SPAMIA: Spam filtering by quantitative profiles

Marián Grendár, Jana Škutová, Vladimír Špitalský

Slovanet a.s., Záhradnícka 151, 821 08 Bratislava, Slovakia
marian.grendar, jana.skutova, vladimir.spitalsky@slovanet.net

Applied Statistics 2012, International conference
September 23 - 26, 2012, Ribno (Bled), Slovenia

This presentation was prepared as a part of the "SPAMIA" project, MŠ 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.

Content

Spam

- an unsolicited email message
- is ussually send in a bulk to spread advert or viruses, or for phishing, scam, verification of email, ...

Existing solutions for spam filtering

Methods

- heuristic rules
- naive Bayes filtering
- text-mining methods

Open-source products

SpamAssassin

Bogofilter

DSPAM

...

Comercial products

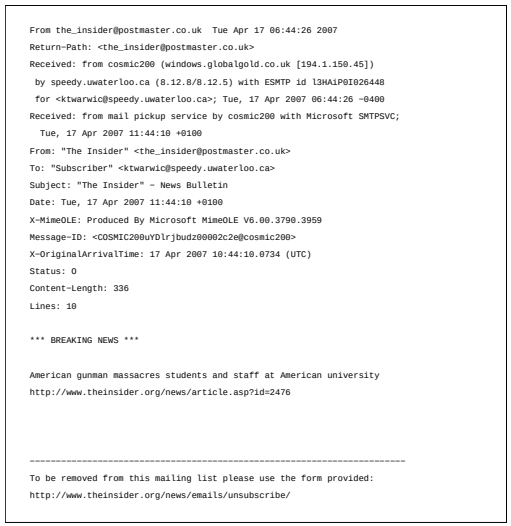
Disadvantages of existing solutions

- language dependence
- heuristic rules are fixed
- necessity to update these rules
- high vulnerability
- high computational costs

Quantitative profile approach

Quantitative profile approach

- an email is represented by an m-dimensional vector of numbers with m fixed in advance
- QPs serve as an input to a classification algorithm



```
From the_insider@postmaster.co.uk  Tue Apr 17 06:44:26 2007
Return-Path: <the_insider@postmaster.co.uk>
Received: from cosmic200 (windows.globalgold.co.uk [194.1.150.45])
 by speedy.uwaterloo.ca (8.12.8/8.12.5) with ESMTP id l3HAiP0I026448
 for <ktwarwic@speedy.uwaterloo.ca>; Tue, 17 Apr 2007 06:44:26 -0400
Received: from mail pickup service by cosmic200 with Microsoft SMTPSVC;
  Tue, 17 Apr 2007 11:44:10 +0100
From: "The Insider" <the_insider@postmaster.co.uk>
To: "Subscriber" <ktwarwic@speedy.uwaterloo.ca>
Subject: "The Insider" - News Bulletin
Date: Tue, 17 Apr 2007 11:44:10 +0100
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.3790.3959
Message-ID: <COSMIC200uYDlrjbudz00002c2e@cosmic200>
X-OriginalArrivalTime: 17 Apr 2007 10:44:10.0734 (UTC)
Status: O
Content-Length: 336
Lines: 10

*** BREAKING NEWS ***

American gunman massacres students and staff at American university
http://www.theinsider.org/news/article.asp?id=2476



---------------------------------------------------------------------------
To be removed from this mailing list please use the form provided:
http://www.theinsider.org/news/emails/unsubscribe/
```

$$QP = (qp_1, qp_2, \ldots, qp_m)$$

Basic quantitative profiles

Binary profile: distances between occurences of special character/characters (only first $k = 100$ occurences for each email)

- LP line: lengths of lines
- WP word: lengths of words
- BRP brackets: distances between brackets
- ...

Histogram binary profile:

- HWP: histogram of lengths of words
- HBRP: histogram of distances between brackets
- ...

