Bayesian Filtering Example

Bayes' Formula

Thomas Bayes was born in 1702 in London, the son of a minister. After being educated privately, he was ordained a minister like his father and was assigned to a chapel in Tunbridge Wells, 35 miles outside of London. After Bayes' death in 1761, his friend Richard Price discovered his theory of probability in his papers. The theory was published by the Royal Society in 1764.

In basic terms, Bayes' Formula allows us to determine the probability of an event occurring based on the probabilities of two or more independent evidentiary events. Mathematically, the general formula is represented as:

$$P\left(E_j \middle| F\right) = \frac{P\left(F \middle| E_j\right) P\left(E_j\right)}{\sum P(F|E_i)\, P(E_i)}$$

Assuming that the variables *a* and *b* are the probabilities of two evidentiary events, the probability would be equal to:

$$\frac{\mathrm{ab}}{\mathrm{ab} + (1 - \mathrm{a})(1 - \mathrm{b})}$$

For three evidentiary events *a*, *b*, and *c*, the formula expands so the probability is equal to:

$$\frac{\mathrm{abc}}{\mathrm{abc} + (1 - \mathrm{a})(1 - \mathrm{b})(1 - \mathrm{c})}$$

In this fashion, the formula can be expanded to accommodate any number of evidentiary events.

This document introduces Bayes' Formula and provides an in-depth example of how a Bayesian filter can be used to classify spam e-mail messages. A more general overview of Bayesian filtering is contained in the *Introduction to Bayesian Filtering* whitepaper, available from Process Software's website at http://www.process.com/.

A Simple Example

Suppose that CheapSkies Airlines flights between Boston and New York City are delayed 75% of the time if it's raining. Also suppose that if a flight is scheduled to leave Boston before noon, it's only delayed 10% percent of the time (rain or shine). If you take a CheapSkies flight from Boston to New York City on a rainy day, and the flight is scheduled to depart before noon, what are the odds your flight will be delayed?

Since there are only two pieces of evidence to consider (the weather conditions and the scheduled departure time), we can use the basic form of Bayes' Formula to solve this problem. The probability that the flight will be delayed on a rainy day (75%, or 0.75) is represented by the variable *a*, and the probability that the flight will be delayed if it's scheduled to leave before noon (10%, or 0.10) is represented by the variable *b*.

Filling in Bayes' Formula from above, we see that the probability is equal to:

$$\frac{(0.75)(0.10)}{(0.75)(0.10) + (1 - 0.75)(1 - 0.10)}$$

Solving this equation yields a probability of 0.25, or a 25% chance that your flight will be delayed.

An important observation from this example is that we're dealing with *independent* events – the probability of one event has no impact on the other event. In the case of our example, there's a 75% chance the flight will be delayed on a rainy day regardless of whether or not it's scheduled to leave before noon. The probability of 75% includes both cases where the flight leaves before noon, and cases where it doesn't. Likewise, the fact that there's a 10% chance of the flight being delayed if it leaves before noon takes into account all flights – not just ones that leave on rainy days.

Using this concept to filter spam messages is known as *naive Bayesian filtering*, because we don't take into account the relationships between the various words contained in email messages. While it may certainly be true that a message containing all three of the words "clinical", "trial", and "Viagra" is never spam, all the naive Bayesian filter knows is that the words "clinical" and "trial" occur mostly in non-spam messages while the word "Viagra" occurs mostly in spam messages.

Spam Filtering Example

In the real world, applications for Bayes' Formula are messier and more complicated than the contrived example in the previous section. Following is a complete example of an e-mail message being filtered by a Bayesian filter similar to the one included in Process Software's PreciseMail Anti-Spam Gateway.

