Spam Corpus Creation for TREC

Gordon Cormack
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Thomas Lynam
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

TREC's *Spam Filtering Track* (Cormack & Lynam, 2005) introduces a standard testing framework that is designed to model a spam filter's usage as closely as possible, to measure quantities that reflect the filter's effectiveness for its intended purpose, and to yield repeatable (i.e. controlled and statistically valid) results. The *TREC Spam Filter Evaluation Toolkit* is free software that, given a corpus and a filter, automatically runs the filter on each message in the corpus, compares the result to the *gold standard* for the corpus, and reports effectiveness measures with 95% confidence limits. The corpus consists of a chronological sequence of email messages, and a gold standard judgement for each message. We are concerned here with the creation of appropriate corpora for use with the toolkit.

It is a simple matter to capture all the email delivered to a recipient or a set of recipients. Using this captured email in a public corpus, as for the other TREC tasks, is not so simple. Few individuals are willing to publish their email, because doing so would compromise their privacy and the privacy of their correspondents. So we are left with the choice between using an artificial public collection of messages and using a more realistic collection that must be kept private.

Artificial collections (spamassassin.org, 2003; Androutsopoulos et al., 2000; Michelakis et al., 2004) may be created by using mailing list messages as opposed to personal email, by selecting non-sensitive messages from a real email collection, by mixing messages from diverse sources, or by obfuscating genuine messages[1]. All of these approaches conflict with our design criteria – that real filter usage be modelled as closely as possible – and may compromise the very information that filters use to discriminate ham from spam, either by removing pertinent details or by introducing extraneous information that may aid or hinder the filter.

[1]The majority of filters we have evaluated exhibit pathologies on the PU obfuscated corpora.

We define spam to be *"Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."* The gold standard represents, as accurately as is practicable, the result of applying this definition to each message in the collection. The gold standard plays two distinct roles in the testing framework. One role is as a basis for evaluation. The gold standard is assumed to be *truth* and the filter is deemed correct when it agrees with the gold standard. The second role is as a source of user feedback. The toolkit communicates the gold standard to the filter for each message after the filter has been run on that message.

Human adjudication is a necessary component of gold standard creation. Exhaustive adjudication is tedious and error-prone; therefore we use a bootstrap method to improve both efficiency and accuracy. The bootstrap method begins with an initial gold standard G_0. One or more filters is run, using the toolkit and G_0 for feedback. The evaluation component reports all messages for which the filter and G_0 disagree. Each such message is re-adjudicated by the human and, where G_0 is found to be wrong, it is corrected. The result of all corrections is a new standard G_1. This process is repeated, using different filters, to form G_2, and so on, to G_n.

One way to construct G_0 is to have the recipient, in the ordinary course of reading his or her email, flag spam; unflagged email would be assumed to be ham. Or the recipient could use a spam filter and flag the spam filter's errors; unflagged messages would be assumed to be correctly classified by the filter. Where it is not possible to capture judgements in real time – as for all public collections to which we have access – it is necessary to construct G_0 without help from the recipient. This can be done by training a filter on a subset of the messages (or by using a filter that requires no training) and running the filter with no feedback.

Experience

We have employed this technique on a collection of 49086 private email messages. G_0 was captured from the recipient's feedback to a spam filter. Figure 1 illustrates the five revision steps forming G_1through G_5, the final gold standard. $S \rightarrow H$ is the number of message classifications revised from spam to ham; $H \rightarrow S$ is the opposite. Note that G_0 had 421 spam messages incorrectly classified as ham. Left uncorrrected, these errors would cause the evaluation kit to overreport the false positive rate of the filters by this a mount – a factor of seventy for the best filters and a factor of 2.4 for the worst. In other words, the results captured from user feedback alone – G_0– are not accurate enough to form a useful gold standard.

	$S \rightarrow H$	$H \rightarrow S$
$G_0 \rightarrow G_1$	0	278
$G_1 \rightarrow G_2$	4	83
$G_2 \rightarrow G_3$	0	56
$G_3 \rightarrow G_4$	10	15
$G_4 \rightarrow G_5$	0	0
$G_0 \rightarrow G_5$	8	421
G_5	$\|H\| = 9038$	$\|S\| = 40048$

Figure 1: Bootstrap Gold Standard Iterations

We have constructed a preliminary gold standard for the Enron Corpus (Klimt & Yang, 2004). As distributed by Carnegie Mellon University, corpus consists of 520,000 files from the email folders of 150 recipients, 216,209 of which are unique. About 2.5% of the messages are spam, with the proportion of spam varying dramatically between recipients – from none to about 30%. G_0 was produced using SpamAssassin 2.63 with its learning component disabled. G_1 was produced using the SpamAssassin with the learning component enabled. G_2 was produced with Bogofilter. In adjudicating these runs, we became aware of a number of technical problems. A very large number of the files were empty, or did not appear to be email messages. Attachments were elided. No original headers were present. We found it very difficult to adjudicate many messages because it was difficult to glean the relationship between the sender and the receiver. In particular, we found a preponderance of sports betting pool announcements, stock market tips, and religious bulk mail that was adjudicated as spam but in hindsight we suspect was not. We found advertising from vendors whose relationship with the recipient we found tenuous. In our adjudication for G_3 we separated the messages by user, which appears to make adjudication easier. G_4 and G_5 involved evaluating 15619 messages from 18 users, of which about 2000 were spam.

During this process, we identified the need to view the messages by sender; for example, once the adjudicator decides that a particular sports pool is indeed by subscription, it would be more efficient and probably more accurate to adjudicate all messages from the same sender at one time. Similarly, in determining whether or not a particular "newsletter" is spam, it is desirable to be able to identify all of its recipients. This observation occasioned us to design a new tool for adjudication – one that would allow us to use full-text retrieval to look for evidence and to ensure consistent judgements. At the same time, we determined that the preponderance of non-email files and lack of headers rendered the CMU version of the corpus unsuitable for the TREC spam track. We chose instead to retrieve the Enron email directly from FERC (FERC, 2003).

The FERC database contains 1.3 million files, 100,652 of which contain original email headers. We created G_6 by running a filter, trained on G_5 on these messages with no feedback. G_7 was created by in the normal way, which G_8 used a preliminary version of our tool to identify senders and domains whose messages had been inconsistently adjudicated. This last iteration reclassified 632 (of 6196) spam to ham and 383 ham to spam.

Construction of a gold standard for the Enron Corpus, and the tools to facilitate that construction, remains a work in progress. We believe that, in spite of the fact that the messages have lost much of their original formatting, and notwithstanding our adjudication problems, the Enron Corpus will form the basis of a larger, more representative public spam corpus than currently exists.

References

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., & Stamatopoulos, P. (2000). Learning to filter spam e-mail. *Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases.*

Cormack, G., & Lynam, T. (2005). Spam track preliminary guidelines - TREC 2005. http://http://plg.uwaterloo.ca/ gvcormac/spam/.

FERC (2003). Information released in enron investigation. http://fercic.aspensys.com/members/manager.asp.

Klimt, B., & Yang, Y. (2004). Introducing the enron corpus. *First Conference on Email and Anti-Spam (CEAS).*

Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G., & Stamatopoulos, P. (2004). Filtron: A learning-based anti-spam filter. *Proceedings of the First Conference on Email and Anti-Spam (CEAS).*

spamassassin.org (2003). The spamassassin public mail corpus. http://spamassassin.apache.org/publiccorpus.