UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

LEARNING-BASED SPAM FILTERS: THE INFLUENCE OF THE TEMPORAL DISTRIBUTION OF TRAINING DATA.

Anton Bryl

May 2006

Technical Report # DIT-06-030

Learning-Based Spam Filters: the Influence of the Temporal Distribution of Training Data.

Anton Bryl
University of Trento, Italy,
Create-Net, Italy
anton.bryl@dit.unitn.it

May 18, 2006

Abstract

The great number and variety of learning-based spam filters proposed during the last years cause the need in complex and many-sided evaluation of them, taking features of the phenomenon of spam into account. This paper is dedicated to the analysis of the dependence of filter performance on the temporal distribution of training data; the cause of this dependence is the changeability of email. Such analysis provides additional information about the filter quality, and also may be useful for organizing more effective training of the filter. The naïve Bayes filter is chosen for evaluation in this paper.

1 Introduction

After the naïve Bayes classifier was proposed for spam filtering by Sahami *et al.* in 1998 [12], great variety of learning-based filters started to appear. The plurality of available filtering methods results in the need for ways of complex evaluation and comparison of them. Anyhow, as we will see in Section 2, today evaluation of filters is largely one-sided, being concentrated on measuring filtering accuracy without considering the changeability of email and possible variations in the training process.

In this paper we propose an additional feature of a spam filter to be evaluated, namely the influence of the temporal distribution of the training data on the filtering accuracy. By the temporal distribution we understand the following two characteristics of training data: how long ago the data was gathered, and how long was the period during which it was gathered. The evaluation of this influence is necessary because of the noticeable changeability of email. Spam is known to be changeable because of different reasons, including variation of topics [7] and efforts of spammers to overcome the existing filters [6]. On the other hand, legitimate mail (often called "ham") also changes, often more abruptly than spam: for example, a user may subscribe to a popular mailing list that will influence the ham statistics greatly; or start active correspondence with a new friend from another country; or just touch upon a hot topic in his blog one day and receive hundreds of comment notifications instead of usual three or four. A spam filter is expected to deal with this "problem of two changeabilities" somehow, with as little user assistance as possible. The evaluation of the dependence of filtering accuracy on time and duration of gathering the training data not only gives an additional dimension of the filter quality, but may also help in organizing more effective and less labor-intensive training of this filter, thus potentially contributing to both evaluation and improvement of the filter's performance. The experiments proposed in this paper do not pretend to be an ample evaluation of the discussed dependence; still they give some interesting results.

Potentially this problem can be stated in a more general way: there is a need for the evaluation of the dependence of filter accuracy from various peculiarities of training process, including not only time, but also sources of training data. Such evaluation will allow to see, how

much efforts an individual user needs to spend in order to achieve reasonable filtering accuracy; thus we may characterize the discussed feature as related to a filter's *user-independence*, i.e. ability of a filter to perform well with minor efforts from the user's side.

We must also underline, that the results obtained for the naïve Bayes classifier will not necessarily be the same or similar for other filters, but may turn out to differ greatly depending on the learning algorithm and the way of the feature extraction.

The rest of the paper is organized as follows: in Section 2 we give an overview of state-of-the art in spam filter evaluation; in Section 3 we describe the data corpus used in this study; in Section 4 we describe the experiments and discuss the results; Section 5 is dedicated to possible directions of the future work; Section 6 is a brief conclusion; and Section 7 contains acknowledgements.

2 Related Work

Evaluation of a spam filter is usually performed by running series of tests with applying the filter to a previously gathered data corpus. A data corpus for spam filter evaluation is a quite large set of email messages sorted into spam and ham. A number of corpora are available online for the public use. A methodology of creation of a data corpus for spam filter evaluation is discussed in [4]. Creation of new public corpora is slowed down by privacy issues; for sure, people are usually unwilling to publish their private email. For this reason many studies use either corpora that are not publicly available, or both private and public corpora. One can also test a filter "in real-life conditions", simply using in for sorting the mail in his mailbox for a while [10], but this way is obviously more time-consuming. A popular way of using a data corpus for testing is splitting it into training and testing data in some way for each run [1, 8]. Another approach is giving messages to a filter one by one and correcting the classification errors immediately, so that the messages are gradually added to the training data [3].

