Standardized Spam Filter Evaluation

Gordon V. Cormack
<gvcormac@uwaterloo.ca>
21 January 2005
Why Standardized Evaluation?

To answer questions!

Is spam filtering a viable approach?
What are the risks, costs, and benefits of filter use?
Which spam filter should I use?
How can I make a better spam filter?

What's the alternative?

Testimonials
Uncontrolled, unrepeateable, statistically bogus tests
Warm, fuzzy feelings
There's no Perfect Test

But a standardized test should

Model real filter usage as closely as possible

Evaluate the filter on criteria that reflect its effectiveness for its intended purpose

Eliminate uncontrolled differences

Be repeatable

Yield statistically meaningful results

Future tests will

Challenge assumptions in the current test
Sponsored by, held at

NIST – National Institute for Standards & Technology

http://trec.nist.gov

Goals

To increase the availability of appropriate evaluation techniques for use by industry and academia, including the deployment of new evaluation techniques more applicable to current systems.

Format

Participants do experiments in one or more tracks
February 21, 2005 (*now!*)
  
  Answer call for participation (SPAM track)

Winter/Spring, 2005
  
  Mailing list discussion, finalize experimental setup

Summer 2005
  
  Prepare and submit filters for evaluation

November 15-18, 2005
  
  TREC Conference, results presented, methods discussed
Spam Filter Usage

Filter Classifies Email

Human addressee

Triage on ham File

Reads ham

Occasionally searches for misclassified ham

Report misclassified email to filter
Simulating Email Stream

Identify user

Secure user's permission (tacit or explicit)

this is the hard part

User's sensitivities

Sender's sensitivities

3rd Parties sensitivities

Privacy legislation & ethics

Capture email exactly as delivered
Spam Filter Evaluation

Simulate (replay) incoming email stream

- single recipient (for now)
- chronological order
- full email message with original headers

Simulate idealized user's behaviour

- reports all misclassifications immediately
  - spam in ham file (spam misclassification, false negative)
  - ham in spam file (ham misclassification, false positive)

Capture filter results

Analyze captured results
Simulating Idealized User

Capture

Filter result for each message (ham/spam)

User's reports of misclassified ham or spam

But Real Users are not Ideal

err and are inconsistent

slow and haphazard in reporting misclassification

Real User involved in pilot evaluation

vets disagreements between user and filter

Gold Standard ham/spam judgement
Filter must implement exactly 3 commands

initialize

All steps necessary to install the software on a clean system and to prepare to classify a user's email.

classify filename

read filename which contains exactly 1 email message
write one line of output:

classification score auxiliary_file

train judgement filename classification auxiliary_file

take note of gold-standard judgement

filename, classification, auxiliary_file from prior classify
Input

User email stream, 1 message per file
Index file, 1 line per message, chronological order:

judgement filename genre

Filter, as 3 commands: initialize, classify, train

Output

Raw Result File, 1 line per message:

file=filename judge=judgement class=classification
genre=genre
initialize

for each judgement, filename, genre in index

classify filename > classification, score, auxiliary_file

train judgement filename classification auxiliary_file

output judgement, filename, classification, score, genre
## Analysis – Contingency Table

### Gold Standard Judgement

<table>
<thead>
<tr>
<th></th>
<th>Ham</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ham</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Spam</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

- **a**: ham (correctly classified) [true negative]
- **b**: spam misclassification [false negative]
- **c**: ham misclassification [false positive]
- **d**: spam (correctly classified) [true negative]

- \(c/(a+c)\): ham misclassification rate
- \(b/(b+d)\): spam misclassification rate
- \((c+d)/(a+b+c+d)\): overall misclassification rate
- \((a+d)/(a+b+c+d)\): accuracy (equivalent to overall misc. rate)
Most filters compute *spamminess*

if \(\text{spamminess} > \text{threshold}\) then classify as spam
else classify as ham

*threshold* value is arbitrary

higher threshold =
fewer ham misclassifications
more spam misclassifications

**ROC (Receiver Operating Characteristic) Curve**

vary \(\text{threshold}\), plot ham misc. vs. spam misc.

Area under curve approaches 1 (perfect filter)
## Some Numbers

### Private Corpus (Mr. X) vs. Public Corpus (Spamassassin)

<table>
<thead>
<tr>
<th>Filters</th>
<th>I-ROCA%</th>
<th>ham%</th>
<th>Spam%</th>
<th>misc%</th>
<th>I-ROCA%</th>
<th>ham%</th>
<th>Spam%</th>
<th>misc%</th>
</tr>
</thead>
<tbody>
<tr>
<td>bogo</td>
<td>0.04</td>
<td>0.07</td>
<td>6.48</td>
<td>5.30</td>
<td>0.19</td>
<td>0.19</td>
<td>23.93</td>
<td>7.61</td>
</tr>
<tr>
<td>spamassassin</td>
<td>0.06</td>
<td>0.06</td>
<td>5.88</td>
<td>4.81</td>
<td>0.18</td>
<td>0.70</td>
<td>6.10</td>
<td>2.39</td>
</tr>
<tr>
<td>spamprobe</td>
<td>0.10</td>
<td>0.41</td>
<td>0.85</td>
<td>0.77</td>
<td>0.28</td>
<td>0.80</td>
<td>3.13</td>
<td>1.52</td>
</tr>
<tr>
<td>crm</td>
<td>0.40</td>
<td>2.24</td>
<td>0.68</td>
<td>1.00</td>
<td>1.14</td>
<td>1.81</td>
<td>4.03</td>
<td>2.50</td>
</tr>
<tr>
<td>dspam</td>
<td>0.91</td>
<td>1.39</td>
<td>0.94</td>
<td>1.02</td>
<td>3.42</td>
<td>1.01</td>
<td>32.79</td>
<td>10.94</td>
</tr>
<tr>
<td>dbacl</td>
<td>2.41</td>
<td>0.65</td>
<td>17.31</td>
<td>14.25</td>
<td>2.95</td>
<td>1.71</td>
<td>11.35</td>
<td>4.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus Size</th>
<th>9038</th>
<th>40048</th>
<th>49086</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Standardized Spam Filter Evaluation, Gordon Cormack, spamconference.org, 21 January, 2005
Misclassification as a function of messages learned

