Combining Global and Personal Anti-Spam Filtering

Richard Segal
IBM Research
Hawthorne, NY 10532

Abstract

Many of the first successful applications of statistical learning to anti-spam filtering were personalized classifiers that were trained on an individual user's spam and ham e-mail. Proponents of personalized filters argue that statistical text learning is effective because it can identify the unique aspects of each individual's e-mail. On the other hand, a single classifier learned for a large population of users can leverage the data provided by each individual user across hundreds or even thousands of users. This paper investigates the trade-off between globally- and personally-trained anti-spam classifiers. We find that globally-trained text classification easily outperforms personally-trained classification under realistic settings. This result does not imply that personalization is not valuable. We show that the two techniques can be combined to produce a modest improvement in overall performance.

1 Introduction

Statistical text classification is at the core of many commercial and open-source anti-spam solutions. Statistical classifiers can either be trained globally with one classifier learned for all users, or personally where a separate classifier is learned for each user. Personally-trained classifiers have the advantage of allowing each user to provide their own personal definition of spam. A user actively refinancing their home can train a personal filter to delete unsolicited stock advice as spam but deliver unsolicited refinancing offers to their Inbox. Another user might train a personal filter to block all unsolicited offers. Personal classifiers can quickly identify terms that are unique to an individual and use them as strong indicators of ham. Phrases such as "pastry-powered," "swiftfile," and "spamguru" appear frequently in the author's ham e-mail but probably have never appeared in any spam, let alone spam directed towards the author's Inbox.

A globally-trained anti-spam filter is trained on spam and ham e-mail from a large collection of users. There are several types of globally-trained anti-spam solutions commonly in use. Anti-Spam services such as BrightMail and IBM E-mail Security use a wide variety of data sources (including spam-trap data, ham-trap data, and customer-labeled data) to create a globally-learned anti-spam model that is shared by all customers. The primary advantage of globally-trained anti-spam solutions is its ability to leverage data from multiple individuals. In a personal system, every training message must be provided by the end user. If a new spam message bypasses several users personally-trained filter, each user must separately update their classifier. In a globally-trained system, only a few users need to label a message for the classifier to be updated for everyone.

Globally-trained systems are much easier to manage and deploy in large organizations. They often take the form of a single SMTP server that can be placed in an organizations SMTP chain and used to filter e-mail for hundreds if not thousands of users. Managing a single server in a central location is vastly easier and less costly than managing and supporting one thousand personal classifiers deployed on the desktop.

It is often assumed that the convenience of globally-trained solutions comes at a cost. Globally-trained solutions cannot take into account individual definitions of spam, cannot easily profile the terms used to each user, and may be easier to attack as a spammer only needs to defeat a single system to reach a broad range of users.

This paper challenges this assumption by empirically comparing personal and global anti-spam filters. We find that globally-trained classifiers substantially outperform personally-trained classifiers for both small and large user communities. We demonstrate that one of the reasons for the success of globally-trained classifiers is that not all personal data gets averaged out when training across a large community. In fact, much of the personal data is retained allowing the globally-trained classifier to benefit from individual preferences.

However, some personal data is indeed lost by a globally-trained classifier. Globally-trained classifiers cannot retain personal preferences when users differ on whether a feature should be treated as an indicator of spam or ham. This loss of personal data can hurt the performance of globally-trained classifiers on diverse datasets. We propose a new algorithm for combining personal and globally-trained data that makes use of both global and personal information to classify each incoming message. We show that this new algorithm offers a substantial improvement in classifier accuracy when the user-base is sufficiently diverse.

In the next section, we describe the Naïve Bayes text classifier that we use as a basis for our research and detail the globally-trained and personally-trained variants used in the experiments to follow. The second section describes *Dynamic Personalization*, our algorithm for combining personally- and globally-trained text classifiers. We then present the results of our empirical evaluation, discuss related work, and then conclude.

2 Naïve Bayes Spam Filtering

We compare personally-trained and globally-trained anti-spam filtering by analyzing their performance in the specific case of a Naïve Bayes text classifier (Lewis, 1998). We believe most of the results and observations in this paper apply equally well to other bag-of-word text classifiers (e.g. linear regression, SVM), but we leave proving this hypothesis to future work.

