On the Relative Age of Spam and Ham Training Samples for Email Filtering

Gordon V. Cormack
University of Waterloo
Waterloo, Ontario, Canada

Jose-Marcio Martins da Cruz
Ecole des Mines de Paris
Paris, France

ABSTRACT

Email spam filters are commonly trained on a sample of spam and ham (non-spam) messages. We investigate the effect on filter performance of using samples of spam and ham messages sent months before those to be filtered. Our results show that filter performance deteriorates with the overall age of spam and ham samples, but at different rates. Spam and ham samples of different ages may be mixed to advantage, provided temporal cues are elided.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]:information filtering

General Terms: Experimentation, Measurement

Keywords: spam, email, filtering, training

1. INTRODUCTION

Spam filters are commonly trained on historical collections of messages, each labeled as spam or ham (non-spam). Their theory of operation assumes that these training messages are a random sample of those to be filtered; an assumption that is clearly not true because, when the filter is trained, the set of messages to be filtered exists only in the future. It is known that future messages are best approximated by recent messages [1]. However, acquiring and labeling recent messages may be impractical, and they may not be plentiful enough for adequate training. Sometimes, it may be more practical to acquire recent examples of one than the other; for example, spam from a spam trap or ham from the client interface. Our objective is to measure the effect on filter performance of using less-than-recent training examples, and training examples in which the ham and spam have different ages.

We collected 160,000 messages (8017 ham; 151,983 spam) addressed to one email recipient over the course of about 8 months. We split the messages by delivery date into 8 equal sets, numbered from most recent (0) to oldest (7), each representing about a month. 0 was used as the test set. 1 was used as the baseline training set, corresponding to new ham and new spam. 2 through 7 were used to measure the effect of using progressively older training examples. In addition, we used the ham from 1 and the spam from 2 through 7 to measure the effect of training on new ham and older spam, and vice versa. The results reported here use an on-line logistic regression filter with byte 4-gram features [3]; similar results were observed with other filters and with email collected from another user on a different continent.

cue	example
header date	**Mon, 4 Dec 2006 13:21:34**
daylight time	**-0400 (EDT)**
server hostname	by **mail1**.institution.net
server config.	**(8.13.1/8.13.1)**
generated ID	**01C7178F**.000D1CD0
seasonal reference	thanks for making **2006** a great year

Table 1: Temporal cues of the sort illustrated here were identified and elided using the spam filter to distinguish new messages from old, instead of spam from ham.

Figure 1 shows the filter error, expressed as $1 - AUC$ (the area above the receiver operating characteristic curve), for all combinations of training sets. Baseline error is $1 - AUC = 0.02\%$, increasing to 0.2% and 0.1% for sets 6 and 7, respectively. Substituting new ham or new spam substantially degrades performance, a result that is on the surface surprising as the average age of the training examples is decreased. Figure 2 provides further insight into this phenomenon: as progressively older ham is combined with new spam, the ham error rate explodes, while the spam error rate vanishes. The complementary effect (not shown) is observed when older spam is combined with new ham. The filter is learning to recognize new messages, not spam.

We posit that the features used to recognize new messages are contained largely in the message headers, which contain explicit timestamp information. The results obtained from removing the headers altogether, shown in figure 3, support this theory by virtue of the fact that the mixtures of new and older training messages outperform strictly older messages. But overall performance is degraded by nearly a factor of ten. Clearly the header is of critical importance to the filter and removing it is not a step toward improved effectiveness.

We therefore investigate the approach of eliding only date-specific information in the header. Eliding explicit dates alone, as shown in the first line of table 1, yields no measurable benefit. But when the other cues shown in 1 are elided, filter effectiveness on new training data is as good as the baseline and on mixed-age training data is improved

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

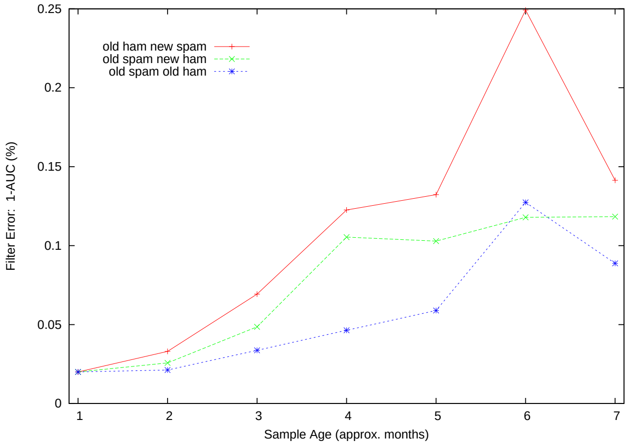

Figure 1: Effect of old and new training samples on filter error. The origin of each curve represents training on the most recent ham and spam available. The three curves represent: substituting progressively older ham, progressively older spam, and progressively older ham and spam of the same age.

