Bayesian spam filtering

Keith Briggs

Keith.Briggs@bt.com

more.btexact.com/people/briggsk2

Pervasive ICT seminar 2004 Apr 29 1500

bayes-2004apr29.tex typeset 2004 May 6 10:48 in pdfLaTeX on a linux system

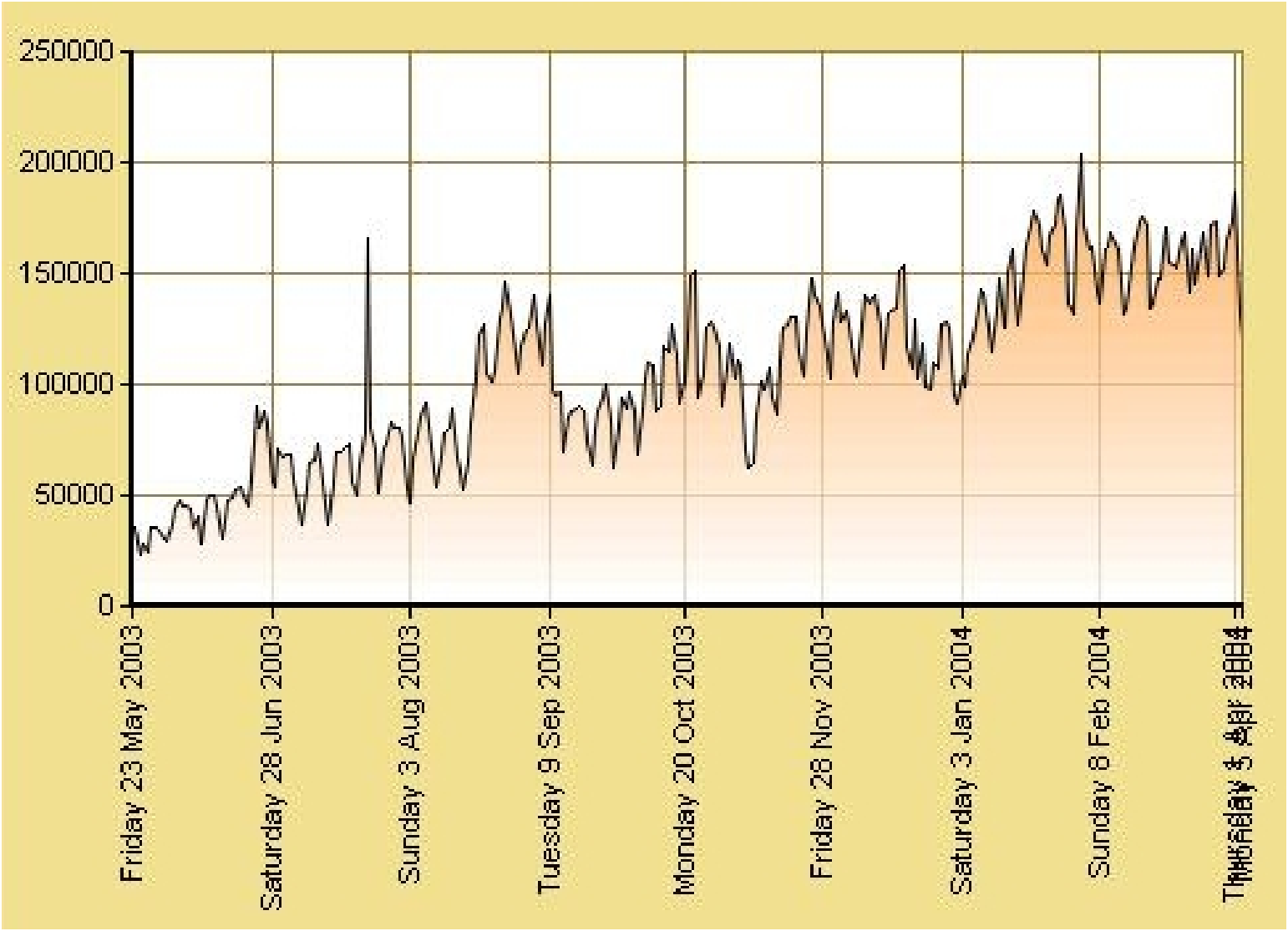
Outline

- the problem
- some 'solutions'
- probability theory
- Bayesian ideas
- . . . the real solution!