Let D denote a document containing words, $w_1 \dots w_n$. Let S denote the event that document D is spam. And, let θ denote the set of labeled training data. A Naïve Bayesian text classifier computes the probability a document D is spam using the formula:

$$P(S|D,\theta) = \alpha P(S|\theta) \prod_{w_i \in D} P(w_i|S,\theta) \qquad (1)$$

where α is a normalization constant chosen to ensure that $P(S|D,\theta) + P(\neg S|D,\theta) = 1$.

We estimate $P(w_i|S,\theta)$ using a variation of Laplacean smoothing that has terms added to reduce the strength of the smoothing for long words.

The difference between globally- and personally-trained Naïve Bayes is the training set θ used to build the classifier. Let θ_u denote a training set consisting of only spam and ham samples from user u. Using this notation, $P(S|D,\theta_u)$ denotes the personally-trained classifier for user u.

A globally-trained classifier is trained on all labeled messages. Let U denote the set of all users. Let θ_* denote the training set consisting of all labeled data regardless of recipient. Then, the globally-trained classifier $P(S|D,\theta_*)$ is the classifier that results from training on θ_*. For the purposes of this comparison, we make the simplifying assumption that all training data comes from individual users. That is,

$$\theta_* = \bigcup_{u \in U} \theta_u. \qquad (2)$$

This assumption accurately models systems where all labeled data originates from user-provided spam and ham samples. It does not accurately characterize labeled data from exogenous sources such as spam traps or the auto-voting of e-mail destined for invalid recipients. We believe this assumption is reasonable as it limits globally-trained classifiers to the same information that is usually available to personally-trained classifiers.

We will compare personally- and globally-trained classifiers on large corpora that contain data from multiple users. This methodology accurately models domain-level anti-spam solutions that can employ either globally-trained or personally-trained classifiers. Let $R : D \rightarrow 2^U$ denote the mapping from messages to message recipients. Let $P = (D,u)|u \in R(D)$ be the set of all message-recipient pairs. The personally-trained Naïve-Bayes classifier *NBP* labels each incoming message-recipient pair using the model learned for the specified recipient:

$$NBP(D,u,\tau) = [P(S|D,\theta_u) \geq \tau].$$

where τ is a threshold used to convert the classifier's probability model into a decision function. Note, *NBP* applies the same classification formula for user u that user u would expect if he or she was using a personally-trained classifier in isolation. The performance of *NBP* is the same as would result from having each user install his or her own personally-trained filter.

The globally-trained Naïve Bayes classifier *NBG* labels each incoming message-recipient pair based on the global model, ignoring the message's intended recipient:

$$NBG(D,u,\tau) = [P(S|D,\theta_*) \geq \tau].$$

We define classification in terms of message-recipient pairs to give each user the chance to benefit from their own personal filter. This introduces the question of how to evaluate the accuracy of the classifier in terms of message-recipient pairs. One option would be to treat each message-recipient pair as a separate event. This would imply that incorrectly classifying a single message for two different users would be treated as two distinct errors. However, this option is not ideal for a comparative study as it tends to produce highly correlated errors. Instead, we divide the weight of each message equally among its recipients. If a message is classified incorrectly for all the recipients in a message, we count it as one full error. If is classified incorrectly for half the recipients, then it is counted as half an error.

3 Dynamic Personalization

The distinction between globally-trained and personally-trained classifiers does not need to be absolute. Varying degrees of personalization can be added to globally-trained systems or vice versa. Personal classification is probably ideal if there is sufficient data to make a judgment. In contrast, global data is best when good personal data is not available.

One method for combining globally-trained and personally-trained classifiers would be to treat them as separate classifiers and combine them using standard classifier-aggregation techniques (Segal, 2005; Dietterich, 2000; Larkey & Croft, 1996). However, this method takes an all-or-nothing approach. Users with insufficient data to train a good personal classifier will have their personalization data ignored as the aggregate system would rely solely on the global classifier. We seek a method that allows even small amounts of personalization data to have an impact when appropriate.

