Spam Filtering by Quantitative Profiles

Marian Grendár¹, Jana Škutová² and Vladimír Špitalský³

¹ Slovanet, a.s.
Záhradnícka 151, 821 08 Bratislava, Slovakia

² Slovanet, a.s.
Záhradnícka 151, 821 08 Bratislava, Slovakia

³ Slovanet, a.s.
Záhradnícka 151, 821 08 Bratislava, Slovakia

Abstract
Instead of the “bag-of-words” representation, in the quantitative profile approach to spam filtering and email categorization, an email is represented by an m-dimensional vector of numbers, with m fixed in advance. Inspired by email shape analysis proposed recently by Sroufe et al., two instances of quantitative profiles are considered: line profile and character profile. Performance of these profiles is studied on the TREC 2007, CEAS 2008 and a private corpuses. At low computational costs, the two quantitative profiles achieve performance that is at least comparable to that of heuristic rules and naive Bayes.

Keywords: Email Categorization, Spam Filtering, Quantitative Profile, Character Profile, Line Profile, Random Forest

1. Introduction
Spam is an unsolicited email message. From the receiver’s perspective, spam is an annoyance and thus it is necessary to block its delivery, for instance, by filtering it out.

Traditional approach to spam filtering and email categorization (cf., e.g. [2-4], [9], [12-14], [16], [22]) that is based on heuristic rules, naive Bayes filtering and/or textmining suffers from several deficiencies. Among shortcomings of the traditional approach there are high computational costs, language dependence (cf., e.g. [1]), necessity to update the heuristic rules, high number of rules, and vulnerability.

Instead of the “bag-of-words” representation, employed in text-mining and naive Bayes filtering, in the quantitative profile (QP) approach that we propose, an email is represented by an m-dimensional vector of numbers with m fixed in advance. Inspired by Sroufe, Phithakkitnukoon, Dantu, and Cangussu [21], two instances of QPs are considered: line profile (LP) and character profile (CP). Informally put, the line profile of an email is a vector of lengths of the first m lines. The character profile is a histogram of characters. Of course, many other QPs are conceivable.

The main advantages of the two considered quantitative profiles are i) sound performance, ii) simple computability, iii) language-independence, iv) robustness to outlying emails, v) high scalability and vi) low vulnerability.

The considered two instances of the quantitative profile approach perform comparably to the naive Bayes filtering and heuristics-based approaches and perform very well also in a multi-language, non-English communication. Furthermore, the satisfactory performance is attained by means of a small set of easily computable quantitative features. A performance study was done on the TREC 2007, CEAS 2008 and a private corpuses. To demonstrate the power of the considered QPs, the profiles are obtained from raw emails, without any preprocessing. Consequently, we intentionally ignore the structure of emails and character encoding.

On the two QPs, the Random Forest algorithm substantially outperforms other classification algorithms (SVM, LDA/QDA, logistic regression). Thanks to the Random Forest, the QP approach gains robustness to emails...
with extreme-valued profiles as well as high scalability.

In our view, classification and email categorization by LP (or CP) should have low vulnerability. For, the lines the lengths of which differentiates between spam and ham, change from corpus to corpus. For instance, in the CEAS 2008 corpus, the most important for deciding between spam and ham are the lengths of the (10, 17, 15, 14, 16)-th lines, whilst in the TREC 2007 corpus they are the (5, 13, 6, 15, 7)-th lines. As the training corpus is usually not available to a spammer, it should be not easy to evade the LP (or CP) filter.

As a by-product of a performance study of CP and LP, we note that in the TREC 2007 and CEAS 2008 corpuses the number of header lines is capable of discriminating between spam and ham, at a rate that is, in our view, too high. Consequently, the corpuses lead to overly optimistic performance of spam filtering methods.

The paper is organized as follows. In the next section we formally introduce the QPs mentioned above. Then, in Section 3, we describe the three email corpuses used for assessment of the QPs’ performance. In Section 4 we describe measures used for performance evaluation. The results are summarized in Section 5. In the concluding section some directions for future research are briefly discussed. All the computations were performed with R [15]. To make the results reproducible, a supplementary material including the source code was prepared, cf. [8].

