"In vivo" spam filtering: A challenge problem for KDD

Tom Fawcett
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA USA
tom.fawcett@hp.com

ABSTRACT

Spam, also known as Unsolicited Commercial Email (UCE), is the bane of email communication. Many data mining researchers have addressed the problem of detecting spam, generally by treating it as a static text classification problem. True *in vivo* spam filtering has characteristics that make it a rich and challenging domain for data mining. Indeed, real-world datasets with these characteristics are typically difficult to acquire and to share. This paper demonstrates some of these characteristics and argues that researchers should pursue *in vivo* spam filtering as an accessible domain for investigating them.

General Terms

spam, text classification, challenge problems, class skew, imbalanced data, cost-sensitive learning, data streams, concept drift

1. INTRODUCTION

Spam, also known as Unsolicited Commercial Email (UCE) and Unsolicited Bulk Email (UBE), is commonplace everywhere in email communication[1]. Spam is a costly problem and many experts agree it is only getting worse [7; 24; 31; 34; 14]. Because of the economics of spam and the difficulties inherent in stopping it, it is unlikely to go away soon.

Many data mining and machine learning researchers have worked on spam detection and filtering, commonly treating it as a basic text classification problem. The problem is popular enough that it has been the subject of a Data Mining Cup contest [10] as well as numerous class projects. Bayesian analysis has been very popular [28; 30; 16; 3], but researchers have also used SVMs [20], decisions trees [4], memory and case-based reasoning [29; 8], rule learning [27] and even genetic programming [19].

But researchers who treat spam filtering as an isolated text classification task have only addressed a portion of the problem. This paper argues that real-world *in vivo* spam filtering is a rich and challenging problem for data mining. By "*in vivo*" we mean the problem as it is truly faced in an operating environment, that is, by an on-line filter on a mail account that receives realistic feeds of email over time, and serves a human user. In this context, spam filtering faces issues of skewed and changing class distributions; unequal and uncertain error costs; complex text patterns; a complex, disjunctive and drifting target concept; and challenges of intelligent, adaptive adversaries. Many real-world domains share these characteristics and would benefit indirectly by work on spam filtering.

Improving spam filtering is a worthy goal in itself, but this paper takes the (admittedly selfish) position that data mining researchers should study the problem for the benefit of data mining. It is unclear whether spam filtering efforts could genuinely benefit from data mining research. On the other hand, one of the persistent difficulties of research in many real-world domains is that of acquiring and sharing datasets. Most companies, for example, do not release datasets containing real customer transactions; we are aware of no public domain datasets containing genuine fraudulent transactions for studying fraud detection. Even sharing such data between partner companies usually requires formal non-disclosure agreements. In other domains datasets may still have copyright or privacy issues. Few datasets involving concept drift or changing class distributions are publicly available. Without such datasets, the ability to replicate results and compare algorithm performance is hindered and progress on these research topics will be impaired. Spam data are easily accessible and shareable, which makes spam filtering a good domain testbed for investigating many of the same issues.

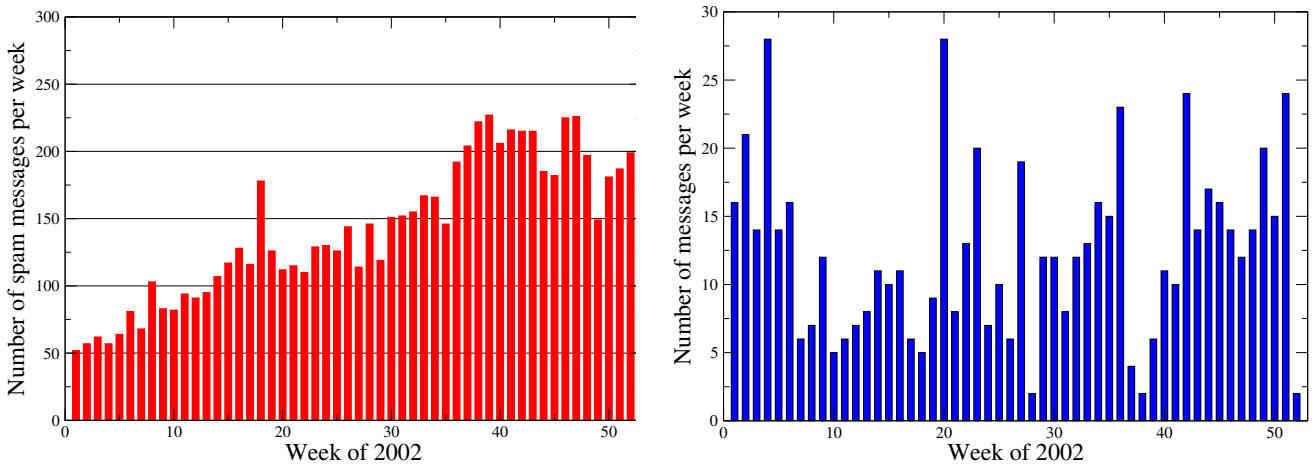
The remainder of the paper enumerates these research issues and describes how they are manifested in *in vivo* spam filtering. The final section of the paper describes how researchers could begin exploring the domain.