Instead, we propose that the trade-off between global and personal data be made at the word or feature level. The basic idea is to apply personal data whenever it provides a better estimate for $P(w_i|S)$ (or $P(w_i|\neg S)$), and to use global data otherwise. We accomplish this trade-off using the following equation:

$$P(w_i|S,\theta_*,\theta_u) \approx \frac{F(w_i,S,\theta_u) + P(w_i|S,\theta_*)K}{F(S,\theta_u)+K} \tag{3}$$

where $F(w_i,S,\theta_u)$ denotes the number of spam documents in θ_u that contain word w_i, and $F(S,\theta_u)$ denotes the total number of spam documents in θ_u. This equation works by applying a Beta prior with a mean of $P(w_i|S,\theta_*)$ to the estimate of $P(w_i|S,\theta_u)$. Using this equation, as the amount of personalization data $F(S,\theta_u)$ grows, the value of $P(w_i|S,\theta_*,\theta_u)$ asymptotically approaches $P(w_i|S,\theta_u)$. When the amount of personalization data is small, the same equation asymptotes to $P(w_i|S,\theta_*)$. Therefore, equation 3 provides a linear transition from a globally-trained classifier to a personally-trained classifier as the amount of personalization data increases. The parameter K determines the relative significance given to personalization data. We use the value of $K = 100$ for our experiments. We define Naïve-Bayes with dynamic personalization *NBDP* as the classifier that uses equation 3 to classify e-mail:

$$NBDP(D,u,\tau) = [P(S|D,\theta_*,\theta_u) \geq \tau]$$

4 Experimental Setup

We use three datasets for our analysis. The first two are the TREC 2005 and TREC 2006 public corpora (Cormack & Lynam, 2006; Cormack, 2007). The third data set is the private SpamGuru 2004 corpus (Segal et al., 2004).

The evaluation of personally-trained classifiers requires that we can identify the intended recipients of each message. As the original SMTP delivery information is not available in any of the corpora, the recipients for each message were extracted from the "To:" and "CC:" headers. Recipients for e-mail domains outside of the domains served by each dataset were filtered out. We also removed any e-mail address that did not receive at least ten spam and ten ham messages to help ensure that only valid e-mail addresses were selected.

Many messages in each corpora either cannot be attributed to an appropriate recipient or are destined for users with less than ten ham or less than ten spam messages. We remove these message from each corpora to ensure that all messages in our test data can be attributed to at least one selected e-mail address. This is important for comparative purposes as personally-trained classifiers cannot classify messages for which no recipient can be identified. Table 1 describes our test corpora, including the number of valid e-mail addresses extracted from each data set.

Our testing methodology is based on the TREC 2005 methodology (Cormack & Lynam, 2006), but adapted for personally-trained classification. We train and test each classifier in an incremental fashion. The dataset is processed in chronological order. Each message is presented to the classifier to be labeled. Once the label is returned, the classifier is given the label of the current message for training. Unlike the original TREC 2005 setup, the classifier is asked to label the message

	SpamGuru 2004	TREC 2005	TREC 2006
Original Spam	130,455	52,790	24,912
Original Ham	42,557	39,399	12,910
Selected Recipients	240	103	25
Selected Spam	120,368	40,805	16,499
Selected Ham	33,886	21,549	9,121

Table 1: Test corpora statistics.

separately for each of its intended recipients. The correct labels for each recipient is separately passed to the classifier.

In personally-trained anti-spam filters, spam is usually defined operationally based on user-supplied spam and ham samples. Each recipient can assign its own label to each message. A message labeled spam by one recipient can be labeled ham by another. Ideally, each message in our test corpora would be assigned a separate label for each user. However, there are no user-specific labels available for any of our test corpora. As we are not aware of any public corpora with judgments stored on a per user basis, we make the unrealistic assumption that each user assigns the same label to each message and assign each user the same judgment, the message's correct label as indicated by the test corpora. As a result of this assumption, the experiments below may underestimate the value of personalized classification. However, we believe the results below are still valid as the large performance differences reported are unlikely to be reversed by the small number of messages that would be labeled differently by individual users.

