A Mail Client Plugin for Privacy-Preserving Spam Filter Evaluation

Mona Mojdeh
Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada N2L 3G1

Gordon V. Cormack
Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada, N2L 3G1

Abstract

We describe a plugin extension to the Thunderbird Mail Client to support standardized evaluation of multiple spam filters on private mail streams. Researchers need not view or handle the subject users' messages and subject users need not be familiar with spam filter evaluation methodology. All that is required of the user is to install the plugin as a standard extension and to run it on his or her mailbox. The plugin evaluates a spam filter, assuming the user's existing classification to be accurate, and sends summary results only to the researcher, after allowing the user to verify exactly what is sent. This plugin addresses an outstanding challenge in spam filter evaluation: that of using a broad base of realistic data while satisfying personal and legislative privacy requirements. Previous efforts have used public data which may not be representative, captured data which may be insufficiently private, and obfuscation techniques which compromise the integrity of the data and may also be insufficiently private. We show preliminary results using the tool to evaluate some filters previously evaluated at TREC.

1 Introduction

Preserving the privacy of personal and corporate data has been the subject of increasing attention in both the public and private sectors. These privacy considerations are often at odds with the conduct of large-scale realistic spam filtering efforts. To date, quantitative spam filter evaluation has been conducted using only data sets that may not be representative of the email for which spam filters are actually deployed. Ling Spam [1], for example, is a synthetic corpus consisting of mailing list messages; much information is removed from the messages. The PU corpora [2] are derived from real email, but obfuscated in a manner that strongly compromises filter performance. Furthermore, in the related field of information retrieval, obfuscation such as that applied to the PU corpora, has been found not adequately to preserve privacy [3] [4]. The SpamAssassin Corpus [5], consists of donated messages from heterogeneous sources. It is not easy to acquire a large corpus of such messages, and in any event they are unrepresentative by virtue of the fact that they are selectively donated. TREC [6] – perhaps the most realistic comparative evaluation to date – uses both public and private corpora. The public corpora are derived from public sources and hand-crafted to approximate realistic data [7]. The private corpora consist of data acquired from actual email users; these users allowed the TREC researchers to archive their email for the purpose of evaluation. These private datasets would more correctly be called semi-private: although TREC participants had no direct access to the email, the researchers running the evaluation did.

To achieve access to a wider range of representative data we believe it is necessary to use the email of subject users without allowing researchers to read or otherwise deduce its content. We describe WATEF, a Mozilla extension that a volunteer subject can easily install and run to test a filter on his or her email. The results of this run are exactly those defined by the TREC interface: simply a text file with one line per message indicating the true (gold standard) classification, the filter's classification, and the filter's *spamminess score*. The subject is given an opportunity to review the file and to approve (or disapprove) of its being transmitted to the researcher.

The tool supports the following experimental design. The researcher would solicit subjects from some user population. Filters implementing the TREC interface would be encapsulated as WATEF extensions, and the

users would be instructed to install and run one or more, resulting in the summary files being returned to the researcher by email. These summary files would then be evaluated using the TREC spam filter evaluation toolkit [6].

2 WATEF

WATEF is an Extension for Mozilla Thunderbird. Thunderbird is a free, open source, cross-platform e-mail and news client developed by the Mozilla Foundation [8]. Extension, also known as add-on, is a great way to extend the functionality of Mozilla Thunderbird by enhancing Mozilla's Foundation's projects [9]. Extensions are easily installed and allow users to modify and personalize Mozilla's environment.

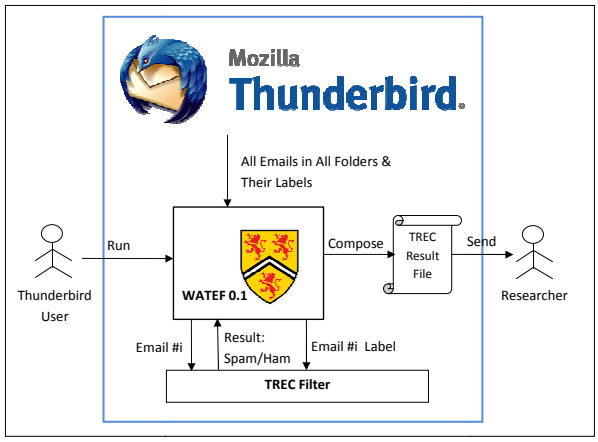

Figure 1: WATEF Design

WATEF extension works as the wrapper of any spam filter written in TREC Spam Track format. TREC's Spam Track uses a standard testing framework that presents a set of chronologically ordered email messages to a spam filter for classification. In the filtering task, the messages are presented one at a time to the filter, which yields a binary judgement (spam or ham). The filter also yields a *spamminess* score, intended to reflect the likelihood that the classified messages is spam [10]. TREC Spam Track has modeled four different user feedbacks(*Immediate*, *Delayed*, *Partial*, and *Active*). For the purpose of this paper we take the *Immediate* feedback.