2. Quantitative profiles

The quantitative profile (QP) of an email is an m-dimensional vector of real numbers that represents the email. The dimension m of the profile is set in advance, and it is the same for all emails. In this paper we consider two particular QPs – the line profile and the character profile. These profiles can be introduced by means of a simple probabilistic model.

An email is represented as a realization of a vector random variable, that is generated by a hierarchical data generating process. The length n of an email is an integer-valued random variable, with the probability distribution F_n. Given the length, the email is represented by a random vector X^n = (X_1, . . . , X_n) from the probability distribution F^n_{X^n} with the support in A^n, where A = \{a_1, . . . , a_m\} is a finite set (alphabet) of size m = |A|.

Then, the character profile (CP) of an email is an m-dimensional random vector CP = (CP_1, . . . , CP_m), where

\[ CP_j = \sum_{i=1}^{n} I_{\{X_i = a_j\}}, \quad j = 1, \ldots , m, \]

and I is the indicator function.

Fig. 1: Graphical representation of the line and character profiles of an email

In order to introduce the other QP, it is necessary to select a special character (or a subset of characters) from the alphabet. Let k be the number of occurrences of the special character in an email and let T_j (j = 1, . . . , k) be the index of the j-th occurrence of the special character; put T_0 = 0. Then the binary profile (BP) of an email is defined as a k-dimensional random vector \( BP = (BP_1, \ldots , BP_k) \), where

\[ BP_j = T_j - T_{j-1} - 1, \quad j = 1, \ldots , \hat{k}. \]

There, \( \hat{k} \) is the maximum allowable number of the occurrences of the special character and it is set in advance. Hence, if \( k > \hat{k} \), the rest of the email is ignored. And if an email has \( k < \hat{k} \), \( BP_j = 0 \) for \( j > k \).

In this work, an email is understood as a stream of bytes without any preprocessing, so that A is taken to be the ASCII character set. The end-of-line is taken for the special character. Consequently, the binary profile becomes the line profile (LP) since \( BP_j \) is the length of the j-th line of an email in bytes. We fix the maximal number \( \hat{k} \) of considered lines to 100.

3. Data sets

To assess the performance of the two quantitative profiles, we consider three email corpuses: the publicly
available TREC 2007 and CEAS 2008 corpuses, and a private corpus.

The TREC 2007 corpus [7] comprises over 75 000 emails (25 220 hams and 50 199 spams), of which approximately 67% is spam, and the rest are ham emails. For the training phase we used the first 50 000 emails. The rest forms the test set. The ratio of spam to ham in the training and test sets is approximately 2:1.

The other publicly available corpus, used for the performance analysis of QPs, is CEAS 2008 [6]. It consists of 137 705 emails (27 126 hams and 110 579 spams). The corpus was hand labeled [18]. To form the training set, we used the first 90 000 emails. The test set comprises the remaining emails. The ratio of spam to ham in the training and test sets is approximately 4:1.

Performance of spam filtering algorithms is typically assessed on English language corpuses, such as the above mentioned TREC and CEAS. When applied to non-English emails, their performance may be different. Due to the support from a Slovak internet services provider, we enjoyed the opportunity of access to a private corpus created in 2010, that comprises mainly non-English emails. Structure of the corpus is summarized in Table 1. The training set we used consists of 11 050 emails and the test set consists of 12 200 emails.

Table 1: Composition of the private corpus

| corpus | s.ham | advert | notify | spam | total |
|--------|-------|--------|--------|------|-------|
| train  | 6837  | 1611   | 1225   | 1377 | 11 050|
| test   | 3650  | 1409   | 5758   | 1383 | 12 200|

The private corpus was hand labeled. Emails were placed into one of the two groups: ham and spam. Moreover, ham was divided into advert, solicited ham (denoted s.ham) and notify. Based on the language of the major part of an email, the emails were placed into one of the four language groups: Slovak/Czech, English, German, and other. The corpus comprises 64% of the Slovak/Czech ham in the training set and 38% in the test set.

4. Classification algorithm and performance measures

Quantitative profiles serve as an input to a classification algorithm. In this work we use the Random Forest classifier, introduced by Breiman [5] and ported to R by Liaw and Wiener [11], with the default settings. We have employed also LDA, logistic regression, LASSO and SVM [10], but these methods performed much worse.