The simplest measure of filter quality is the classification accuracy, i.e. percentage of mail correctly classified [8]. More precise evaluation may consider false positives and false negatives separately, using such measures as spam recall (the percentage of spam correctly detected), spam precision (the percentage of spam in the whole amount of blocked email), and error rates. Androutsopoulos *et al.* [1] propose to use relative cost of the two types of errors as a variable parameter, and three measures based on this parameter are introduced: weighted spam recall, weighted spam precision, and TCR (Total Cost Ratio) measure, that represents the relative cost of using no filter at all (and so having all the spam classified as ham, but no ham classified as spam) to using the filter (and so having some false positives and some false negatives). For the filters that allow to modify the "spam-likeness" threshold for blocking spam, a receiver operating characteristics (ROC) curve may be built, that shows the combinations of the spam detection rate and the false positive rate obtained for different thresholds [9]. Sometimes a previously known filter is involved in testing to provide a quality baseline (the naïve Bayes filter is often used for this purpose).

Apart from the accuracy measures, a number of other features are evaluated from time to time. Drucker *et al.* in [5] give a judgement on the training and classification speed. Boykin *et al.* [2] analyze possible countermeasures that spammers may take to cheat their filter. Androutsopoulos *et al.* [1] evaluate the dependence of performance on training data size and attribute set size. Cormak *et al.* [3] use learning curves to see how filter performance changes with time in the assumption that the user continues to retrain the filter all the time by correcting most of the classification errors. In several papers [1, 8] the effect of data preprocessing (different combinations of stemming and stoping) is considered.

To sum up what is said above, we can claim that the evaluation of the filter accuracy is quite well-developed and widely applied; at the same time the use of other quality measures is occasional and unsystematic, though anyhow present in the literature.

3 Data Corpus

For our study we prepared a special data corpus, based on the author's private mailbox. The corpus contains messages received from September 2005 till March 2006, sorted manually into spam and ham. The sources of the legitimate mail are quite various, including private communications, one mailing list, and notifications from several

	09.05	**10.05**	**11.05**	**12.05**	**01.06**	**02.06**	**03.06**	***Total***
Number of messages	555	527	242	306	360	503	677	*3170*
Spam rate	67%	52%	62%	43%	72%	73%	72%	64%

Table 1: Corpus statistics.

websites. The messages are in four different languages (Belarusian, Russian, English, and Italian), of which the first two are highly inflectional, i.e. have great number of forms for the most of the words. For this reason we chose to analyze headers, not bodies of the messages. Some corpus statistics are presented in the table 1. The fields of the headers added by the local mail server's spam filter were deleted, as well as the "spam" and "possible spam" marks added by this filter to the subject lines. We must notice, that a corpus of this size will allow us to see the short-time effects of the email changeability, but in case of experiments related to the long-time effects a larger corpus will be needed.

4 Experiments

During this study we have performed two experiments, both of them using the naïve Bayes classifier. Basic measures used in the experiment were spam recall (number of correctly recognized spam divided by total number of spam) and false positive rate (number of ham messages classified as spam divided by total number of ham).

The goal of the first experiment was to see how the performance changes with time after the last retraining of the filter. For this purpose a series of tests was performed. In each test the filter was trained on a set of messages from one month (further called *training month*), and then tested on messages of one of the following months (further called *testing months*). Tests for all possible pairs of training and testing months were held. For training the first 240 messages of the training month were taken in each case. For testing all the messages of the testing month were used. The results of this experiment are presented in Table 2.