2. CHALLENGES

2.1 Skewed and drifting class distributions

Like most text classification domains, spam presents the problem of a skewed class distribution, *i.e.*, the proportion of spam to legitimate email is uneven. There are no generally agreed upon class priors for this problem. Gómez Hidalgo [15] points out that the proportion of spam messages reported in research datasets varies considerably, from 16.6% to 88.2%. This may be simply because the proportion varies considerably from one individual to another. The amount of spam received depends on the email address, the degree

[1]The term "spam" is sometimes used loosely to mean any message broadcast to multiple senders (regardless of intent) or any message that is undesired. Here we intend the narrower, stricter definition: unsolicited commercial email sent to an account by a person unacquainted with the recipient.



(a) Spam messages (b) Legitimate messages

Figure 1: Weekly variation in message traffic, spam versus legitimate email

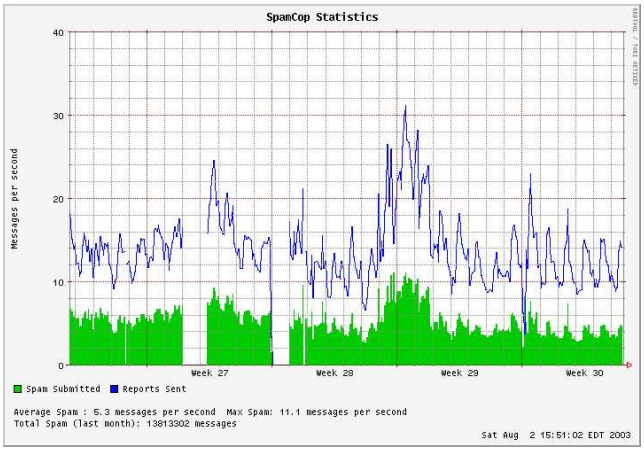


Figure 2: SpamCop: Spam forwarded and reports sent

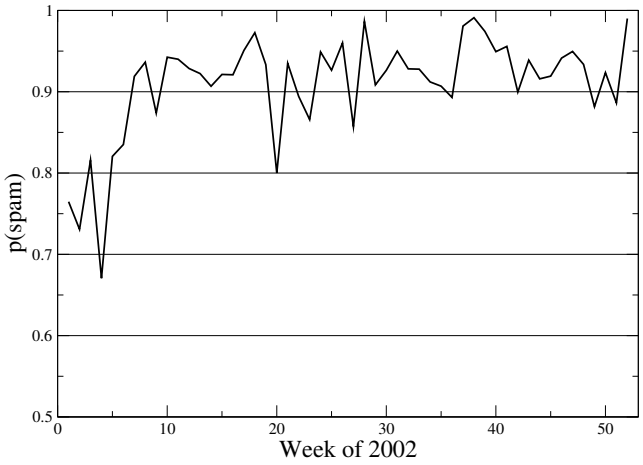


Figure 3: Drifting priors: weekly estimates of p(spam) taken from data in figure 1.

of exposure, the amount of time the address has been public and the upstream filtering. The amount of legitimate email received similarly varies greatly from one individual to another.

Perhaps more importantly, spam varies over *time* as well. This was demonstrated dramatically in 2002 when a large number of open relays and open proxies were brought on-line in Asian countries, primarily Korea and China. Such a large new pool of unprotected machines provided great opportunities for spammers, and soon email servers throughout the world experienced a huge surge in the amount of spam they forwarded and received. The problem became so bad that for a brief time all email from certain Asian countries was blocked completely by some ISPs [9].

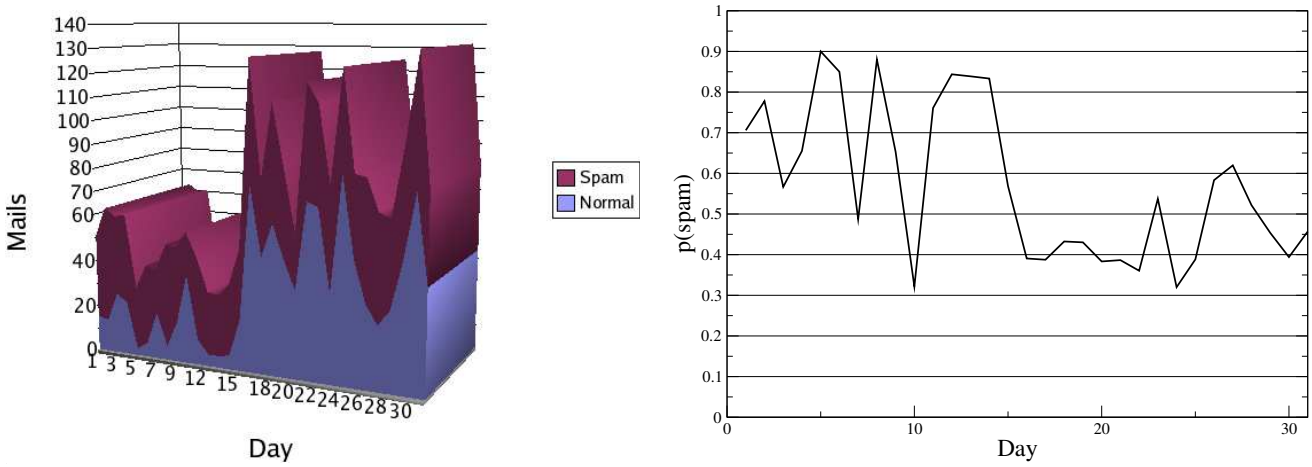
In spite of claims that spam is generally increasing [7; 24; 5], the volume varies considerably and non-monotonically on a daily or weekly scale. Calculating spam proportion even approximately is difficult. Although some public spam datasets are available (see Appendix A), we are aware of no personal email datasets arranged over time, so it is difficult to match the two to establish priors. Nevertheless, using several datasets we can make a case that spam priors change significantly over time.