The aim:

to understand why Bayesian methods work so well

The problem



Total `bt.com` spam blocked per day

RFC 706

Network Working Group Jon Postel (SRI-ARC)
Request for Comments: 706
NIC #33861

On the Junk Mail Problem

In the ARPA Network Host/IMP interface protocol there is no mechanism for the host to selectively refuse messages. This means that a Host which desires to receive some particular messages must read all messages addressed to it. Such a host could be sent many messages by a malfunctioning host. This would constitute a denial of service to the normal users of this host. Both the local users and the network communication could suffer. The services denied are the processor time consumed in examining the undesired messages and rejecting them, and the loss of network thruput or increased delay due to the unnecessary busyness of the network.

Request for Comments: 706 Nov 1975

Solutions?

Method	version	pro	con
legal	CAN-SPAN		just politics
legal	opt-in		you're kidding!
legal	opt-out		"
blacklist	spamhaus etc.		centralized
whitelist			new emailers must register
financial penalty	micropayment		who administers?
computational penalty			new protocol
tripoli [5]		?	new protocol
challenge-response	e.g. about.mailblocks.com		lists
			attacks
filter	keyword		it doesn't work
filter	probabilistic	it works!	
honeypot	beat them at		admin load
	their own game		
commercial solutions	?	?	?

Prediction

- 00000000000000000000000000000000000000000000
- 01010101010101010101010101010101010101010101
- HHHTHHHHHHTHHHTTHHHHHHHTHHHHHHHHTHHHHHHHHHHH
- GAATTCTAGGCTTTCTTTGAAGAGGTAGTAATCTGTAGCCC
- BT is Britain's greatest company

Probability vs. degree of belief

- $P(\text{event}) \equiv \lim_{n\to\infty} \frac{\#\text{events}}{n}$
  - *objective*
  - *must be able to repeat the experiment indefinitely*
  - *rate of convergence of limit unspecified*
  - *strictly speaking, this rules out using this definition in the real world*

- 'degree of belief' B is more or less subjective
  - *meaningful for a single, non-repeatable event*
  - *your B might not be the same as my B*
  - *chance of rain tomorrow*
  - *chance of horse winning a race*
  - *spamminess of an email*

- Cox's axioms are a set of reasonable assumptions under which there exists a function mapping beliefs to probabilities

Probability theory

★ conditional probability

$$P(x=a|y=b) \equiv \frac{P(x=a,y=b)}{P(y=b)}$$

★ product rule

$$P(x,y|\mathcal{H}) = P(x|y,\mathcal{H})P(y|\mathcal{H}) = P(y|x,\mathcal{H})P(x|\mathcal{H})$$

★ marginalization

$$\begin{aligned} P(x|\mathcal{H}) &= \sum_y P(x,y|\mathcal{H}) \\ &= \sum_y P(x|y,\mathcal{H})P(y|\mathcal{H}) \end{aligned}$$

Bayes' theorem

★ Bayes' theorem - is just the product rule:

$$P(y|x,\mathcal{H}) = \frac{P(x|y,\mathcal{H})P(y|\mathcal{H})}{P(x|\mathcal{H})}$$

★ . . . with y interpreted as the data D, and x interpreted as parameters θ:

$$P(\theta|D,\mathcal{H}) = \frac{P(D|\theta,\mathcal{H})P(\theta|\mathcal{H})}{P(D|\mathcal{H})}$$

Prior, likelihood and posterior

- we can think of Bayes' rule as:

$$\boxed{\text{posterior} \propto \text{likelihood*prior}}$$

- for example, a single bit s sent twice over a noisy channel, received as $r_1 r_2$:
  - $P(s=1\,|\,r_1r_2) = P(r_1r_2\,|\,s=1)P(s=1)/P(r_1r_2)$
  - $P(s=0\,|\,r_1r_2) = P(r_1r_2\,|\,s=0)P(s=0)/P(r_1r_2)$
- that is, your prior (degree of belief before you observed that data r), is updated by the information the data provides about the value of s (the likelihood), to provide your posterior degree of belief

Thomas Bayes (1702-1761)

I now send you an essay which I have found among the papers of our deceased friend Mr Bayes, and which, in my opinion, has great merit. . . In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times. — Richard Price.