We perform several evaluations with only a subset of all recipients. For these evaluations, we only train and test on those message-recipient pairs that contain the target recipients.

5 Empirical Results

Figure 1 compares the performance of globally-trained and personally-trained anti-spam filters on each of our test corpora. For this experiment, we learned separate personal classifiers for each of the recipients in the database that had received ten on more spam and ten or more ham examples. Messages that could not be attributed to recipients with enough examples were excluded from the experiment. The results for the personally-trained classifier were computed by classifying each message-recipient pair with the personally-trained classifier for the pair's recipient. We also evaluated a single, global classifier that was trained on all messages and used to classify every message-recipient pair.

The results reveal the limitations of personally-trained classifiers. The personally-trained classifiers performed very poorly, providing at best twice the false-negative rate of a globally-trained system at a false positive rate of 1%, and missing as much as 10 times as much spam at a false positive rate of 0.1%. The reasons for this failure in retrospect are obvious. Many of the recipients in these experiments have fewer than 100 training examples. For instance, in the TREC 2005 corpus, 7.6% of all ham comes from users with 100 or fewer ham examples. The personally-trained classifiers for these users are unlikely to achieve good performance at a 0.1% false positives as the data is just not there. The poor performance on users with insufficient data makes it virtually impossible for personally-trained classifiers to perform well in this experiment.

It is unclear whether personally-trained classifiers should be held accountable for users that cannot provide sufficient training data. Personally-trained classifiers can certainly be effective for users with sufficient data. But, there will always be users which either do not have sufficient messages or time to train the system appropriately. For these users, a globally-trained classifier is likely to be a better alternative.

We can get a better idea of the potential of personally-trained classifiers by limiting our evaluation to users with sufficient data. Figure 2 shows the same comparison applied to the four highest ranked users from each corpora, where users are ranked based on the number of spam messages they receive or the number of ham messages, whichever is *smaller*. We use this ranking as the data set contains several virtual spam traps and ham traps that receive a large number of one type of message, but almost none of the other. All the recipients in this experiment have at least 500 spam and 500 ham examples available for training. It is unlikely that all but the most active of e-mail users with have much more than this easily available to train a personal classifier.

The results of this second experiment are surprising. The personally-trained classifier performs reasonably well, catching 99.7% of spam on the TREC 2006 database with at a false positive rate of 0.1%. For as good as this is, the globally-trained classifier is even better, catching close to 99.85% at the same false positive rate. Overall, the globally-trained classifier let through about half as many spam messages as did the personally-trained classifier across most of each ROC curve.

These results suggest there is little benefit to be gained from personalized classification. But even the results for the complete corpora are for a relatively small set of users. The results cannot be directly applied to large

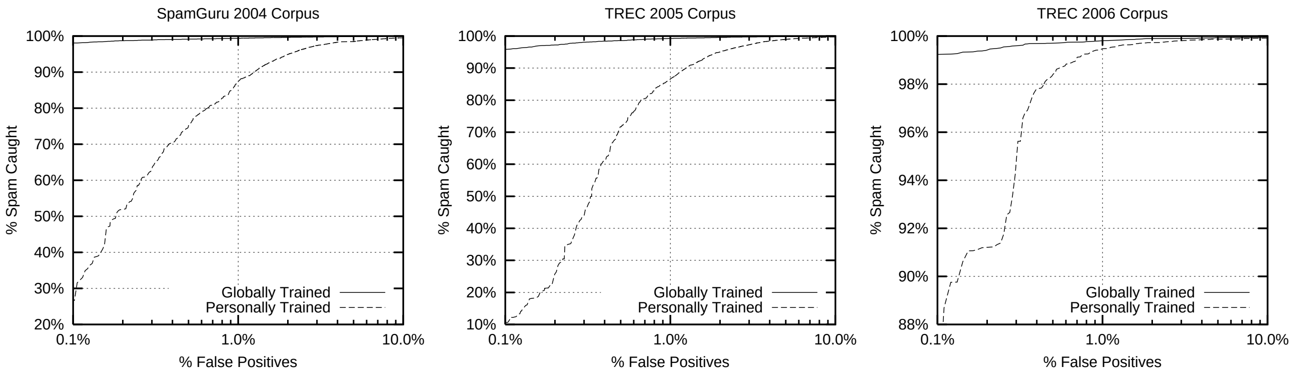

Figure 1: Comparison of globally- and personally-trained classifiers.