Figure 1a shows a graph of spam volume received in 2002 by Paul Wouters of Xtended Internet[2]. In 2002 the spam volume was 146 ± 55 messages per week, indicating a great deal of variation in spite of its upward trend. For most people, the volume of the legitimate email received varies as well. Figure 1b shows a graph of the number of legitimate messages saved by the author over the weeks in 2002. The volume is 12.3 ± 6.4 messages per week.

Figure 2 shows the volume of reports issued from SpamCop's website[3] This graph also demonstrates some of spam's episodic nature. SpamCop is a service used by many people to filter spam and to submit reports (complaints) to the originators of spam. Both the amount of spam submitted and the number of reports sent show clear episodic behavior. These graphs show time variation in both the volume of

[2] `http://spamarchive.xtdnet.nl/`

[3] `http://www.spamcop.net/spamstats.shtml`



(a) (b)

Figure 4: Email volume from Eide's trial, (a) absolute volume, (b) resulting prior p(spam).

spam and the volume of legitimate email received, something that researchers have not generally acknowledged. Since the two sources of email—senders of spam and senders of legitimate email—are independent parties with little in common, we can expect their variation to be statistically uncorrelated, and the class priors will vary over time. No fixed prior will be correct.

How much could we expect class priors to vary? If we assume that a user received the spam shown in figure 1a and the legitimate email shown in figure 1b, we can estimate the class prior p(spam) simply as the proportion of weekly messages that are spam. Figure 3 shows a graph of this value, which ranges between about .67 to .99.

A further demonstration of changing priors appears in Kristian Eide's study of bayesian spam filters [12]. In evaluating these filters he measured the volume of spam and legitimate email he received over the course of one month. These volumes are graphed in figure 4a, and the computed daily spam prior is graphed in figure 4b. The prior ranges from .32 to .9, showing greater variation than in figure 3, though the skew is not as high.

Variation in class priors may be problematic for researchers because it makes solution superiority more difficult to establish. A classifier that performs better than another on a dataset with 80% spam may perform worse on one with 40% spam [26].

Should researchers be concerned about these varying class priors? This question is difficult to answer conclusively because it depends on classifier performance as well as error cost assumptions (discussed in section 2.2). But by employing the cost curve framework of Drummond and Holte [11], we can answer a related question, *How much of cost space is influenced by this variation?* This question can be answered by calculating the span of the Probability Cost Function (PCF), which is the x axis of a cost curve. The PCF ranges from zero to one and is a function of the class prevalence and the ratio of misclassification error costs. In the case of spam:

$$\mathrm{PCF}_{\mathrm{spam}} = \frac{p(\mathrm{spam}) \cdot \mathrm{cost}(FN)}{p(\mathrm{spam}) \cdot \mathrm{cost}(FN) + p(\mathrm{legit}) \cdot \mathrm{cost}(FP)}$$

If we assume that the cost of a false positive (that is, of classifying a legitimate message as spam) is about ten times that of a false negative, this reduces to

$$\mathrm{PCF}_{\mathrm{spam}} = \frac{p(\mathrm{spam})}{p(\mathrm{spam}) + p(\mathrm{legit}) \times 10}$$

Using the p(spam) range from figure 4b, the PCF range of interest for spam filtering is $.04 \leq \mathrm{PCF}_{\mathrm{spam}} \leq .47$. Since the entire PCF range is [0,1], this means nearly half of cost space is influenced by this variation in priors. Any classifier whose performance lies within this 44% could be a competitive solution. This is a wide range, and it is reasonable to expect classifier superiority to vary within it.

The purpose of this analysis is not to call into question the validity of prior work, but to point out that changing class distributions are a reality in this domain and their influence on solutions should be tested. Conversely, researchers investigating skewed and varying class distributions would do well to study the spam filtering problem.

Exactly how a researcher should best track and adjust class priors is an open question and will require research. Time series work in statistics should provide some strategies, for example, using an exponentially decayed average of recent priors. However estimation is done, researchers should acknowledge that priors change and static values are unrealistic.