€~~€~~ biased?

★ Guardian Friday 2002 January 04:

> When spun on edge 250 times, a Belgian €1 coin came up heads 140 times and tails 110. '*It looks very suspicious to me*', said Barry Blight, a statistics lecturer at the LSE '*If the coin were unbiased the chance of getting a result as extreme as that would be less than 7%*'

★ we compare the models $\mathcal{H}_0$ (the coin is fair) and $\mathcal{H}_1$ (the coin is biased), with uniform prior $P(p|\mathcal{H}_1)=1$

★ the likelihood ratio is:

$$\frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_0)}=\frac{\frac{140!110!}{251!}}{1/2^{250}}\approx 0.48$$

★ thus the data give weak evidence (2.08/1) in favour of $\mathcal{H}_0$!

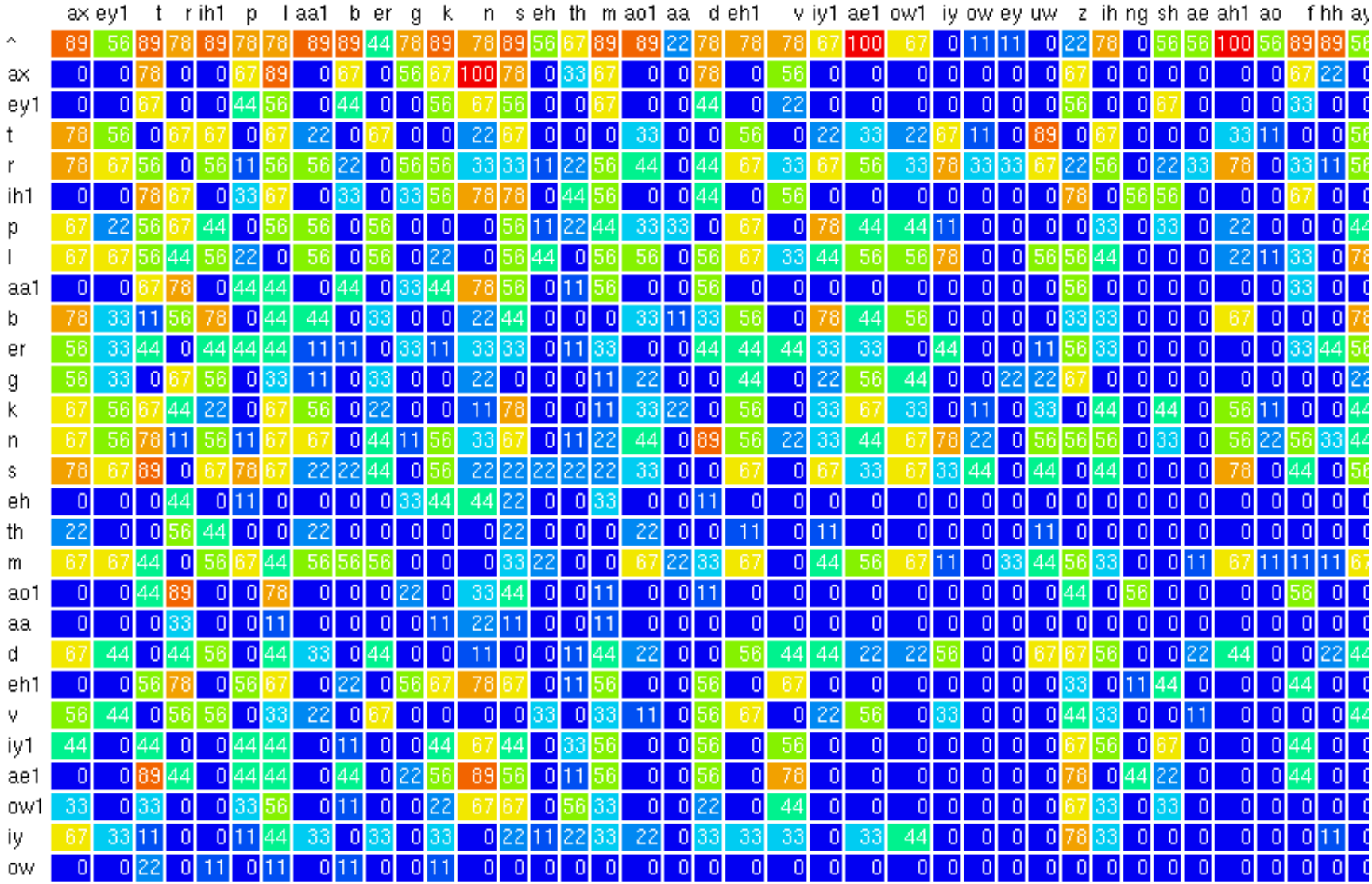
Text classification theory

- could be based on various choices of *features*: words, or n-grams
- corpora $C_1, C_2, \ldots, C_k$
- priors $\pi_1, \pi_2, \ldots, \pi_k$
- models $\mathbb{P}_{C_1}, \mathbb{P}_{C_2}, \ldots, \mathbb{P}_{C_k}$
- if x is an unknown document, the posterior probability that x belongs to C_j is $P(C_j|x) \propto \mathbb{P}_{C_j}\, \pi_j$
- decision rule: choose j to maximize $P(C_j|x)$
- other uses
  - *sorting emails - work or personal*
  - *forwarding emails in French or German etc. to the right person*

Digram measure

- word $w = w_1 w_2 \ldots w_k$
- reference measure $R_C(w) \equiv p_C(\wedge, w_1) p_C(w_1, w_2) \ldots p_C(w_k, \$)$
  - *this is naïve - it assumes adjacent digrams are statistically independent*
- Dirichlet digram measure $p_C(u, v) = \frac{\#(v|u) + \alpha\mu(v)}{\sum_r \#(r|u) + \alpha}$
- α is a hyperparameter, and the optimum α should be chosen from tests on various corpora
- for spam we use two corpora, spam and ham
- the latest softwares claim 99.9% accuracy
- false positives are much worse than false negatives, so we can adjust our algorithm to account for this

Example digram measure



Example using `dbacl` [1]

```
# learn...

kbriggs:~/Bayes> dbacl -l ham  ham/*
kbriggs:~/Bayes> dbacl -l spam spam/*