To evaluate QPs’ performance, we calculate the false positive rate \( fpr = FP/(TN + FP) \) and the false negative rate \( fnr = FN/(TP + FN) \), where \( TP \) (TN) stands for the number of true positive (true negative) emails, i.e. the correctly recognized spam (ham) emails; and \( FP \) (\( FN \)) stands for the number of false positive (false negative) emails, i.e. the incorrectly recognized ham (spam) emails, respectively. We also present the receiver operating characteristic (ROC) curve [19], i.e. the graph of the true positive rate vs. the false positive rate, obtained as functions of the decision threshold. The area \( AUC \) under the ROC curve is also reported.

5. Results

In this section we summarize performance of the basic QPs on the three corpuses mentioned above.

5.1 Comparison of quantitative profiles with SpamAssassin and Bogofilter

For the sake of comparison, we report also results for SpamAssassin (SA) [20], version 3.3.1, off-line and without the Bayes filter, and Bogofilter (BF) [17], version 1.2.2 with the default configuration. On the private corpus, the output from SA was processed by the Random Forest classification algorithm, as it attains much better performance than SA with the default weights. In addition to much better performance, it allows for email categorization, which is impossible with the default SA. On the public corpuses, however, the default SA classification performs better. Bogofilter allows a binary classification only. BF was learnt in the batch mode.

![ROC curves](image)

(a) private corpus  (b) TREC 2007 corpus

**Fig. 2: ROC curves**
Table 2: fnr (%) at fixed fpr = 0.5% or fpr = 1%

| filter | private at 0.5% | private at 1% | TREC07 at 0.5% | TREC07 at 1% | CEAS08 at 0.5% | CEAS08 at 1% |
|--------|----------------|--------------|---------------|--------------|----------------|--------------|
| CP     | 14.39          | 11.64        | 2.53          | 0.49         | 4.38           | 4.25         |
| LP     | 21.33          | 20.10        | 0.30          | 0.13         | 0.39           | 0.27         |
| SA     | 12.68          | 10.10        | 35.87         | 30.51        | 76.14          | 69.92        |
| BF     | 13.05          | 7.38         | 0.40          | 0.06         | 0.47           | 0.36         |

On the private corpus, for fpr = 0.5% the best performance is attained by SA and BF, and CP performs only slightly worse, cf. Table 2 and Figure 2. LP is slightly less effective, and it attains fnr around 21% at fpr = 0.5%.

With the exception of SA, on the public corpuses all the studied filters attain much better performance than on the private corpus, see also Section 5.6. The two QPs perform much better than SA as Table 2 as well as Figure 2 indicates. The line profile attains better performance (smaller fnr) than BF at the 0.5% level of fpr.

5.2 Email categorization for the private corpus

On the private corpus, performance of CP and especially LP in spam filtering (i.e. binary categorization) is worse than that of SA and BF. However, in categorization of emails into one of the four categories, both CP and LP perform much better in the most interesting category of solicited ham. Misclassification table for CP and SA is in Table 2.

5.3 Comparison with email shape analysis

Sroufe et al. [21] suggest to filter spam by means of its shape, which the authors define (using our terminology) as a smoothed line profile of email body, where smoothing is performed by the kernel smoother. Sroufe et al. also report the total error of 30%, based on a preliminary study on the TREC corpus. Further, the authors find the performance very good, 'considering that no content or context was even referenced', cf. [21], p. 26. The line profile, that is inspired by the email shape analysis, attains on the TREC corpus fnr = 4.23% a fnr = 17.00%, when the threshold is not optimized and email headers are intentionally not taken into account, cf. Section 5.6.

This gives the total error around 12.29% and indicates that smoothing is unnecessary.

5.4 Choice of LP dimension

The dimension k of the line profile is a free parameter. In order to choose the value of k it is necessary to explore the dependence of the LP filter’s performance on k. Figure 3 depicts fnr for different values of k and the two values 0.5% and 1% of fpr, for the private corpus. As fnr stabilizes at its lowest values for k ≈ 40, the choice of k = 100 is justified on the ‘safe bet’ ground.