2.2 Unequal and uncertain error costs

A further complication of *in vivo* filtering is the asymmetry in error costs. Judging a legitimate email to be spam (a false positive error) is usually far worse than judging a spam email to be legitimate (a false negative error). A false negative simply causes slight irritation, *i.e.*, the user sees an undesirable message. A false positive can be critical. If spam is deleted permanently from a mail server, a false positive can be very expensive since it means a (possibly important) message has been discarded without a trace. If spam is moved to a low-priority mail folder for later human scanning, or if the address is only used to receive low priority email, false positives may be much more tolerable.

In an essay on developing a bayesian spam filter, Paul Graham [16] describes the different errors in an insightful comment:

> False positives seem to me a different kind of error from false negatives. Filtering rate is a measure of performance. False positives I consider more like bugs. I approach improving the filtering rate as optimization, and decreasing false positives as debugging.

Ken Schneider, CTO of the mail filtering company BrightMail, makes the same point more starkly [31]. He argues that filtering even a small amount of legitimate email defeats the purpose of filtering because it forces the user to start reviewing the spam folder for missed messages. Even a single missed important message may cause a user to reconsider the value of spam filtering. This argues for assigning a very high false positive error cost.

Regardless of the exact values, these asymmetric error costs must be acknowledged and taken into account by any acceptable filtering solution. Judging a spam filtering system by accuracy (or, equivalently, error rate) is unrealistic and misleading [26]. Some researchers have measured precision and recall without questioning whether metrics for information retrieval are appropriate for a filtering task.

Fortunately, most researchers have acknowledged these asymmetric costs, but methods of dealing with them have been ad hoc. The 2003 Data Mining Cup Competition [10] required that learned classifiers have no more than a 1% false positive rate, but the organizers gave no justification for this cut-off. Graham [16] simply double-counted the tokens of his legitimate email, essentially considering the cost of a false positive to be twice that of a false negative. Sahami et al. [28] used a very high probability threshold of .999 for classifying a message as spam. Androutsopoulos et al. [2] performed more careful experiments across cost ratios of 1, 10 and 100, exploring two orders of magnitude of cost ratios.

These approaches suggest a deeper issue: true costs of filtering errors may simply be unknown to the data mining researcher, or may be known only approximately. Only the end user will know the consequences of filtering mistakes and be able to estimate error tradeoffs. *In vivo* filtering requires flexibility of solutions: the user should be able to specify the approximate costs (or relative severity) of the errors and the run-time filter should accommodate. Admittedly this requirement complicates research evaluation since the superiority of an approach may not extend throughout a cost range, and multiple experiments may have to be performed.

Such uncertainty is actually common in real-world domains, where experts may have difficulty stating the exact cost of an erroneous action, or the cost of the action may vary depending on external circumstances. This situation motivated development of a framework based on ROC analysis for evaluating and managing classifiers when error costs are uncertain [25]. In the case of spam filtering, the uncertainty of error costs may not change temporally but they do vary between users. Gómez Hidalgo [15] used this framework for developing and evaluating spam filtering solutions, and found it useful. Drummond and Holte [11] have also developed a cost curve framework that extends ROC analysis and serves much the same purpose. Whatever technique is used for evaluating classifier performance, researchers should be prepared to demonstrate a solution's performance over a range of costs.

2.3 Disjunctive and changing target concept

Section 2.1 made the case that the amount of spam drifts over time, so class distributions vary. It is also true that the *content* of spam changes over time, so class-conditioned feature probabilities will change as well.

Some spam topics are perpetual, such as advertisements for pornography sites, offers for mortgage re-financing, and moneymaking schemes. Other topics are bursty or occur in epidemics.

One notorious example of a spam ploy coming into vogue is the "Nigerian Money" scam, a get-rich-quick scam in which help was solicited to transfer money from a Nigerian bank account [32]. The details varied, but the sender usually claimed to be responsible for a large bank account and requested assistance in "liberating" the funds from the Nigerian government. The sender was willing to pay generously for access to a foreign bank account into which the money would be transfered. This account was usually drained of funds once access was granted. Eventually the people responsible for the scam were arrested, and spam of that type declined quickly (unfortunately, variants continue to circulate as other people adopt the general idea). Prior to this scam, keywords such as `nigeria` and `assistance` were not strong predictors of spam.

A more dramatic episode occurred in April of 2003 when decks of playing cards depicting "Iraq's Most Wanted" were made available for sale. These cards were advertised primarily via spam. The advertising campaign created such a spam blizzard that its story—and the campaign's success — were written up in the New York Times [18]. This campaign abated quickly and few of the terms uniquely associated with this episode retain much predictive power now.

The point for researchers is that spam content changes over time so the "spam" concept should drift inevitably. Some components (disjuncts) of the concept description should remain constant or change only slowly. Others will spike during epidemics, as specific scams or merchandising schemes come into vogue. Even perpetual topics do not exhibit constant term frequencies.

It is difficult to estimate how much we can expect "spam" as a concept to drift over time, in part because no metric of concept drift has been adopted by the community. It is beyond the scope of this paper to present a rigorous investigation of concept drift in spam, but a simple technique can demonstrate significant word frequency variation.