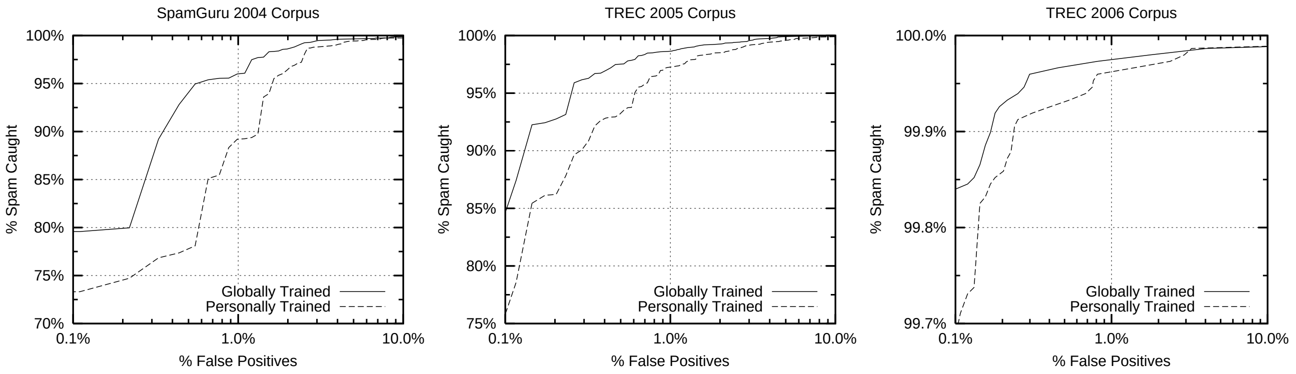

Figure 2: Comparison of globally- and personally-trained classifiers for the four largest users in each dataset.

organization and ISP's where the user base can include thousands or even hundreds of thousands. This is important as there is still a legitimate concern that aggregating across too many users may degrade the effectiveness of globally-trained classifiers.

Figure 3 shows how the performance of globally-trained Naïve Bayes scales as we increase the number of users. Rather than show the complete ROC curve for each data point, we summarize the results of each experiment using one minus the area under the ROC curve (Fawcett, 2003). The results for the SpamGuru 2004 corpus, which is the larger and possible more realistic corpus, show globally-trained classification only improving with the number of included recipients. The results on the TREC 2006 corpus is similar, showing a strict improvement in overall performance as the number of included recipients increase. The results on these two corpora are suggestive of globally-trained classifiers scaling well to hundreds if not thousands of users. However, the results on the TREC 2005 corpus show performance only improving for the first 75 users, and shows performance decreasing as the number of users increase from 75 to 100. The reason for this drop in performance is unclear and warrants further study.

Figure 4 compares the performance of dynamic personalization to a strictly globally-trained classifier. The results are mixed. Dynamic personalization offers a modest improvement on the SpamGuru 2004 corpus for false positive rates above 0.1% , a modest improvement on the TREC 2006 corpus for false positive rates above 0.4%, and offers almost identical performance to a globally-trained classifier on the TREC 2005 corpus. On all three datasets the performance of dynamic personalization drops below the globally-trained classifier for the smallest false-positive rates we measure.

The modest gains shown for dynamic personalization on two of the three corpora at a false positive rate of 1% is suggestive that it may offer real advantages on very large datasets. But, the consistent drop in performance at low false positive rates is definitely of concern.