5.5 Reduction of the feature space

It is also important to know to what extent the dimension of the QP feature space can be reduced without substantive reduction of the classifier’s performance. To this end the top 20 and the top 50 features were considered, where the ranking of features was provided by the Random Forest’s measure of the mean decrease of accuracy. The study was done on the private corpus and solely the email body was considered.

In the binary classification the top 20 features of the line profile attain essentially the same performance as the entire line profile of the length 100. In the case of CP, to attain the full-set performance, the top 50 features out
Table 3: CP and SA confusion tables for categorization, private corpus

| count | CP advert | s.ham | notify | spam | SA advert | s.ham | notify | spam |
|-------|-----------|-------|--------|------|-----------|-------|--------|------|
| advert| 530       | 837   | 10     | 32   | 515       | 830   | 17     | 47   |
| ham   | 26        | 3597  | 27     | 0    | 95        | 3290  | 248    | 17   |
| notify| 20        | 237   | 5499   | 2    | 24        | 235   | 5498   | 1    |
| spam  | 37        | 221   | 12     | 1113 | 44        | 161   | 14     | 1164 |

of 256 are needed. The same holds for SA. However, in the case of the email categorization it is not possible to reduce the dimension of SA features without substantive decrease in accuracy.

5.6 Why are TREC and CEAS misleading?

All the considered email filters except of SA attain much better performance on the public corpuses than on the private one; cf. Table 2. In search for explanation we have noted that in the public corpuses, unlike to the private corpus, the number of header lines contains information that is substantive for spam filtering. Figure 4 depicts the distribution of the number of lines in header, for spam and ham, in the TREC corpus.

Fig. 4: Distribution of emails with respect to the number of header lines, for spam and ham, in the TREC 2007 corpus

Once the email header is not taken into account, and solely the email body is processed, performance of LP and CP worsens and becomes comparable to that on the private corpus; cf. Table 4. In Table 4, LPH (CPH) denotes the line (character) profile of email header and LPB (CPB) denotes the line (character) profile of email body, respectively.

The decline of performance of CP and LP caused by exclusion of email headers supports the hypothesis that in the TREC 2007 and CEAS 2008 corpuses the profiles of headers carry a substantive information for discriminating between spam and ham.

5.7 Summary of the performance study

The empirical study implies that the simple and easily obtainable line and character profiles attain at least comparable performance as the optimally tuned SpamAssassin, which is based on hundreds of fixed rules, and the performance of character profiles is close to that of Bogofilter. Particularly, on the public corpuses LP is better than BF and SA. On the private corpus CP attains comparable performance as BF and SA, and LP is slightly worse.

6. Conclusions

Motivated by Sroufe et al. [21], we have proposed the quantitative profile approach to email classification. In this communication we explored two quantitative profiles, the line profile and the character profile. The profiles are obtained from raw emails, without any preprocessing. The computational costs of the two profiles are minimal. Performance of the profiles was studied on the TREC 2007, CEAS 2008 corpuses and a private, multi-lingual corpus. The two quantitative profiles attained at least comparable performance as the optimally tuned SpamAssassin and the batch-mode learnt Bogofilter. Besides the good performance, the two quantitative profiles are language independent and the resulting filter is robust to outlying emails, highly scalable and has low vulnerability.

As a by-product, we have noted that the number of header lines in the TREC 2007 and CEAS 2008 corpuses contain rather strong information on the email class. The corpuses thus lead to overly optimistic performance of spam filters.

In the near future we plan to explore quantitative profiles based on size and structure of emails, on the symbolic dy-
Table 4: fnr (%) at fixed fpr = 0.5% or fpr = 1%

| filter | private at 0.5% | at 1% | TREC07 at 0.5% | at 1% | CEAS08 at 0.5% | at 1% |
|--------|----------------|-------|----------------|-------|----------------|-------|
| CPH    | 16.51          | 13.71 | 0.19           | 0.05  | 0.78           | 0.30  |
| LPH    | 18.06          | 15.11 | 1.78           | 0.12  | 0.68           | 0.36  |
| CPB    | 14.24          | 11.92 | 15.05          | 5.47  | 4.70           | 4.51  |
| LPB    | 21.26          | 18.58 | 45.01          | 43.67 | 7.07           | 6.35  |

namic, and another instances of the binary profile. Also, the profiles are worth employing in a semi-supervised email categorization.