Swan and Allan [33] employed a χ^2 test to discover "bursty" topics in daily news stories. Their test was designed to determine whether the appearance of a term on a given day was statistically significant. This test can be applied to weekly groups of spam messages in Wouters' archive. The results are shown in figure 5, with selected terms listed on the left side and a column for every week (1–52) of 2002 extending to the right. The height of a bar at a term-week is proportional to the term's frequency in that week. The special symbol "⊠" denotes a term burst: it appears in a term's row if that term appeared more than four times in the week and the χ^2 test succeeded at $p < 0.01$.

Figure 5 shows that spam has complex time-varying behavior. Some terms recur intermittently, such as adult, click, free, hot and removed. Others are episodic, *e.g.*, the terms

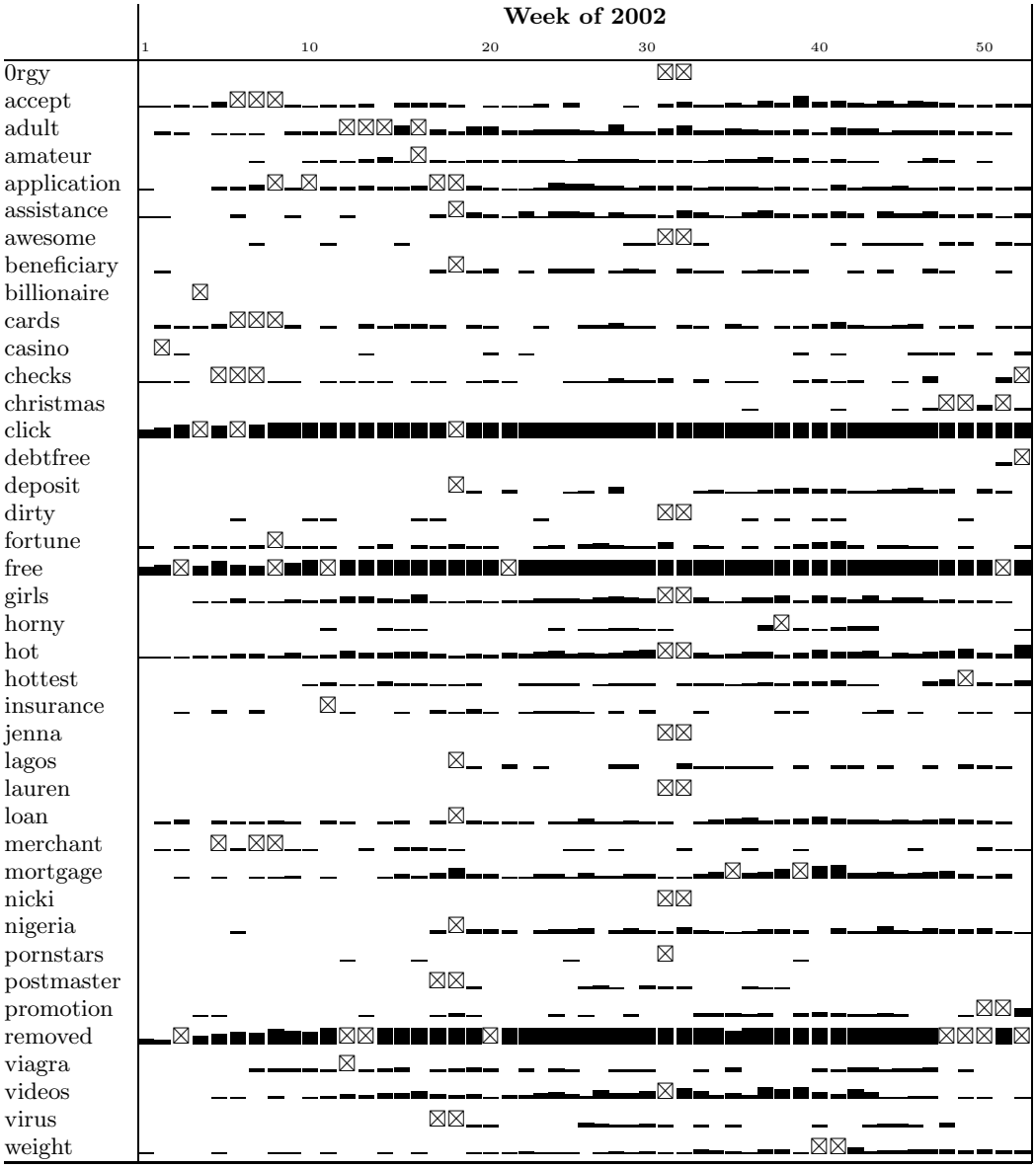


Figure 5: Frequency and burstiness of spam terms.

common in a "Nigerian scam" burst in week 18 (nigeria, lagos, assistance, beneficiary) and terms in a "pornstar videos" burst in weeks 31–32 (0rgy, awesome, pornstars, jenna, lauren, nicki). The term christmas bursts late in the year and presumably reappears every year around the same time.