# test...

kbriggs:~/Bayes> echo "meet for lunch?" | dbacl -v -c ham -c spam
ham

kbriggs:~/Bayes> echo "viagra" | dbacl -v -c ham -c spam
spam

kbriggs:~/Bayes> echo "mortgage" | dbacl -v -c ham -c spam
spam
```

Can they beat us?

- the old trick of deliber@te spel1ing erors and füñny diàçrítics now works in our favour!
- dictionary salad: lots of random words
  - fails because the spammer doesn't know our model
- it's a long story: some genuine text - a possible danger
- habeas haiku: copyright poem, attempt at legal protection - now a strong spam indicator

Honesty in signalling - some philosophy

- spammers cannot violate the rules of any *machine* protocol like SMTP. . .
- but they violate the rules of human netiquette all the time. . .
- so perhaps we have to write some software which emulates human behaviour to detect the breach
- normal human communication has evolved mechanisms to detect sincerity: tone of voice, body language. . .
- and in particular, complex rules of grammar which although they impose a cost on the speaker, the consequence is that the listener *knows* the speaker is making an effort, and is thus worth listening to
  - *does this imply that any solution to the spam problem (or any other network security problem) must involve some similar mechanism?*
- in any case, I now understand why correct spelling and grammar are so important in written documents!

References

[1] L Breyer *dbacl - a digramic Bayesian classifier* `http://dbacl.sourceforge.net/`

[2] W Yerazunis *The spam-filtering accuracy plateau at 99.9% and how to get past it* `http://crm114.sourceforge.net/Plateau_Paper.pdf`

[3] D J C MacKay & L C Peto *A hierarchical Dirichlet language model* `http://www.inference.phy.cam.ac.uk/mackay/BayesLang.html`

[4] P Gburzynski & J Maitan *Fighting the spam wars: a remailer approach with restrictive aliasing* ACM Trans. Internet Tech. **4** 1-30 (2004)

[5] Tripoli `http://www.pfir.org/tripoli-overview`

[6] J D M Rennie & T Jaakkola *Automatic feature induction for text classification* MIT AI Lab, September 2002