Basic quantitative profiles

Character profile: the number of occurrences of the characters

- CP: characters from $\mathcal{A}$ (ASCII character set)

Grouped character profile: the number of occurrences of the groups of characters

- CPG9: numbers, spaces, brackets, operators, separators, upper/lower-case letters, forbidden characters, other
- CPG11: as CPG9, separately ! a $

d-gram grouped character profile:

- 2CPG11: pairs of groups of characters
- 3CPG11: triples of groups of characters
- ...

Basic quantitative profiles

Moving window profile: CPGs for each parts of email

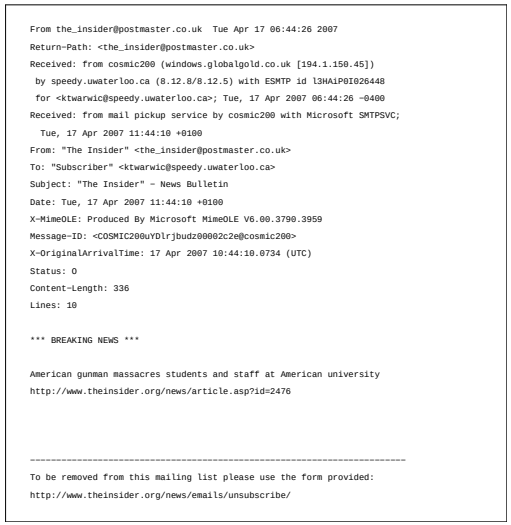
- MWPCPG11

Size profile:

- size of email
- sizes of selected headers
- sizes of parts of email according to content-type
- (optional) CPG of headers and parts

- SP
- SPCPG11

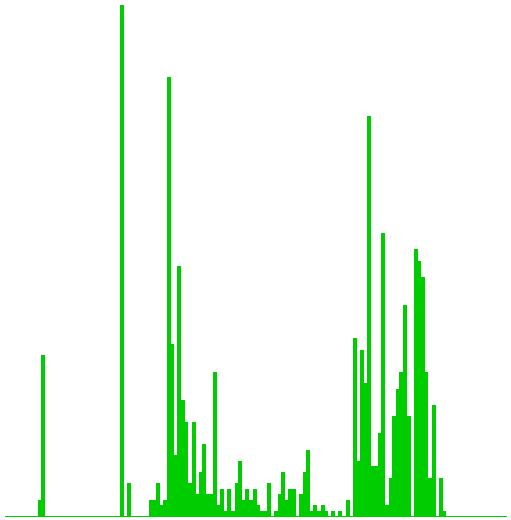
Graphical representation of line and character profile



```
From the_insider@postmaster.co.uk  Tue Apr 17 06:44:26 2007
Return-Path: <the_insider@postmaster.co.uk>
Received: from cosmic200 (windows.globalgold.co.uk [194.1.150.45])
 by speedy.uwaterloo.ca (8.12.8/8.12.5) with ESMTP id l3HAiP0I026448
 for <ktwarwic@speedy.uwaterloo.ca>; Tue, 17 Apr 2007 06:44:26 -0400
Received: from mail pickup service by cosmic200 with Microsoft SMTPSVC;
  Tue, 17 Apr 2007 11:44:10 +0100
From: "The Insider" <the_insider@postmaster.co.uk>
To: "Subscriber" <ktwarwic@speedy.uwaterloo.ca>
Subject: "The Insider" - News Bulletin
Date: Tue, 17 Apr 2007 11:44:10 +0100
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.3790.3959
Message-ID: <COSMIC200uYDlrjbudz00002c2e@cosmic200>
X-OriginalArrivalTime: 17 Apr 2007 10:44:10.0734 (UTC)
Status: O
Content-Length: 336
Lines: 10

*** BREAKING NEWS ***

American gunman massacres students and staff at American university
http://www.theinsider.org/news/article.asp?id=2476



---------------------------------------------------------------------
To be removed from this mailing list please use the form provided:
http://www.theinsider.org/news/emails/unsubscribe/
```