Further analysis suggests that Dynamic Personalization did not offer a large benefit on these corpora partially because of the ability of globally-trained classifiers to retain a large amount of personalization information. In a globally-trained classifier, the statistics for each term is based on a sum of the statistics for

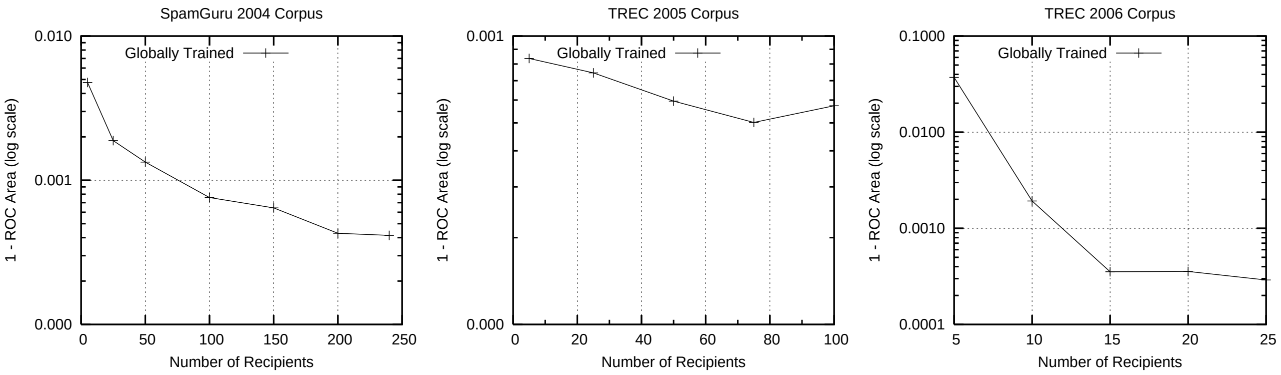

Figure 3: Scaling of a globally-trained classifier.

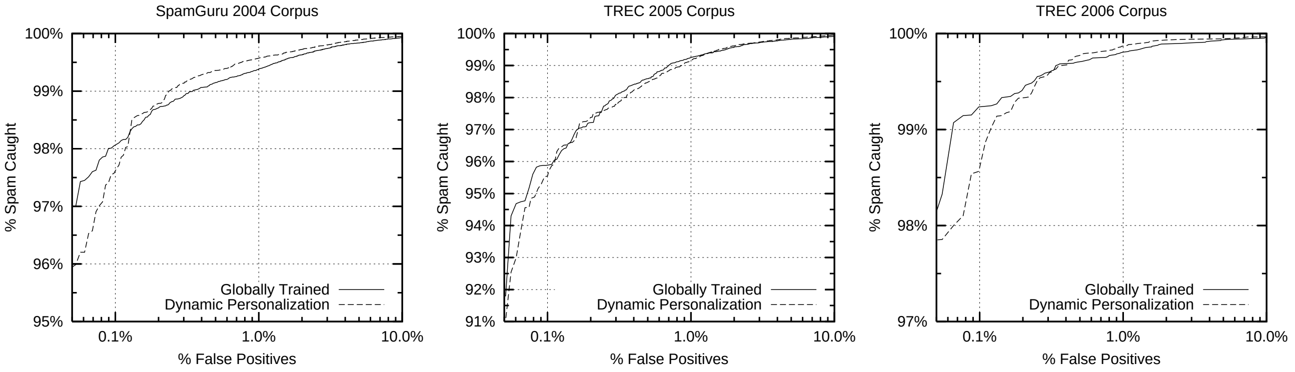

Figure 4: Comparison of a globally-trained classifier and dynamic-personalization.

each personal classifier. That is,

$$F(w_i, S, \theta_*) = \sum_{u \in U} F(w_i, S, \theta_u).$$

If a term only occurs for a single user, then the global statistics for that term will be the same as the personal statistics. As a result, one of the key advantages of personalization – the ability to identify unique aspects of each user's e-mail – is largely retained in a globally-trained classifier.

We can test this hypothesis by modifying our globally-trained classifier to ignore the data unique to each individual. That is, we classify each user's e-mail with a classifier trained on $\theta_{u-} = \theta_* - \theta_u$. If the globally-trained classifier did not benefit from personalization effects, we would expect this classifier to perform similarly to the classifier learned on θ_*. Figure 5 shows the results of this experiment for the SpamGuru 2004 and the TREC 2005 corpora. We omit the TREC 2006 dataset as the two largest users represent over 80% of the dataset. The results show that eliminating the personal information from a globally-trained classifier doubles the error rate across the ROC-curve. This is strong evidence that globally-trained classifiers make effective use of personal data.