For our example, we're going to use the following "Nigerian spam" message. Note that we're looking at the complete message – headers and all.

```
Received: from unknown (HELO incamail.com) (209.11.24.18)
  by venice.example.com with SMTP; 4 May 2003 14:15:35 -0000
Received: from [10.1.1.27] (HELO app2.incamail.com)
  by incamail.com (CommuniGate Pro SMTP 4.0.6)
  with ESMTP id 2217203; Sun, 04 May 2003 10:12:16 -0400
Message-ID: <6549662.1052057538895.JavaMail.tomcat@app2.incamail.com>
From: BUMA SARO WIWA <bsarowiwa@incamail.com>
To: bsarowiwa@incamail.com
Subject: URGENT ASSISTANCE PLEAse
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-Priority: 3
X-Suffix: INBOX
```

```
Date: Sun, 04 May 2003 10:12:16 -0400
Content-Length: 2388

   Princess Buma Saro-Wiwa
101 Younde avenue YD
2390 Cameroun.
bsarowiwa@incamail.com OR b_sarowiwa@yahoo.com.au

Dear Friend,

I got your contact from a directory in a library in one of our
international school in my country and my instinct tells me to write you
and i feel It will be a great pleasure to be in contact with someone like
you.
frist, let me introduce myself, my name is PrincessBuma Nene Saro Wiwa Ken.
I am 27 years old from a royal family of Ken sarowiwa Kings hence I bear
the tittle "PRINCESS" I am single and the only duagther of my parents.my
father was a royal king of OGONI a prominent community in Rivers state
Nigeria who was killed through hanging by the order of late Gen sani Abacha
because of his community inheritance which are ( crude oil) that the F.G.N
has taken possession of it.
We are only two, I and my younger brother KEN SARO WIWA[jnr],after one year
death of my father, my mother died of High Blood preasure (HBP).Meanwhile,
we inherited some fortune in form of cash which I will reveal to you when
we get your response.Our old family friends have been very dishonest with
us since the death of our parents, they have duped us of virtually all cash
in the banks with different stories and reason. As such we decided to cut
off relationship from people around us because we find out that they have
on motive to squander what is left. We had to leave Nigeria to stay in
neighbuoring cameroun republic with the assistance of our family lawyer in
Nigeria, we are here now for three years and would like to move out to
another continent.I am interested to enter into strong relation with you as
a friend and partner after i have gotten good information about you on
internet.To be frank, we need someone who is kind and sincere that will
assist us.
We are interested to invest and live in your country therefore, it will be
our pleasure if you can be of help to us by assisting us to handle the
investment and planing of our fortune we inherited, to enable us build a
new home for safekeeping of our lives.
Please let me receive your response urgently.My kindest compliments.

Yours Faithfully,
Princess B. Saro-Wiwa.
bsarowiwa@incamail.com OR b_sarowiwa@yahoo.com.au

------------------------------------------------------------
Tired of spam and email overload?
Get a FREE 6MB email account at http://www.incamail.com
```

The first thing a Bayesian filter must do is split the message into tokens and build a table of all the tokens it intends to use in the decision making process. For our sample message, the table would be:

10.1.1.27	209.11.24.18	abacha	about
account	after	all	and
another	app2.incamail.com	are	around
assist	assistance	assisting	avenue
banks	bear	because	been
bit	blood	brother	bsarowiwa
build	buma	cameroun	can
cash	charset	communigate	community
compliments	contact	content-length	content-type
continent.i	country	crude	cut
dear	death	decided	died
different	directory	dishonest	duagther
duped	email	enable	enter
esmtp	f.g.n	faithfully	family
father	feel	find	for
form	fortune	frank	free
friend	friends	frist	from
gen	get	good	got
gotten	great	had	handle
hanging	has	have	hbp
helo	help	hence	here
high	his	home	http
inbox	incamail.com	information	inheritance
inherited	instinct	interested	international
internet.to	into	introduce	invest
investment	jnr	ken	killed
kind	kindest	king	kings
late	lawyer	leave	left
let	library	like	live
lives	may	meanwhile	mime-version
mother	motive	move	myself
name	need	neighbuoring	nene
new	nigeria	now	off
ogoni	oil	old	one
only	order	our	out
overload	parents	parents.my	partner
people	plain	planing	please
pleasure	possession	preasure	princess
princessbuma	pro	prominent	reason
receive	received	relation	relationship
republic	response	response.our	reveal
rivers	royal	safekeeping	sani
saro	saro-wiwa	sarowiwa	school
since	sincere	single	smtp
some	someone	spam	squander
state	stay	stories	strong
subject	such	sun	taken

tells	text	that	the
therefore	they	three	through
tired	tittle	two	unknown
urgent	urgently.my	us-ascii	
venice.example.com			
very	virtually	was	what
when	which	who	will
with	wiwa	would	write
www.incamail.com	x-priority	x-suffix	yahoo.com.au
year	years	you	younde
younger	your	yours	

Once the Bayesian filter has the list of tokens in the message, it searches the spam and non-spam token databases for these tokens. These databases of tokens are created and updated whenever the Bayesian filter is "trained" on a new message.

If a token from the message is found in the databases, the Bayesian filter calculates the token's spamicity based on the following variables:

- The frequency of the token in spam messages that the filter has been trained on
- The frequency of the token in ham messages that the filter has been trained on
- The number of spam messages the filter has been trained on
- The number of ham messages the filter has been trained on

The algorithm used to calculate a token's spamicity from these pieces of information is as follows:

```
Ham probability = Token frequency in ham messages / Number of ham messages
trained on

Spam probability = Token frequency in spam messages / Number of spam messages
trained on

If either Ham probability or Spam probability are greater than 1.0, set them
equal to 1.0.

Spamicity = Spam probability / (Ham probability + Spam probability)
```

If a token has occurred less than 5 times total in both ham and spam messages, the token is assigned a default spamicity of 0.4. The following example and table use a set of sample token databases generated by live mail feed on a test system at Process Software. The Bayesian filter was trained on 19,977 spam messages and 5,141 ham messages.

An example of this algorithm, using the token "after" from the example spam message and frequency values in the above tables is:

```
Ham probability = 1184 / 5141 = 0.230305
Spam probability = 1134 / 19977 = 0.056765
Spamicity = 0.056765 / (0.056765 + 0.230305) = 0.197740
```

This tells us that there's only a 19.8% chance that a message containing the word "after" is a spam message.