(a) Email

(b) Line profile



(c) Character profile

Results

Test corpuses

TREC 2007 corpus of 75 419 emails (spam 66.6%)

- train: 50 000 (68.3%)
- test: 25 419 (63.1%)

CEAS 2008 corpus of 137 705 emails (spam 80.3%)

- train: 90 000 (81.2%)
- test: 47 705 (77.9%)

Performance measures and classification algorithm

Performance measures

- false negative rate *fnr* (the ratio of misclassified spam) at fixed low values of false positive rate *fpr* (the ratio of misclassified ham)
- the receiver operating characteristic (*ROC*) curve, i.e. the graph of the true positive rate vs. the false positive rate, obtained as functions of the decision threshold

Classification algorithm

- Random Forest classifier

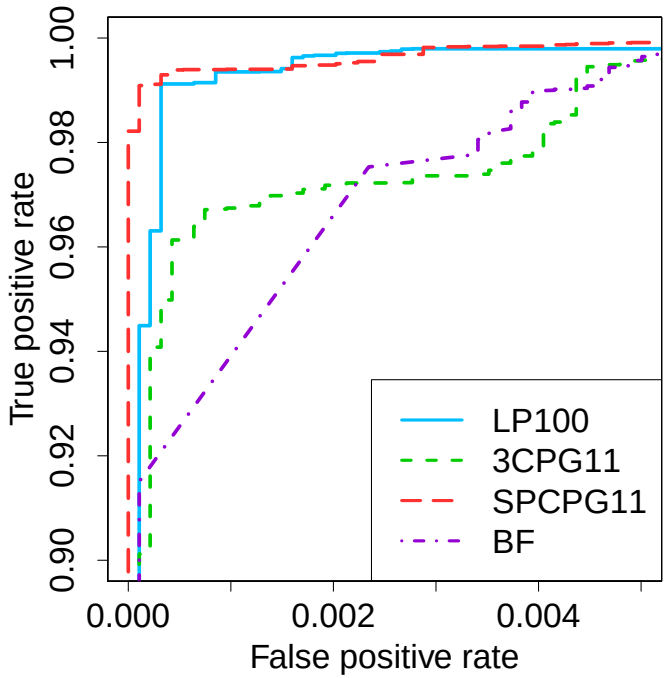
Performance of quantitative profiles

fnr (%) at fixed $fpr = 0.1\%$

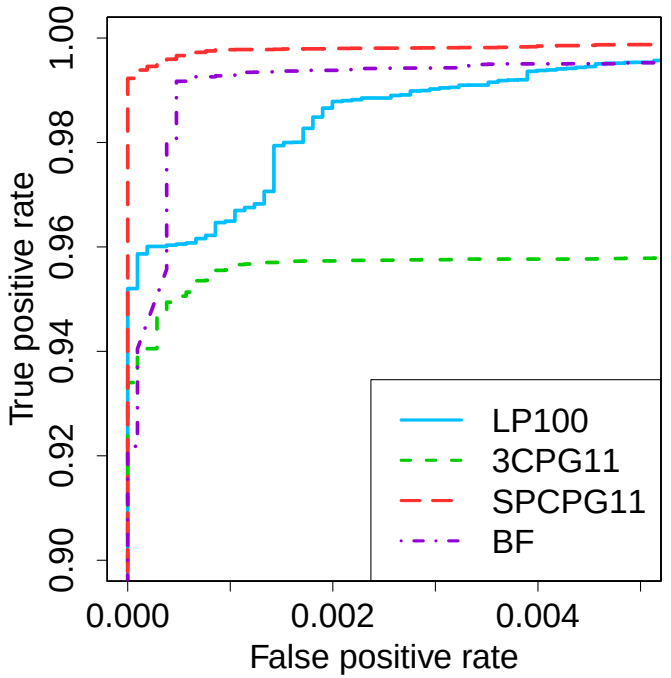
filter	TREC 2007	CEAS 2008
LP	0.65	3.46
WP	0.52	8.89
BRP	6.22	4.88
CP	14.61	4.98
3CPG11	3.26	4.42
MWPCPG11	17.26	5.45
SP	4.33	0.51
SPCPG11	0.60	0.22
SpamAssassin-RF	66.06	92.23
Bogofilter	7.98	0.71

