Predicting the Future
Finite State Machines
Testing, Probability, Statistics
and other unpleasant things

Gordon V. Cormack

University of Waterloo
The Task

Create a compressed version (self-extracting archive) of the 100MB file \texttt{enwiki8} of less than about 16MB. More precisely:

News: Alexander Ratushnyak is also the second Winner! Congratulations!

... the contest continues ...

Being able to compress well is closely related to intelligence as explained below. While intelligence is a slippery concept, file sizes are hard numbers. Wikipedia is an extensive snapshot of Human Knowledge. If you can compress the first 100MB of Wikipedia better than your predecessors, your (de)compressor likely has to be smart(er). The intention of this prize is to encourage development of intelligent compressors/programs.
Honeywell MK III CMU

DMC

Information theory
Automata theory
Markov processes
Probability & stats
Machine learning
250 lines of code
Evaluation & measurement
Application (avionic telemetry)

theory + practice + evaluation + application
Predict Human Actions

Cormack, *Spam and other unpleasant things*, October 2007
Thanks gcormac@uwaterloo.ca! On the last email you said **ham** and the spam filters said **ham**
So far you've found 0 possible filter errors and verified 8 emails

Click one of these buttons:

- This is Spam
- I'm not sure
- This is Ham
- Flag message as funny

As displayed by Microsoft Outlook

You have been identified as a current subscriber to our Reuter's Cobra or I-Finance Market Data. We are currently in the process of re-negotiating these licenses. Please respond no later than Tuesday, May 28th with your Co# and RC. Future ongoing services will be allocated to the appropriate cost centers.

This information will allow us to negotiate in an informed manner.
Individual costs will be determined at the time we have gathered all subscriber details.

I appreciate your assistance with this effort.

Paula Corey
Predictive models work for

data compression
spam detection
  viruses, phishing, IM, SMS, blog, Web spam
  insensitive to language, alphabet, coding method
  heterogeneous, multimedia, metadata
plagiarism detection, authorship attribution
intrusion detection
game playing

Need well defined tasks and evaluation!

Cormack, Spam and other unpleasant things, October 2007
Data Compression

Given a stream of bits
Represent the stream in fewer bits

Trick:
predict each bit in turn (as a probability p)
encode as \(-\log_2(p)\) bits (on average)

arithmetic coding
optimal given p

Measure success! Compress some data!
What is Spam?

Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient.

Depends on sender/receiver relationship

Not “whatever the user thinks is spam.”
Spam and non-spam examples

Spam

Hi,
=20
MeR/D/A
V/aGRA
PROz&C
Amox/ciIl/n
CiAL/S
VAL/uM
Tr&madoI
Amb/EN
X&nax
LeV/TRA
S0m&
=20
http://www.prosebutis.com

Non-spam

Dear Gord:

Your C program has solved Ok the problem 11102 (Moonshine)
in 0.514 seconds using as much as 420 kbytes of virtual memory.
Congratulations!

--

PS: Check the board at http://acm.uva.es/board/

The Online Judge (Linux acm.uva.es 2.4.18-27.7.x i686)
Judge software version 2.8 [http://acm.uva.es/problemset/]
Wed May 24 23:19:30 UTC 2006
Hi,

Me R / D / A
V / a G R A
P R O z & C
A m o x / c i I l / n
C i A L / S
V A L / u M
T r & m a d o I
A m B / E N
X & n a x
L e V / T R A
S o m &


Dear Gord:

Your C program has solved ok the problem 11102 (Moonshine) in 0.514 seconds using as much as 420 kbytes of virtual memory. Congratulations!

--

PS: Check the board at http://acm.uva.es/board/

The Online Judge (Linux acm.uva.es 2.4.18-27.7.x i686) Judge software version 2.8 [http://acm.uva.es/problemset/]
Wed May 24 23:19:30 UTC 2006
How is the coloring used?

Filter Classifies Email

Human addressee

Triage on ham File

Reads ham

Occasionally searches for misclassified ham

Report misclassified email to filter

Cormack, Spam and other unpleasant things, October 2007
Questions to answer

Method to color spam & non-spam (ham)?
How well does the method color?
How well is the overall purpose met?