Spam behavior is not simply a matter of one concept drifting to another in succession, but instead is a superimposition of constant, periodic and episodic phenomena. Researchers in data mining have studied classification under concept drift but it remains an open problem. Work in the field of Topic Detection and Tracking [1] is likely to be relevant to spam classification, though technically it addresses a different problem. No detailed study of real-world concept drift has yet been undertaken, and to the best of our knowledge there are no standard datasets for studying it. A longitudinal spam dataset would be an excellent testbed for investigating issues in concept drift and stream classification.

2.4 Intelligent adaptive adversaries

The spam stream changes over time as different products or scams, marketed by spam, come into vogue. There is a separate reason for concept drift: spammers are engaged in a perpetual "arms race" with email filters [7; 35].

Over time spammers have become increasingly sophisticated, both in their methods for sending out large volumes of email undetected and in their techniques for evading filtering [23]. In its early days, spam would have predictable subject lines like `MAKE MONEY FAST!` and `Refinance your mortgage`. Messages would be addressed to `Undisclosed_recipients` or `nobody`. As basic header filtering became common in e-mail clients, these obvious text markers were simple to filter upon so spam could be discarded easily. As message body scanning became common, fragments such as `viagra` and `click`

here could be checked for as well.

To circumvent simple filtering, spammers began to employ content obscuring techniques such as inserting spurious punctuation, using bogus HTML tags and adding HTML comments in the middle of words. It is now common to see fragments such as these:

- `v.ia.g.ra`
- `100% Mo|ney Back Guaran|tee!`
- `Our pro<br2sd9/>duct is doctor reco<br2sd9 />mmen<br2sd9/>ded and made from 100% natu<br2sd9/>ral ingre<br2sd9/>dients.`
- `C<!--7udzl53l5spp6-->lic<!--yajiwn1xnbecx2-->k he<!--ehc0aj2pvwu-->re</a>`
- `Inċrëäsë tëstöstërönë by 254%`

When rendered, these are recognizable to most people but they foil simple word and phrase filtering. To counter this, some filters remove embedded punctuation and bogus HTML tags before scanning, and consider them to be additional evidence that the message is spam.

Spammers have also realized that filters use bayesian word analysis and content hashing. In response, they often pepper their messages with common English words and nonsense words to foil these techniques [23]. Messages are designed so these words are discarded when the text is rendered, or are rendered unobtrusively. Graham-Cumming [17] maintains an extensive catalog of the techniques used by spammers to confuse filters.

Whatever new filtering capabilities arise, it is just a matter of time before spammers find ways to evade them. Already at least one company[4] provides a service for bulk commercial emailers to determine whether (and why) a prospective email message will be filtered. In machine learning terms, spammers have a strong interest in making the "spam" and "legitimate" classes indistinguishable. Because the discrimination ability of spam filters improves continually, the resulting concept drifts.

Because of this text distortion, *in vivo* spam filtering diverges significantly from most text classification and information retrieval problems, where authors are not deliberately trying to obfuscate content and defy indexing. Researchers should expect that they will have to develop techniques unknown in these related fields. Much of the effort of developing spam filters will probably shift from feature combining (*i.e.*, experimenting with different induction algorithms) to feature *generation* (*i.e.*, devising automatic feature generation methods that can adapt to new distortion patterns).

Such an arms race is not uncommon. Co-evolving abilities appear often when access to a desired resource is simultaneously sought and blocked by intelligent, adaptive parties. Fraud analysts observe criminals developing increasingly sophisticated techniques in response to security improvements [13]. With spam, the desired resource is the attention of email users, and spam may be seen as a way of illicitly gaining access to it. Another example of an arms race occurs in e-commerce. As pricing schemes have become more sophisticated, consumers have become more adept at gaming the systems. Sophisticated "shop bots" have been developed, and on-line merchants have had to develop ways to keep pricing information from them. Both sides continue to improve their techniques. Finally, the well-publicized conflict between the music swapping networks and the American music industry shows characteristics of an arms race, as both sides develop more sophisticated methods in their battle over access to copyrighted material [21].

Such co-adaptation of intelligent agents is foreign to most data mining researchers: the data are mined and the results are deployed, but the data environment is not considered to be an active entity that will react in turn. With the internet, much information is freely and automatically available by all parties, and interactivity is the rule. I propose that the future will bring more scenarios involving feedback and co-adaptation. Data miners may have to consider the effects of mining on their task environment, and perhaps incorporate such concerns into the data mining process. Possible strategies include concealing one's deployed techniques from adversaries, incorporating deception into techniques, or simply speeding up the deployment cycle to adapt more quickly to adversaries' moves. Spam filtering could be a useful domain in which to explore such strategies.

3. MEETING THE CHALLENGE

This paper has made the case that *in vivo* spam filtering can be a complex data mining problem with difficult challenging characteristics:

- Skewed and changing class distributions
- Unequal and uncertain error costs
- Complex text patterns requiring sophisticated parsing
- A disjunctive target concept comprising superimposed phenomena with complex temporal characteristics
- Intelligent, adaptive adversaries

Researchers wishing to explore these issues would do well to study *in vivo* spam filtering. Controlled laboratory datasets exhibiting these characteristics are often difficult to acquire and to share. Spam filtering, on the other hand, is an excellent domain for investigating these problems.