Performance of quantitative profiles

ROC curves



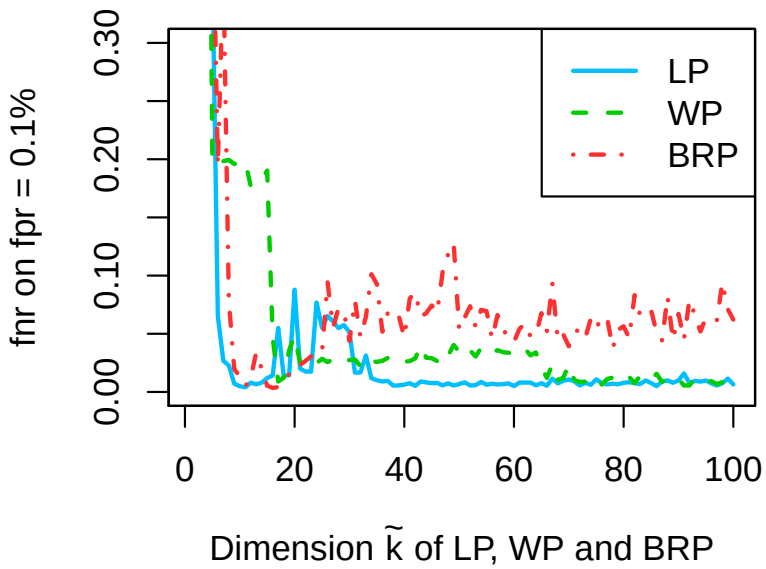
(a) TREC 2007



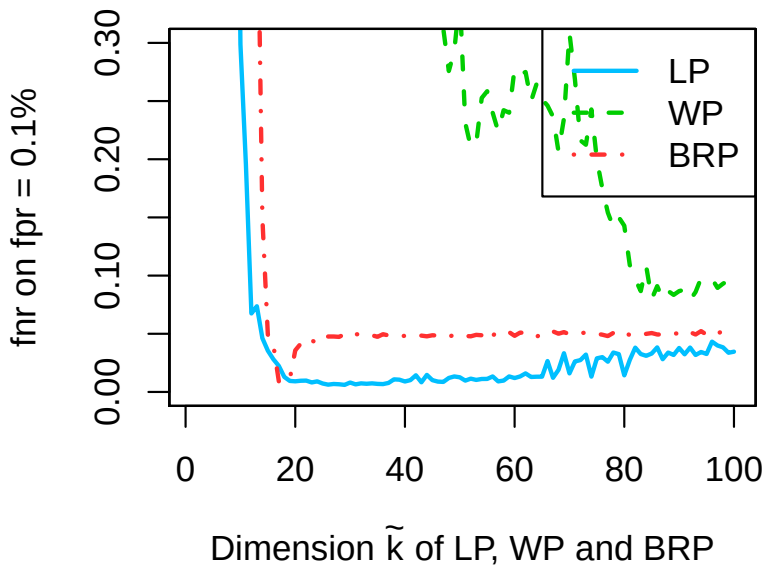
(b) CEAS 2008

Dimension of binary profiles

Dependence of BPs performance on its dimension



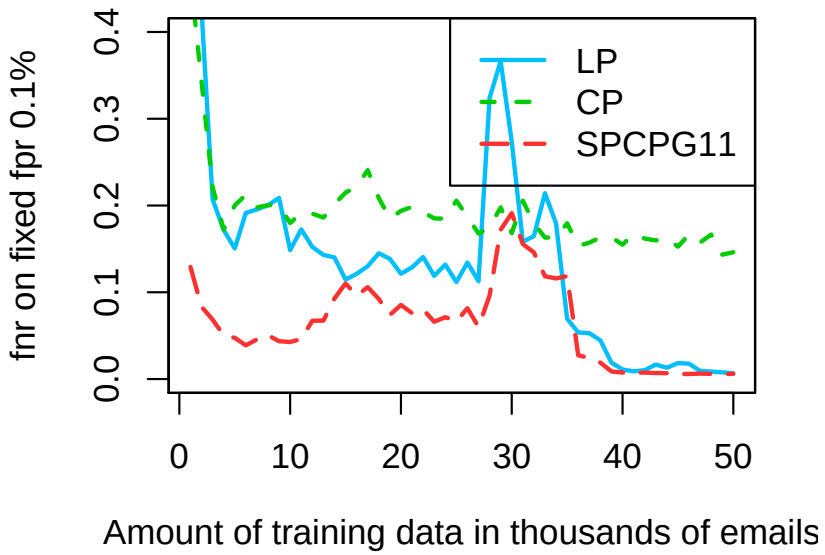
(a) TREC 2007



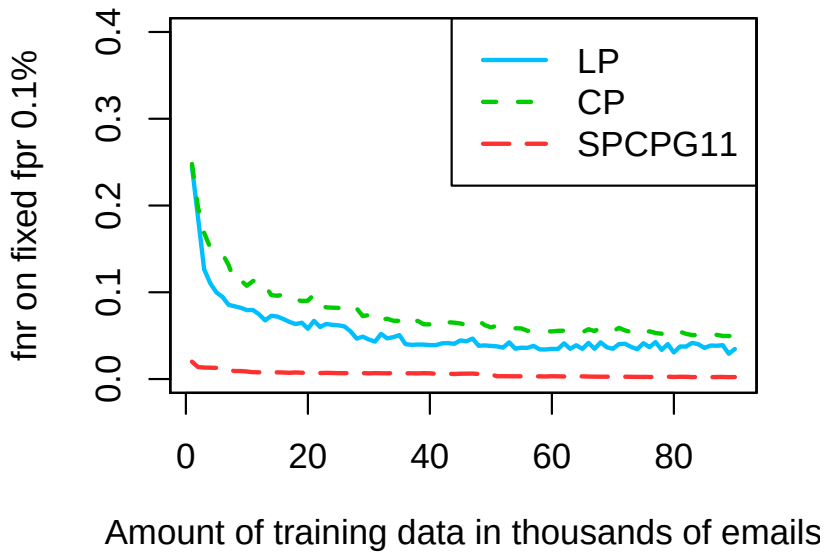
(b) CEAS 2008

Learning curves

Dependence of QPs performance on the size of training set



(a) TREC 2007



(b) CEAS 2008

Conclusions

Conclusions

- quantitative profiles based Random Forest classifiers attain very good performance, at least comparable or better to that of Bogofilter and much better than optimized SpamAssassin
- the resulting filters are:
  - highly scalable
  - easy to parallelize (thanks to RF)
  - independent of language
  - easy to combine with other filters (thanks to QPs)

References


Breiman, L. (2001) *Random forests*. Machine Learning, 45, 5-32.


Grendár, M., Škutová, J. and Špitalský, V. (2011) *Spam filtering by quantitative profiles*. Appear in IJCSI Volume 9, Issue 5, September 2012. `http://arxiv.org/pdf/1201.0040v1.pdf`


Grendár, M., Škutová, J. and Špitalský, V. (2012) *Email categorization and spam filtering by random forest with new classes of quantitative profiles*. Proceedings of COMPSTAT 2012, 283–294.


R Development Core Team (2010) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org/`


Sroufe, P., Phithakkitnukoon, S., Dantu, R. and Cangussu, J. (2010) *Email shape analysis*. In *Distributed Computing and Networking*, Lecture Notes in Computer Science, K. Kant et al. (eds), 5935/2010, pp. 18-29.

Thank you for your attention.