Facilitating delivery of good email
“filtering spam” is just a means to the end
Given a sequence of bits, predict the next one ($x$)

1011011011011011011x

$x$ is probably 0

0101101110111101111x

$x$ is probably 1

How 'probably'?

$$\text{Prob}(x = 0 \text{ following } 10110110110110110111)$$

$$\text{Prob}(x = 1 \text{ following } 01011011101111011111)$$

Model

abstracts the string of bits; used to predict behavior
Count the number of zeros & the number of ones:

1011011011011011011x

zeros: 6  ones: 13

Use the proportion of ones to estimate

\[ \text{Prob}(x = 1) = \frac{13}{19} = 0.68 \]

Doesn't seem like such a good estimate

how can we validate it?

intuition
testimonial
faith
experiment
1\textsuperscript{st} order Markov model

Count the number of \textit{ones} and \textit{zeros} following a 0, and the number following a 1

\[
1011011011011011011011011\times
\]

following 0: \textit{zeros}: 0 \textit{ones}: 6

following 1: \textit{zeros}: 6 \textit{ones}: 6

Use the proportion of \textit{ones} following 1 to estimate

\[
\text{Prob}(x = 1 \text{ following } 1) = 6/12 = 0.5
\]

Still doesn't seem like such a good estimate

but better than 0\textsuperscript{th} order
Count the number of *ones* and *zeros* following 00, and 01, and 10, and 11.

101101101101101101101101101111x

following 00: *zeros*: 0  *ones*: 0

following 01: *zeros*: 0  *ones*: 6

following 10: *zeros*: 0  *ones*: 6

following 11: *zeros*: 5  *ones*: 0

Use the proportion of *ones* following 11 to estimate

\[ \text{Prob}(x = 1 \text{ following } 11) = \frac{0}{5} = 0 \]

Overconfident!  (Overfitted model)
This example implements a 1\textsuperscript{st} order Markov model

A means \textit{following} 0; B means \textit{following} 1

Outputs $f$ on edges are frequencies

\[
\text{Prob}(1 \text{ following } A) = \frac{4}{2 + 4} = 0.667
\]

$f$ incremented after each transition
State A, input 1, Prob 0.67
B visited 16 times previously
- 4 from A; 12 from elsewhere
B should be cloned because it is visited from distinct contexts several times

B cloned to create B'
f divided in 4:12 ratio in proportion to previous visits
f incremented as usual

Cormack, Spam and other unpleasant things, October 2007
Predict each bit in turn

DMC

Construct optimal code

arithmetic coding

the more probable the shorter the representation

-\log_2 \text{prob}

but how do you do a fraction of a bit?

many bits at a time

Google for dmc.c
Likelihood Ratio

Likelihood of a bit (say 0) in spam

10110110110110110110 0 \quad \text{Prob}(x = 0)

Likelihood of same bit in non-spam

0101101110111101111 0 \quad \text{Prob}(x = 0)

log-likelihood ratio

\text{spamminess} = \log(\text{Prob}(x = 0) / \text{Prob}(x = 0))

coloring method

spam if \text{spamminess} > 0; otherwise non-spam

more generally

spam if \text{spamminess} > t; otherwise non-spam
Combining Likelihoods

\[ \text{spamminess}(x_1 x_2 x_3 \ldots x_n) \]

\[ = \log\left( \frac{\text{Prob}(x_1 x_2 x_3 \ldots x_n)}{\text{Prob}(x_1 x_2 x_3 \ldots x_n)} \right) \]

\[ = \log\left( \frac{\text{Prob}(x_1)}{\text{Prob}(x_1)} \right) + \]
\[ \log\left( \frac{\text{Prob}(x_2)}{\text{Prob}(x_2)} \right) + \]
\[ \log\left( \frac{\text{Prob}(x_3)}{\text{Prob}(x_3)} \right) + \ldots \]

\[ \ldots + \log\left( \frac{\text{Prob}(x_n)}{\text{Prob}(x_n)} \right) \]
Email spamminess