Acknowledgments

Stimulating feedback from Peter Farkaš, Ján Gallo, Peter Kočík, Erik Lehotský, Dušan Slivoň and Stanislav Záriš is gratefully acknowledged. This paper was prepared as a part of the project “SPAMIA”, MS SR - 3709/2010-11, supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic, under the heading of the state budget support for research and development.

References

[1] T. A. Almeida, and A. Yamakami, "Content-based spam filtering", in Proceedings of the 23rd IEEE International Joint Conference on Neural Networks (IJCNN’10), Barcelona, Spain, 2010.

[2] T. A. Almeida, and A. Yamakami, "Advances in spam filtering techniques", in Computational Intelligence for Privacy and Security. Book Series: Studies in Computational Intelligence, Springer, Berlin/Heidelberg, 2012, to appear.

[3] I. Androutsopoulos, G. Paliouras, and E. Michelakis, "Learning to filter unsolicited commercial e-mail", National Centre for Scientific Research "Demokritos", Athens, Greece, Tech. Rep. 2004/2, (2004).

[4] P. S. Bogawar, and K. K. Bhoyar, "Email mining: a review", IJCSI International Journal of Computer Science Issues, Vol. 9, No. 1, 2012.

[5] L. Breiman, "Random forests”, Machine Learning, Vol. 45, No. 1, 2001, pp. 5-32.

[6] G. V. Cormack, and T. R. Lynam, "CEAS 2008 corpus”, 2008.

[7] G. V. Cormack, and T. R. Lynam, "TREC 2007 corpus”, 2007.

[8] M. Grendár, J. Škutová, and V. Špitalský, "Supplement to “Spam filtering by quantitative profiles”", 2011.

[9] T. Guzella, and W. Caminhas, "A review of machine learning approaches to spam filtering”, Expert Systems with Applications, Vol. 36, 2009, pp. 10206-10222.

[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, New York, 2009.

[11] A. Liaw, and M. Wiener, "Classification and regression by randomForest”, R News, Vol. 2, No. 3, 2002, pp. 18-22.

[12] T. M. Mahmoud, and A. M. Mahfouz, "SMS spam filtering technique based on artificial immune system", IJCSI International Journal of Computer Science Issues, Vol. 9, No. 1, 2012.

[13] A. N. Pathak, M. Seghal, and D. Christopher, "A study of selective data mining algorithms", IJCSI International Journal of Computer Science Issues, Vol. 8, No. 2, 2011.

[14] J. Pruthi, and E. Kumar, "Anti-trust rank: fighting web spam”, IJCSI International Journal of Computer Science Issues, Vol. 8, No. 1, 2011.

[15] R Development Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, (2010).
[16] M. Rafi, S. Hassan, and M. S. Shaikh, "Content-based text categorization using Wikitology", IJCSI International Journal of Computer Science Issues, Vol. 9, No. 4, 2012.

[17] E. S. Raymond, D. Relson, M. Andree, and G. Louis, "Bogofilter", (2004).
http://Bogofilter.sourceforge.net

[18] R. Segal, A. Bratko, and G. Cormack, "CEAS 2008 Spam Filter Challenge", conference talk, 2008.
http://www.ceas.cc/2008/challenge/results.pdf

[19] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: Visualizing the performance of scoring classifiers", R package version 1.0-4., 2009.

http://CRAN.R-project.org/package=ROCR

[20] SpamAssassin.
http://spamassassin.apache.org

[21] P. Sroufe, S. Phithakkitnukoon, R. Dantu, and J. Cangussu, "Email shape analysis", in Distributed Computing and Networking, K. Kant, S. V. Pemmaraju, K. M. Sivalingam, J. Wu (Eds.), Lecture Notes in Computer Science 5935/2010, Springer, New York, 2010, pp. 18-29.

[22] J. Zdziarski, Ending spam: Bayesian Content Filtering and the Art of Statistical Language Classification, No Starch Press, 2005.