Define $p = \text{Prob( misclassification after } n \text{ messages )}$

Logistic Regression

Assume

$$\log\left(\frac{p}{1-p}\right) = an + b$$

Compute $a, b$ that best fit data:

*maximum likelihood estimate*

Piecewise Estimate

Divide $1..n$ into 50 interval and compute $p$

Coalesce adjacent intervals so when $p = 0$
Not all types of ham are equal!

Some more likely misclassified
higher likelihood of ending up in spam filter
Some more likely missed if filtered
can be retrieved from spam file
Some more valuable
consequences of non-receipt vary dramatically

Overall downside risk depends on all these factors

Spam can similarly be classified
Subject: Booking Confirmation
From: "Destina.ca" <confirmation@destina.ca>
Date: Thu, 13 Jan 2005 22:18:07 +0000 (GMT)
To: gvcormac@uwaterloo.ca

***** PLEASE DO NOT REPLY TO THIS EMAIL *****

Your booking is confirmed. Thank you for choosing Destina.ca.
Please print this itinerary/receipt for your reference.

Main Contact Information

Name: Doctor Gordon V Cormack
Email: gvcormac@uwaterloo.ca
Phone 1: 1-905-627-2457
Credit Card#: xxxxx-xxxxx-xxxxx-5359

Booking Reference: KUDM2M

Customer Care
Call our Customer Care 1-866-871-4747

Air Canada Flight Info
1-888-247-2255

Electronic Ticketing confirmed.
This is your official itinerary/receipt.

Flight Itinerary

Flight From To Stops Duration Aircraft Fare Type

Toronto (YYZ) Boston (BOS) 0 1hr32 319 Tango

Boston (BOS) Toronto (YYZ) 0 1hr46 CRJ Tango

More Travel Options

Save on Hotels
Earn 200 Aeroplan Miles.

Save on Cars
Earn 100 Aeroplan Miles.

Add a Flight
Earn 1 mile for $3 spent within North America.

Add Travel Insurance
Choose the travel insurance that best suits your needs.
From: "T. E. Simos" <tsimos@mail.ariadne-t.gr>
Date: Thu, 13 Jan 2005 04:37:10 +0200
To: secretary@ieccs.net
CC: tsimos@mail.ariadne-t.gr

Dear Colleagues

This year we organise the International e-Conference on Computer Science 2005 (IeCCS 2005) from 12 to 17 May 2005.

Please circulate the following announcement, call for papers, sessions and minisymposia and leaflet to your colleagues.

If you want leaflets and posters for IeCCS 2005, please send your request to secretary@ieccs.net

Sincerely yours

Professor Dr. T.E. Simos
Chair and Organiser IeCCS 2005

--

Professor Dr. T.E. Simos
President of the European Society of Computational Methods in Sciences and Engineering (ESCMES)
Active Member of the European Academy of Sciences and Arts
Corresponding Member of the European Academy of Sciences
Corresponding Member of European Academy of Arts, Sciences and Humanities
Fellow of the Royal Society of Chemistry (FRSC)
URL: http://www.ucp.gr/~simos
Editor-in-Chief and Founder
Journal of Computational Methods in Sciences and Engineering (JCMSE)
IOS Press. URL: http://www.iospress.nl/html/14727976.html
Editor-in-Chief and Founder
Applied Numerical Analysis and Computational Mathematics (ANACM)
ISSN 1611-8170
Wiley-VCH. URL: http://www3.interscience.wiley.com/cgi-bin/jhome/106571062
Editor-in-Chief and Founder
Computing Letters (COLE)
ISSN 1974-0404
VSP/Brill Publishing Company. URL: http://www.vspub.com/journals/jn-ComLet.html
Series Editor: Lecture Series on Computer Science and Computational Science
ISSN 1573-4196
VSP/Brill Publishing Company
Official Address:
Department of Computer Science and Technology,
Faculty of Sciences and Technology,
University of Peloponnese, GR-221 00 Tripolis, GREECE.
Postal Address:
26 Menelaou Street, Amfithea - Paleon Faliron, GR-175 64 Athens, GREECE.
E-mail: tsimos@mail.ariadne-t.gr

----------------------------------------------------------
Ham Genre Examples

Personal email from a regular correspondent
Personal email from first-time correspondent
Advertising (acceptable to recipient)
Email delivery failure notices
Mailing lists (formal and informal)
News clipping services
Internet transactions
Spam Genre Examples

Advertising
Scams
Demographic-targetted
Backscatter
Virus
Further Resources

TREC - trec.nist.gov
- Call for participation
- Description of tracks
- Past proceedings

Spam Track – plg.uwaterloo.ca/~gvcormac/spam
- Preliminary guidelines
- Test jig, analysis tools, sample filters
  - Linux prototype only – will evolve

Methodology -
plg.uwaterloo.ca/~gvcormac/spamcormack