Researchers wishing to pursue this domain should begin collecting longitudinal data in a controlled manner. Spam is notoriously easy to attract. Several studies have measured the extent to which various activities attract spam [22; 6], and this information may be useful. It is easy to create ad hoc email addresses (for example, through Hotmail or Yahoo) and to advertise them in a controlled manner to attract spam. Such addresses are sometimes called "spam traps" and are used by email filtering companies such as BrightMail to obtain a continuous clean feed of spam for analysis.

A more difficult problem is that of obtaining shareable corpora of non-spam email, which often contain personal details that people want to keep private. Two general approaches have been taken:

1. Researchers who have contributed personal email have sought ways to anonymize it. The contributors of the UCI "spambase" dataset achieved this by reducing the

[4] `www.assurancesys.com`

original messages to word frequencies and performing feature selection upon the set. Unfortunately, this makes it difficult for other researchers to experiment with alternative feature selection or text processing operations on the data.

Androutsopoulos et al. [2] have developed a basic encoding technique for sharing data without compromising privacy. Their software and several of their datasets are available; see Appendix A.

2. Androutsopoulos et al. [3] have suggested using messages from websites and public mailing lists as proxies for personal email. Their "Ling-spam" corpus uses messages from the moderated Linguist list. Other researchers have suggested that such messages may not be representative of the email most people receive. Whether this renders mailing list data ineffective for exploring *in vivo* spam filtering remains to be studied.

However researchers decide to generate such corpora, they should consider making their data publicly available.

This article has outlined the challenges of *in vivo* spam filtering and explained how pursuing such challenges could help data mining. It is hoped that this article stimulates interest in the problem. The appendix and references should serve as useful resources for researchers wishing to pursue it.

Acknowledgements

The opinions in this paper are those of the author and do not necessarily reflect the policies or priorities of the Hewlett-Packard Corporation.

I wish to thank Melissa McDowell for providing spam and email feeds. Thanks to Rob Holte and Chris Drummond for help on cost curves. Thanks to Julian Haight for his continuing work on SpamCop and for allowing use of figure 2; thanks to Kristian Eide for providing the data in figure 4; and thanks to Paul Wouters and the people at SpamArchive.org for making their data publicly available.

Much open source software was used in preparing this paper. I wish to thank the authors and maintainers of XEmacs, LaTeX, Grace, Perl and its many user-contributed packages, and the Free Software Foundation's GNU Project.

APPENDIX

A. SOURCES OF SPAM DATA

There are several sources of spam data on the internet, though researchers should be aware of their limitations.

1. Several static databases have been used by the machine learning community. The UCI database "spambase"[5] has a featurized version of spam and legitimate email. Androutsopoulos et al. [3] have made available several of their corpora containing both spam and personal email. All are available for download from `http://www.iit.demokritos.gr/skel/i-config/`. Note that messages in these databases are not reliably timestamped so they are not useful for measuring time-varying aspects of spam.

2. Paul Wouters, of Extended Internet, has an extensive archive of spam available at `http://spamarchive.xtdnet.nl/`. His archive covers several years. His messages from 2002 were used to produce figures 1a and 5.

3. Richard Jones of Annexia.Org has made a longitudinal spam archive available at `http://www.annexia.org/spam/index.msp`. Although his messages are carefully timestamped, note his explanation about the large drop around mid-2002, attributed to deleting a number of old mail accounts. For this reason I avoided making inferences about spam volume from his dataset.

4. `SpamArchive.org` is a "community resource used for testing, developing, and benchmarking anti-spam tools. The goal of this project is to provide a large repository of spam that can be used by researchers and tool developers." Current SpamArchive has over 200K spam messages and receives about 5000 messages per day.

5. Bruce Guenter has a longitudinal database of spam available at `http://www.em.ca/~bruceg/spam/`. See the caveat below about measuring spam volume.

Note that these datasets are archives of spam saved over time and were not designed to be controlled research datasets. It is important to understand the limitations of measuring spam volume from any of them. They are kept by owners of entire sites rather than individual accounts so the spam may be extracted from several mailboxes. The mailboxes may include `admin` and `webmaster`, which are believed to receive more spam than average. Some of these administrators even use "spam trap" addresses deliberately to attract spam. Finally, note that being active on Usenet or the web can often get a user added to spamming lists—as can making a spam archive available on the web. For all of these reasons, these spam archives may contain more spam than the average email user typically gets.

[5] `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spambase`

B. REFERENCES

[1] J. Allan, editor. *Topic Detection and Tracking: Event Based Information Retrieval*. Kluwer Academic Press, 2002.

[2] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. Spyropoulos. An evaluation of naive bayesian anti-spam filtering. In G. Potamias, V. Moustakis, and M. van Someren, editors, *Proceedings of the Workshop on Machine Learning in the New Information Age*, pages 9–17, 2000.

[3] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of SIGIR-2000*, pages 160–167. ACM, 2000.

[4] 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.

[5] CAUBE.AU. Spam volume statistics. Web page: `http://www.caube.org.au/spamstats.html`, 2002.