The goal of the second experiment was to see if an increase of the length of training period without changing the amount of training data influences the filter accuracy. In each test in this experiment the filter was trained on a set of messages from a pair of subsequent months (further called *training two-month*), and then tested on messages of one of the following months. The training data was prepared in the following way: the first 240 messages were taken from each of the months in the training two-month, so that the total amount of selected data equalled to 480 messages; then every second of this messages was deleted from the training corpus, so that the amount of training data became equal to 240 messages. For testing all the messages of the testing month were used, as in the first experiment. The results of the second experiment are presented in Table 3.

Looking at the results of the first experiment, we can see that the spam recall is quite high and stable (never below 97%), but the false positive rate is high (up to 17%), and varies greatly. Training on the data gathered in December in all cases lead to higher false positive rate than training on the data gathered in November or January; also in all cases the false positive rate is higher in the tests for January than in the tests for December or February. In general, it seems that the changeability of the input influences the filter greatly, but it results not in the gradual monotonic decrease of performance, but in quite unpredictable jumps, that are likely to be dependant on local features of the data of different months (e.g. noticeable amount of messages from previously unknown correspondents in January) rather than on some general rules.

In the second experiment the filter shows clearly better results. The false positive rate decreases, while the spam recall remains approximately the same; often training on the combination of two months leads to lower false positive rate than training on any of them. More accurately, in 5 cases out of the 15 the training on the two-month outperforms the training on any of the months in the pair; in one case the two-month result equals to the best of the two one-month results; in 5 cases the two-month results is between the one-month results and better than the average of them; in 3 cases the two-month results is between the one-month results and worse than the average of them;

		Tested on											
		Oct'05		**Nov'05**		**Dec'05**		**Jan'06**		**Feb'06**		**Mar'06**	
		SR	*FP*	*SR*	*FP*	*SR*	*FP*	*SR*	*FP*	*SR*	*FP*	*SR*	*FP*
Trained on	**Sep'05**	98.9	2.4	99.3	3.3	97	3.5	100	7.9	99.7	3.6	100	1.6
	Oct'05			100	2.2	97	5.2	100	16.8	99.7	2.2	100	2.1
	Nov'05					97	1.7	100	5	99.7	2.2	100	0
	Dec'05							100	17.8	99.7	15.2	100	13.2
	Jan'06									99.7	3.6	100	0
	Feb'06											100	4.2

Table 2: Experiment 1: spam recall (SR, %) and false positive rate (FP, %).

		Tested on									
		Nov'05		**Dec'05**		**Jan'06**		**Feb'06**		**Mar'06**	
		SR	*FP*	*SR*	*FP*	*SR*	*FP*	*SR*	*FP*	*SR*	*FP*
Trained on	**Sep'05 - Oct'05**	99.3	1.1	97	2.3	100	5.9	99.5	1.4	100	4.2
	Oct'05 - Nov'05			97	2.9	100	7.9	99.5	0	100	0.5
	Nov'05 - Dec'05					100	13.9	99.7	12.3	100	9.5
	Dec'05 - Jan'06							99.7	3.6	100	0.5
	Jan'06 - Feb'06									100	0.5

Table 3: Experiment 2: spam recall (SR, %) and false positive rate (FP, %).

and in one case the two-month result is worse than the one-month results.

From the results of the experiments we can conclude that training the naïve Bayes classifier on the messages gathered for a longer time period may give performance improvement in comparison to a shorter period, even if the number of messages is the same. The most possible reason for this is that during the longer time period the greater variety of spam and ham appears in one's mailbox, so the filter has less date-specific information to learn on. A useful consequence of this observation is that the performance evaluation based on splitting a corpus into training and testing data randomly, like when using the 'Bow' toolkit[11], may lead to delusive results: in such an experiment the filter may perform better than in reality due to the fact that the training messages are chosen randomly from all over the corpus, and thus from all over the time interval.