The design of WATEF is shown in figure 1. On one side, WATEF can include any TREC formatted filter and on the other, it accesses all the folders and emails of Thunderbird through pre-defined components in Mozilla. WATEF also gets the judgment of Thunderbird on the emails; this judgment is basically how Thunderbird has labeled the email at the time of running the extension. Thunderbird assigns a *junk score* to each email, which is usually the score the built-in spam filter has assigned. This score might also be the result of the user explicitly marking a misclassified message as spam or ham, in order to train the filter. We treat these judgements as the ground truth, with the idea that the user almost always corrects misclassifications of Thunderbird.

When run, WATEF first compiles and initializes the spam filter. It, then, accesses emails in Thunderbird's different folders and passes them one by one, in a chronological order, to the TREC filter, and gets the result of the filter, indicating whether the email is classified as spam or ham. When all the emails are evaluated by the TREC filter, WATEF collects all the results, and makes a TREC formatted result file. Then an email is composed with this result file as the attachment. This email is sent to a specific address which can be modified in the preference option of WATEF extension. The default value for this address is the author's email address. The Thunderbird user can easily open the result file before sending the email to make sure no private data is sent out. This result file has, for each email, the spamminess score, the filter's classification, and Thunderbird's label, without indicating any sensitive date about the email itself. The extension also contains the XML and Java source codes, so the user can be certain about what the program does.

Figure 2 shows how the extension looks like after being installed, and the option box after selecting preferences of WATEF extension from the extensions list.

3 Results

In order to evaluate the extension, we used three different filters; Logistic Regression (LR) [11], Dynamic Markov Compression (DMC) [12], LR and DMC Fusion [13]. For each filter to be applied, we simply made a new version of WATEF xpi file by placing the filter in the extension, and then upgrading the extension in Thunderbird. This way we got the three result file.

The result files were tested using the evaluation script in the TREC Spam Kit. The measure used is the Receiver Operating Characteristic (ROC) Curves, and (1-ROCA)(%) as the area above ROC curve, indicating the probability that a random spam message will receive a lower spamminess score than a random ham message [10]. Table 1 shows the (1-ROCA)% of the three filters applied on the current author's set of emails in Thunderbird. Number of emails at the time of running the filters is 4800. The table also shows the results of the filters on the TREC'07 Public Corpus as published in [10].

As can be seen in table 1, the three filters have not

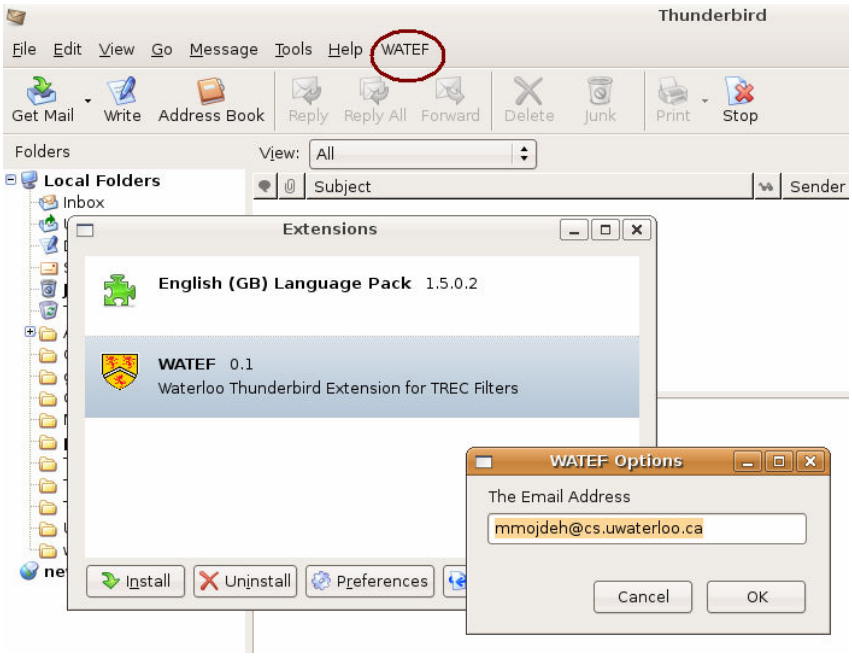

Figure 2: WATEF Extension

performed to their best when run within Thunderbird. Recall that the lower the (1-ROCA)(%), the better the filter's performance. The first explanation for this problem is that we are taking the judgements of Thunderbird on each email as the ground truth of the email's classification, while this is not always true. Some users never train the filter or they miss some emails which are misclassified by Thunderbird's built-in spam filter. This is specially the case for false positives, as users usually don't check their junk folder for legitimate emails. [cite cormack and lynam 2005 ceas]

