam corpus: Using email spam to identify web spam automatically," in *Proceedings of CEAS 2006 — Third Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[189] West Coast Labs, http://www.westcoastlabs.com.

[190] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[191] G. L. Wittel and S. F. Wu, "On attacking statistical spam filters," in *CEAS 2004 — The Conference on Email and Anti-Spam*, 2004.

[192] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, *Weka: Practical Machine Learning Tools and Techniques with Java Implementations.* 1999.

[193] M. Wong and W. Schlitt, *Sender Policy Framework (SPF) for Authorizing Use of Domains in E-mail.* Vol. RFC 4408, 2006.

[194] W. Yerazunis, *Correspondence with Paul Graham.* http://www. paulgraham.com/wsy.html, 16 October 2002.

[195] W. S. Yerazunis, *CRM114 — the Controllable Regex Mutilator.* http://crm114.sourceforge.net/, 2004.

[196] W. S. Yerazunis, "The spam-filtering accuracy plateau at 99.9% accuracy and how to get past it," in *2004 MIT Spam Conference*, January 2004.

[197] W. S. Yerazunis, "Seven hypothesis about spam filtering," in *Proceedings 15th Text REtrieval Conference (TREC 2006)*, NIST, Gaithersburg, MD, November 2006.

[198] W. S. Yerazunis, "Seven Hypothesis about Spam," in *Sixteenth Text REtrieval Conference (TREC-2007)*, Gaithersburg, MD: NIST, 2007.

[199] W. Yih, J. Goodman, and G. Hulten, "Learning at low false positive rates," in *Proceedings of the 3rd Conference on Email and Anti-Spam*, 2006.

[200] X. Yue, A. Abraham, Z.-X. Chi, Y.-Y. Hao, and H. Mo, "Artificial immune system inspired behavior-based anti-spam filter," *Soft Computing*, vol. 11, pp. 729–740, 2007.

[201] L. Zhang, J. Zhu, and T. Yao, "An evaluation of statistical spam filtering techniques," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 4, pp. 243–269, 2004.

[202] W. Zhao and Z. Zhang, "An email classification model based on rough set theory," in *Active Media Technology, 2005. (AMT 2005)*, 2005.

[203] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *18th Annual Conference on Neural Information Processing Systems*, 2003.

[204] X. Zhu, *Semi-supervised Learning Literature Survey.* Vol. TR 1530, University of Wisconsin, 2007.

[205] A. Zien, "Semi-supervised support vector machines and application to spam filtering," in *Oral Presentation, ECML/PKDD Discovery Challenge Workshop*, Berlin, 2006.

[206] A. Zinman and J. Donath, "Is Britney Spears spam?," in *Proceedings of CEAS 2007 — Fourth Conference on Email and Anti-Spam*, Mountain View, CA, 2007.