As we already mentioned above, one practical outcome of such evaluation is having hints to better ways of training the evaluated filter. In our case (for naïve Bayes) it is likely that the same amounts of data gathered during the time period of different length lead to different performance, namely the shorter training period gives worse results. Anyhow, to find the best way of gathering of training data for the filter, more experiments on a larger data corpus (or better several corpora) are needed.

5 Future Work

The experiments presented in this paper are far from being exhaustive, and there are thing to do to extend and generalize the results. Possible future work in this direction includes: performing the same kind of testing for other algorithms and/or ways of feature selection; performing tests on a corpus gathered during a longer period of time to see the effects of gradual improvement of spamming technologies; finding a systematic way of evaluation of dependence of filtering accuracy on various peculiarities of training process; uncovering the events in the training period that influence the filter accuracy most seriously (it may be a holiday that causes a great number of greetings from lots of different people; or a local increase of activity on a mailing list; or whatever).

6 Conclusion

In this paper we have introduced an attempt to evaluate dependence of naïve Bayes spam filter accuracy on the temporal distribution of training data. Two experiments were performed, with the results presented in this paper. From the results of the experiments we can make the following conclusions:

1. Due to the changeability of email the data gathered during a short period of time may be too specific for training the filter, so that it may fail to show reasonable accuracy even in the next month; actually, the performance of the filter learned in this way may well be called unpredictable.

2. Temporal distribution of the training data influences the performance distinctly (for naïve Bayes training on data gathered during two months gives better results than training on the same amount of data gathered during one month). One consequence of this is that random splitting of an experimental corpus into training and testing data, that provides unrealistic temporal distribution of training data (training messages both "in the past" and "in future" relative to testing messages), may lead to delusive results. Another consequence is that evaluation of this dependence may be helpful in organizing more effective filter training.

7 Acknowledgments

I would like to thank my supervisors, Prof. Fabio Massacci and Prof. Enrico Blanzieri, for their support throughout my studies.

References

[1] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos. 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), pp. 9-17, 2000.

[2] P. O. Boykin and V. P. Roychowdhury. Leveraging Social Networks to Fight Spam, Computer, Vol. 38, No. 4, pp. 61-68, April 2005.

[3] G. Cormack and T. Lynam. TREC 2005 Spam Track Overview, *available at* http://plg.uwaterloo.ca/ gvcormac/trecspamtrack05/

[4] G. Cormack and T. Lynam. Spam Corpus Creation for TREC, CEAS'2005 http://www.ceas.cc/, 2005.

[5] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization, IEEE Transactions on Neural networks, v. 10(5), pp. 1048 - 1054, September 1999.

[6] T. Fawcett. "In vivo" spam filtering: a challenge problem for KDD, SIGKDD Explor. Newsl., v. 5, pp. 140–148, 2003.

[7] G. Hulten, A. Penta, G. Seshadrinathan, and M. Mishra. Trends in Spam Products and Methods, CEAS'2004 http://www.ceas.cc/, 2004.

[8] C.-C. Lai and M.-C. Tsai. An empirical performance comparison of machine learning methods for spam e-mail categorization, in proceedings of HIS 2004, Fourth International Conference on Hybrid Intelligent Systems, pp. 44 - 48, 2004.

[9] B. Leiba, J. Ossher, V. T. Rajan, R. Segal, and M. Wegman. SMTP Path Analysis, CEAS'2005 http://www.ceas.cc/, 2005.

[10] E. Michelakis, I. Androutsopoulos, G. Paliouras, G. Sakkis, and P. Stamatopoulos. Filtron: A Learning-Based Anti-Spam Filter, CEAS'2004 http://www.ceas.cc/, 2004.

[11] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, http://www.cs.cmu.edu/ mccallum/bow, 1996.

[12] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian Approach to Filtering Junk E-Mail, Learning for Text Categorization: Papers from the 1998 Workshop, 1998.