The above analysis is also suggestive of when Dynamic Personalization would work best. Globally-trained classifiers cannot benefit from personalization data when users differ on how each term should be categorized. The term "mortgage" in a globally-trained classifier will either increase the spam score for everybody, decrease the spam score for everybody, or have little effect. If every user agrees that this term is indicative of spam, then a globally-trained classifier will treat the term as in indicator of spam for all users. But if half the users consider mortgage-related e-mail as spam and half the users consider it ham, then a globally-trained classifier will average these opinions and will treat "mortgage" as neither an indicator of spam or ham. Dynamic personalization will therefore be most useful for disparate user communities that have differing opinions of what is spam.

To test this hypothesis, we created a fourth dataset that is a combination of our three test corpora. We used roughly equal number of examples from each corpora to maximize the diversity within the combined corpus. We use the entire TREC 2006 corpora as it is the smallest. We sub-sampled the SpamGuru 2004 and TREC 2005 datasets by selecting the e-mail for the largest M users in each data set, where M was

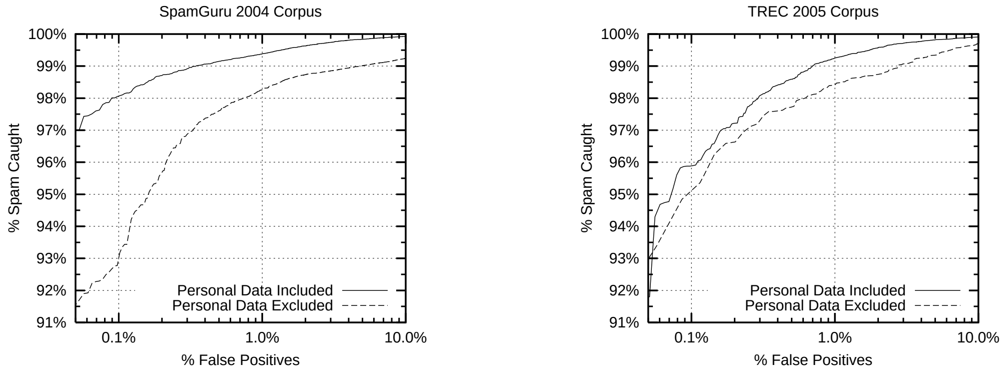

Figure 5: Comparison of a globally-trained classifier with and without personalization data.

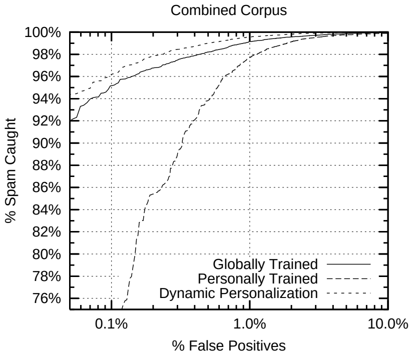

Figure 6: Comparison of a globally-trained classifier, a personally-trained classifier, and dynamic personalization on the combined SpamGuru 2004, TREC 2005, and TREC 2006 corpus.

chosen such that the number of selected messages was as close as possible to the number of messages in the TREC 2006 corpora. This process produced a new combined corpora with 77,935 examples and 58 users.

Figure 6 compares the performance of dynamic personalization and global classification on the combined dataset. The results show that dynamic personalization can be very effective for diverse user communities. Dynamic personalization achieved a 20% reduction in the number of missed spams across the entire ROC curve. The figure also shows the performance of a personally-trained classifier on this more diverse dataset. The personally-trained classifier cannot match the performance of the globally-trained classifier, let alone the performance of dynamic personalization. Interestingly, dynamic personalization makes effective use of personal classification despite the overall poor performance of personal classification on this dataset.