[6] Center for Democracy and Technology. Why am I getting all this spam? Unsolicited commercial email six month report. Available: http://www.cdt.org/speech/spam/030319spamreport.shtml, March 2003.

[7] L. F. Cranor and B. A. LaMacchia. Spam! *CACM*, 41(8):74–83, August 1998.

[8] P. Cunningham, N. Nowlan, S. J. Delany, and M. Haahr. A case-based approach to spam filtering that can track concept drift. In *The ICCBR'03 Workshop on Long-Lived CBR Systems*, Trondheim, Norway, June 2003. Available: http://www.cs.tcd.ie/publications/tech-reports/reports.03/TCD-CS-2003-16.pdf.

[9] M. Delio. Not all asian e-mail is spam. *Wired News*, Feb 19 2002. Available: http://www.wired.com/news/politics/0,1283,50455,00.html.

[10] Data Mining Cup 2003. Contest data, instructions and results available from: http://www.data-mining-cup.com/2003/Wettbewerb/1059704704/, 2003.

[11] C. Drummond and R. C. Holte. Explicitly representing expected cost: An alternative to ROC representation. In R. Ramakrishnan and S. Stolfo, editors, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 198–207. ACM Press, 2000.

[12] K. Eide. Winning the war on spam: Comparison of bayesian spam filters. Available: http://home.dataparty.no/kristian/reviews/bayesian/, August 2003.

[13] T. Fawcett and F. Provost. Fraud detection. In W. Kloesgen and J. Zytkow, editors, *Handbook of Knowledge Discovery and Data Mining*. Oxford University Press, 2002. CeDER Working Paper #IS-99-18, Stern School of Business, New York University, NY, NY 10012.

[14] J. Gleick. Tangled up in spam. *New York Times*, Feb 9 2003. Available: http://www.nytimes.com/2003/02/09/magazine/09SPAM.html.

[15] J. M. Gómez Hidalgo. Evaluating cost-sensitive unsolicited bulk email categorization. In *Proceedings of SAC-02, 17th ACM Symposium on Applied Computing*, pages 615–620, Madrid, ES, 2002.

[16] P. Graham. Better bayesian filtering. Available: http://www.paulgraham.com/better.html, Jan 2003.

[17] J. Graham-Cumming. *Field Guide to Spam*. ActiveState, 2003. Available from http://www.activestate.com/Products/PureMessage/Field_Guide_to_Spam/ and updated periodically.

[18] S. Hansell. E-mail message blitz creates what may be fastest fad ever. *New York Times*, June 2003.

[19] H. Katirai. Filtering junk e-mail: A performance comparison between genetic programming & naive bayes. Available: http://members.rogers.com/hoomank/papers/katirai99filtering.pdf, 1999.

[20] A. Kolcz and J. Alspector. SVM-based filtering of e-mail spam with content-specific misclassification costs. In *Proceedings of the TextDM'01 Workshop on Text Mining - held at the 2001 IEEE International Conference on Data Mining*, 2001.

[21] B. Krebs. Online piracy spurs high-tech arms race. *Washington Post*, June 26 2003. Available: http://www.washingtonpost.com/ac2/wp-dyn/A34439-2003Jun26.

[22] M. Lake. The great CNET spam-off. *CNET Reviews*, July 26 2001. Available: http://reviews.cnet.com/4520-3534_7-5020441-1.html.

[23] S. Machlis. Uh-oh: Spam's getting more sophisticated. *Computerworld*, Jan 17 2003.

[24] S. Olsen. Spam: It's completely out of control. *CNET News.com*, March 21 2002. Available: http://zdnet.com.com/2100-1106-865442.html.

[25] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, Mar. 2001.

[26] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453, San Francisco, CA, 1998. Morgan Kaufmann.

[27] J. Provost. Naive-bayes vs. rule-learning in classification of email. Technical Report AI-TR-99-284, University of Texas at Austin, Artificial Intelligence Lab, 1999.

[28] 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, 1998. AAAI Technical Report WS-98-05.

[29] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, and P. Stamatopoulos. A memory-based approach to anti-spam filtering. Tech Report DEMO 2001, National Centre for Scientific Research "Demokritos", 2001.

[30] K. Schneider. A comparison of event models for naive bayes anti-spam e-mail filtering. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, 2003.

[31] K. Schneider. Fighting spam in real time. In *Proceedings of the 2003 Spam Conference*, Jan 2003. Available: http://www.brightmail.com/press/2003_MIT_Spam_Conference/.

[32] W. Sturgeon. Nigerian spam scam fraudsters arrested. *Silicon.com*, Aug 20 2003. Available: http://www.silicon.com/news/500022/1/1035205.html.

[33] R. Swan and J. Allan. Extracting significant time varying features from text. In *Proc. 8th Intl. Conf. on Information Knowledge Management*, pages 38–45. ACM, 1999.

[34] L. Weinstein. Inside risks: Spam wars. *CACM*, 46(8):136, Aug 2003.

[35] D. P. Willis. Spammed. *Daily Record*, July 15 2003. `http://www.dailyrecord.com/morrislife/morrislife5-spam.htm`.