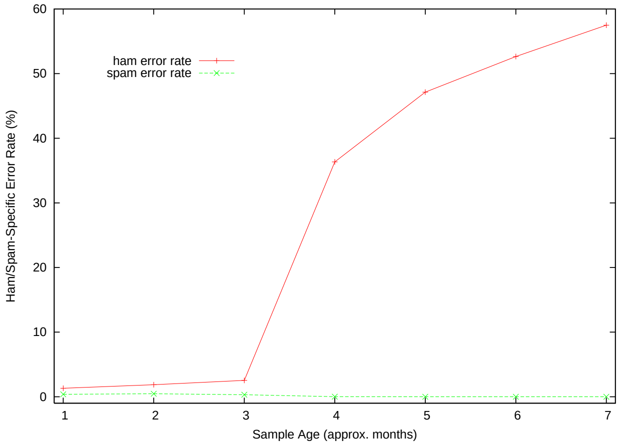

Figure 2: Separate ham and spam error rates training with progressively older ham and new spam. Spam error rate vanishes while ham error rate increases dramatically, even for 1- and 2-month-old ham.

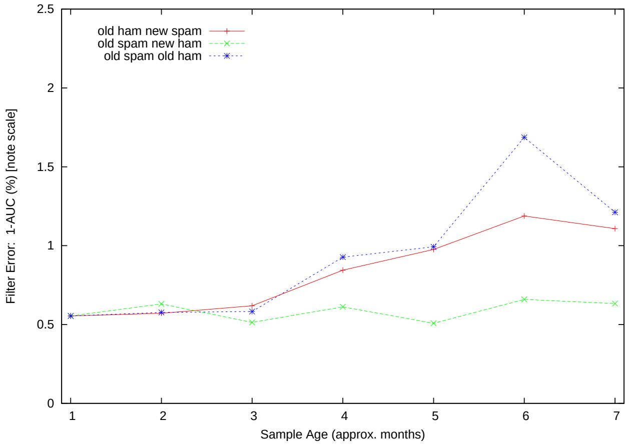

Figure 3: Effect of removing email headers. Overall error is increased tenfold but the effect of age disparity between training examples disappears.

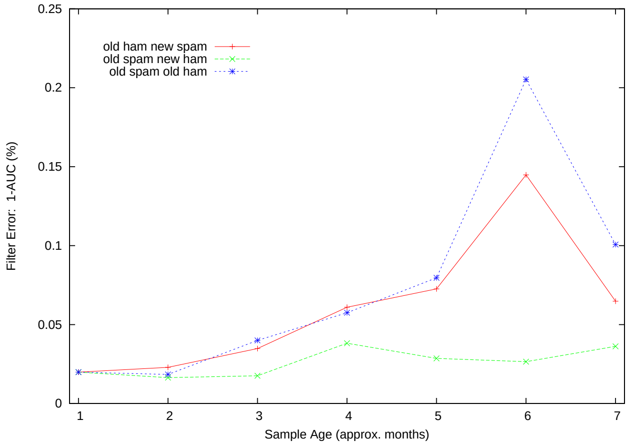

Figure 4: Effect of eliding features to mitigate temporal effects. Effectiveness on new training sample is restored to that of Figure 1 while the effect of age disparity disappears.

dramatically (figure 4). In particular, old spam and new ham works nearly as well as new spam and new ham, and much better than old spam and old ham.

The temporal cues were discovered with the aid of the spam filter itself, trained to classify messages as *new* (belonging to set 1) or *old* (belonging to set 7) rather than as spam or ham. Once the most discriminative features were identified, it was not difficult to write ad hoc scripts to eliminate them from the header. Table 1 is a complete list of the sorts of cues we found: inappropriate use of daylight saving time, server hostnames and software that were reconfigured over time, and timestamp-derived message IDs and MIME delimeters.

2. DISCUSSION

The use of old training data degrades performance, but not nearly so much as the use of raw training data in which the ham and spam have different ages. If age cues are removed, training data of mixed age may provide improved performance in the situation where only new ham or new spam is available. Header removal is too radical as it dramatically compromises overall performance. If a few tell-tale temporal cues are identified and elided, substituting newer training data for one class of messages appears to yield improved effectiveness over using old for both.

Our approach to identifying the training cues was not entirely automatic, and not entirely blind to the training data (but definitely blind to the test data). We believe it is a good candidate to be automated. And even if effected manually, it is much more efficient than labeling a new training set. The cues we discovered closely match those mentioned by the authors of the TREC 2005 Spam Corpus [2].

References

[1] Cormack, G. V., and Bratko, A. Batch and on-line spam filter evaluation. In *CEAS 2006*.

[2] Cormack, G. V., and Lynam, T. R. Spam corpus creation for TREC. In *CEAS 2005*.

[3] Goodman, J., and tau Yih, W. Online discriminative spam filter training. In *CEAS 2006*.