6 Related Work

Many personally-trained anti-spam filters offer the ability to leverage globally-trained data sources such as real-time blacklists and spam signature databases. For instance, SpamAssassin supports a variety of DNSRBL, URIBL, and signature databases (SpamAssassin, 2006). SpamAssassin combines the predictions from each external source with its own classification data using a fixed weighting scheme. Dynamic personalization differs in that it uses the specific structure of bag-of-word classifiers to dynamically choose between global and personal data at the level of individual tokens.

Cosoi combines globally- and personally-trained text classifiers by dynamically learning the optimal weights to combine the outputs of each classifier (Cosoi, 2007). Cosoi focuses on a filtering model in which the personally-trained classifier is continuously updated, but the user only infrequently retrieves the latest globally-trained classifier, say once a month. As a result, Cosoi focuses on the problem of how the weight assigned to the globally-trained classifier should be adapted over time to adjust to the timeliness of the global model.

Yerazunis discusses the importance of “inoculating” a personally-trained text classifier by sharing spam examples across users (Yerazunis, 2004). Inoculation works by training each user’s personal classifier on the spam examples forwarded by trusted colleagues. Our results confirm the value of sharing data across multiple users, and even across large collections of dissimilar users.

Graham has argued that personally-trained statistical filters are the key to an effective solution to the spam problem (Graham, 2003b; Graham, 2003a). He reasons that spammers will find it difficult to tune their content to bypass each individual user’s classifier. Our results suggest that globally-trained anti-spam filters may actually be a more effective solution due to the larger performance gains afforded by sharing all training data across a large user community. Dynamic Personalization may be better yet as it both leverages training data across all users and creates user-specific classifiers that may be more difficult for spammers to target.

7 Conclusion

We compared the performance of globally-trained and personally-trained text classifiers on three separate corpora. Our results show that globally-trained classifiers easily outperform personalized text classifiers for both small and large user communities. Furthermore, we demonstrate that globally-trained classifiers appear to scale well with performance improving as we increase the number of e-mail accounts being aggregated. We show that the ability of globally-trained classifiers to scale to large, diverse user communities is partially due to its ability to retain and use a substantial amount of personal data.

However, globally-trained classifiers cannot make use of all available personal data. A globally-trained classifier cannot honor the preferences of two users if the preferences of those users contradict. Dynamic personalization combines globally-trained and personally-trained text classifiers at the level of individual words to allow globally-trained text classifiers to take advantage of any available user-specific training data. Our results demonstrate that Dynamic Personalization offers a modest performance improvement on test corpora that includes diverse user communities.

References

Cormack, G. (2007). TREC 2006 Spam Track overview. *The Fifteenth Text REtrieval Conference (TREC2006) Notebook.*

Cormack, G., & Lynam, T. (2006). TREC 2005 Spam Track overview. *The Fourteenth Text REtrieval Conference (TREC 2005) Notebook.*

Cosoi, C. (2007). Combining antispam filters. *Proceedings of the MIT Spam Conference.*

Dietterich, T. G. (2000). Ensemble methods in machine learning. *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems* (pp. 1–15). Springer-Verlag.

Fawcett, T. (2003). Roc graphs: Notes and practical considerations for data mining researchers.

Graham, P. (2003a). Better bayesian filtering. *Proceedings of the MIT Spam Conference.*

Graham, P. (2003b). A plan for spam. *Proceedings of the MIT Spam Conference.*

Larkey, L., & Croft, W. (1996). Combining classifiers in text categorization. *SIGIR-96: 19th ACM International Conference on Research and Development in Information Retrieval* (pp. 289–297). Zurich.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *Proceedings of ECML-98, 10th European Conference on Machine Learning.*

Segal, R. (2005). Combining multiple classifiers. *Virus Bulletin.* February, 2005.

Segal, R. B., Crawford, J., Kephart, J. O., & Leiba, B. (2004). SpamGuru: An enterprise anti-spam filtering system. *Proceedings of the First Conference on Email and Anti-Spam.*

SpamAssassin (2006). The spamassassin open-source spam filter. http://www.spamassassin.org.

Yerazunis, W. (2004). The spam filtering plateau at 99.9% accuracy and how to get past it. *Proceedings of the MIT Spam Conference.*