To show that the relatively low (1-ROCA)(%) in table 1 are due to mistakes in the ground truth judgements, we went through all the emails with different filter classification and Thunderbird judgements, and corrected the misjudgements. Table 2 shows the results of filters applied on the new modified email judgements. For this table we have also tested Relaxed Online SVM (ROSVM) filter implemented in [14].

4 Future Work

As described in the introduction section, WATEF extension provides a mean to evaluate spam filters in a more real environment. In order to make WATEF publicly usable, some features need to be added to the current version. First, users must be convinced that WATEF is easy and fast (or not time consuming). Therefore, we have to make sure WATEF runs in the background, letting user run his own processes. Also, to preserve users's privacy, WATEF must be signed.

One feature that can be included to motivate users, is double checking with the user the judgement of the emails that have big gap between their Thunderbird assigned *junk score* and the filter's *spamminess* scores. In this way, user can identify false positives and negatives, and also the researcher will be sure about the judgements sent out in the composed result file.

Table 1: Results: 1-ROCA(%)

Method	**Thunderbird**	**TREC07 Pub.**
LR	2.4368	0.0057
DMC	3.6107	0.0077
LR & DMC Fusion	2.5877	0.0055

Table 2: Results After Correcting Thunderbird Misjudgements

Method	**1-AUC(%)**
LR	0.2079
DMC	0.3207
LR & DMC Fusion	0.1222
ROSVM	0.5309

5 Conclusion

This paper introduces a new approach to evaluate spam filters in a more genuine settings. The main objective is to provide a reliable and convenient way to assess filters written in TREC format on a real user mail box, without asking users to donate their emails.

WATEF is an extension (or add-on) for Mozilla Thunderbird email client. It includes some TREC formatted spam filter which is run on the user's client machine, and composes a file for TREC Spam Kit. This file contains the result of the filter applied on the emails, together with the judgements Thunderbird has assigned to the them. So far, WATEF has been used to evaluate four different filters, for which the results are described in the paper.

References

[1] Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., and Spyropoulos, C., An evaluation of naive bayesian anti-spam filtering, In Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML-2000), Spain.

[2] Androutsopoulos, I., Paliouras, G., Michelakis, E., Learning to Filter Unsolicited Commercial E-Mail, Technical report 2004/2, NCSR "Demokritos".

[3] http://www.securitypronews.com/insiderreports/insider/spn-49-20060807AOLPublishesWithdrawsUserSearchData.html

[4] http://www.securityfocus.com/news/11497

[5] SpamAssassin. 2004. The spamassassin public mail corpus. http://spamassassin.apache.org/publiccorpus.

[6] http://plg.uwaterloo.ca/ gvcormac/spam/

[7] Cormack G. V., Lynam, T. R., Spam Corpus Creation for TREC, In The Second Conference on Email and Anti-Spam (CEAS-2005), Stanford University, CA, USA, 2005.

[8] https://addons.mozilla.org/en-US/thunderbird/

[9] http://www.mozilla.org/projects/thunderbird/specs/extensions.html

[10] Cormack, G. V., TREC 2007 Spam Track Overview, In Sixteenth Text REtrieval Conference (TREC-2007), Gaithersburg, MD, 2007, NIST.

[11] Goodman, J., and tau Yih, W., Online discriminative spam filter training, In The Third Conference on Email and Anti-Spam (CEAS-2006), Mountain View, CA, USA, 2006.

[12] Bratko, A., Cormack, G. V., Filipic, B., Lynam, T. R., and Zupan, B., Spam filtering using statistical data compression, Journal of Machine Learning Research, 6: 2673-2698, 2006

[13] Cormack, G. V., University of Waterloo Participation in the TREC 2007 Spam Track, In Sixteenth Text REtrieval Conference (TREC-2007), Gaithersburg, MD, 2007, NIST.

[14] Sculley, D., and Wachman, G. M., Relaxed Online SVMs in the TREC Spam Filtering Track, In Sixteenth Text REtrieval Conference (TREC-2007), Gaithersburg, MD, 2007, NIST.