Let

$S$ be a string consisting of all known spam
$N$ be a string consisting of all known non-spam
$E$ be an email message

Define spamminess

$log \left( \frac{\text{Prob}(E \text{ following } S)}{\text{Prob}(E \text{ following } N)} \right)$
Measuring success

Collect email stream

adjudicate as spam or ham

gold standard

Filter email to

spam file if spamminess > t

ham file otherwise

Idealized user

reports errors immediately

Measure

false positive rate

spam misclassification rate
Receiver Operating Characteristic Curve

Cormack, *Spam and other unpleasant things*, October 2007
Summary statistics

<table>
<thead>
<tr>
<th>Filter</th>
<th>1-AUC (%)</th>
<th>SMR at 1% FP</th>
<th>SMR at 0.1% FP</th>
<th>SMR at 0.01% FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMC</td>
<td>†0.013 (0.010 – 0.018)</td>
<td>0.22%</td>
<td>1.17%</td>
<td>14.47%</td>
</tr>
<tr>
<td>PPM</td>
<td>†0.019 (0.015 – 0.023)</td>
<td>0.36%</td>
<td>1.78%</td>
<td>9.89%</td>
</tr>
<tr>
<td>dbacl&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.037 (0.031 – 0.045)</td>
<td>0.45%</td>
<td>5.19%</td>
<td>19.77%</td>
</tr>
<tr>
<td>Bogofilter&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.048 (0.038 – 0.062)</td>
<td>0.33%</td>
<td>3.41%</td>
<td>10.39%</td>
</tr>
<tr>
<td>SpamAssassin&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.059 (0.044 – 0.081)</td>
<td>0.37%</td>
<td>2.56%</td>
<td>7.81%</td>
</tr>
<tr>
<td>SpamProbe</td>
<td>0.059 (0.049 – 0.071)</td>
<td>0.65%</td>
<td>2.77%</td>
<td>15.30%</td>
</tr>
<tr>
<td>CRM114&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.122 (0.102 – 0.145)</td>
<td>0.68%</td>
<td>4.52%</td>
<td>17.17%</td>
</tr>
<tr>
<td>SpamBayes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.164 (0.142 – 0.189)</td>
<td>1.63%</td>
<td>6.92%</td>
<td>12.55%</td>
</tr>
</tbody>
</table>

† improves on best TREC result (p < .05)

<sup>b</sup> TREC 2005 result
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*

Standardized evaluation of well-defined tasks
Hello,

I was going through the pictures on this web site and ran across yours. I found you to be attractive, and your profile interesting. I sent along mine, take a look and see if there is any interest. I hope there is and I hear from you later.

Paul

Click the following link to view the sender's profile:

The following is my profile text:

Hello, I'm a down to earth person that likes to laugh and have a good time. I work in downtown Denver as a commodities trader for a large energy trading house. I do enjoy my career but understand life needs to be kept in prospective and be balanced. I enjoy activities over sitting and watching, but I have been known to sit in front of the TV and watch football or college basketball all day. I'm looking for someone who enjoys life and doesn't take themselves too seriously.

*******************************
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
More information?  Google!

cormack spam
TREC spam
DMC spam
DMC compression
ECML challenge
ROC curve
Markov model
PPM spam
OSBF Lua
Bogofilter

spamorham.org
spam conference
e-mail anti-spam
likelihood ratio
machine learning
text classifier

Cormack, Spam and other unpleasant things,  October 2007
Prediction by Partial Matching (PPM)

For each class:

- left context occurrences
- left context + prediction
- log-likelihood estimate
- compressed length

**Smoothing/backoff:**
- zero occurrence problem

**Adaptation:**
- increment counts
  - assuming in-class

---

**Context** (509 spam, 1 ham)

- `ai.stanford.?`

**Prediction** (0 spam, 1 ham)

- `ai.stanford.e`

**Prediction** (509 spam, 0 ham)

- `ai.stanford.E`