Repeating this process for each of the tokens in our sample message, we get the following frequencies and spamicities:

Token	Spam Frequency	Ham Frequency	Spamicity
10.1.1.27	0	0	0.400000
209.11.24.18	0	0	0.400000
abacha	14	2	0.643038
about	3301	2578	0.247848
account	585	563	0.210984
after	1134	1184	0.197740
all	9767	3759	0.400717
and	32109	12353	0.500000
another	1305	784	0.299898
app2.incamail.com	0	0	0.400000
are	13555	6130	0.404241
around	433	480	0.188409
assist	256	46	0.588847
assistance	386	171	0.367453
assisting	6	4	0.278509
avenue	70	25	0.418797
banks	238	8	0.884474
bear	80	12	0.631763
because	5114	973	0.574936
been	3233	2036	0.290097
bit	4296	2292	0.325398
blood	383	53	0.650312
brother	171	171	0.403703
bsarowiwa	0	0	0.400000
build	3364	576	0.600475
buma	0	0	0.400000
cameroun	0	0	0.400000
can	8083	4568	0.312889
cash	1318	49	0.873771
charset	9300	3324	0.418608
communigate	16	61	0.063232
community	70	76	0.191612
compliments	58	58	0.788651
contact	1552	760	0.344489
content-length	0	0	0.400000
content-type	26907	5054	0.504267
continent.i	0	0	0.400000
country	316	62	0.567406

Token	Spam Frequency	Ham Frequency	Spamicity
crude	19	0	0.990000
cut	272	199	0.260218
dear	752	113	0.631350
death	118	37	0.450768
decided	205	107	0.330228
died	44	31	0.267542
different	593	704	0.178152
directory	57	401	0.035289
dishonest	0	0	0.400000
duagther	0	0	0.400000
duped	0	0	0.400000
email	13820	2097	0.629081
enable	65	97	0.147084
enter	753	139	0.582309
esmtp	7239	7152	0.265983
f.g.n	0	0	0.400000
faithfully	35	0	0.990000
family	3255	172	0.829646
father	75	38	0.336835
feel	2269	299	0.661350
find	2966	854	0.471956
for	29946	14355	0.500000
form	2721	258	0.730756
fortune	211	16	0.772404
frank	47	85	0.124571
free	13077	948	0.780215
friend	456	110	0.516164
friends	1215	181	0.633362
frist	0	0	0.400000
from	65251	18549	0.500000
gen	63	14	0.536620
get	10853	2876	0.492677
good	1426	1752	0.173185
got	946	998	0.196101
gotten	49	35	0.264860
great	1761	556	0.449061
had	1202	1709	0.153260
handle	201	103	0.334309
hanging	39	51	0.164434
has	3661	2693	0.259176
have	11235	7113	0.359958
hbp	0	0	0.400000
helo	1855	1473	0.244761
help	2364	1406	0.302014
hence	36	16	0.366699
high	2032	265	0.663674
his	815	712	0.227545
home	3510	650	0.581532
http	57485	4233	0.548432

Token	Spam Frequency	Ham Frequency	Spamicity
inbox	74	91	0.173055
incamail.com	0	0	0.400000
information	4197	1490	0.420252
inheritance	0	0	0.400000
inherited	0	5	0.010000
instinct	0	0	0.400000
interested	592	237	0.391291
international	1392	165	0.684648
internet.to	0	0	0.400000
into	1359	1268	0.216187
introduce	53	20	0.405458
invest	139	7	0.836338
investment	657	31	0.845059
jnr	0	0	0.400000
ken	0	0	0.400000
killed	10	25	0.093331
kind	130	266	0.111720
kindest	0	0	0.400000
king	210	117	0.315960
kings	8	24	0.079005
late	181	221	0.174078
lawyer	31	9	0.469894
leave	141	189	0.161066
left	9847	488	0.838522
let	1007	987	0.207959
library	242	274	0.185197
like	6794	2752	0.388500
live	667	166	0.508366
lives	106	47	0.367248
may	4255	2102	0.342510
meanwhile	3	13	0.056058
mime-version	17646	4370	0.509602
mother	76	45	0.302956
motive	0	0	0.400000
move	403	336	0.235861
myself	103	110	0.194178
name	10101	1624	0.615480
need	2714	1813	0.278103
neighbuoring	0	0	0.400000
nene	0	0	0.400000
new	9051	2191	0.515291
nigeria	132	2	0.944398
now	8920	2034	0.530203
off	3061	835	0.485437
ogoni	0	0	0.400000
oil	64	42	0.281685
old	949	731	0.250427
one	8722	2995	0.428388
only	4954	2298	0.356824

Token	Spam Frequency	Ham Frequency	Spamicity
order	4442	680	0.627015
our	16869	1634	0.726535
out	5565	2829	0.336092
overload	0	5	0.010000
parents	119	61	0.334237
parents.my	0	0	0.400000
partner	509	39	0.770574
people	1808	828	0.359768
plain	954	3206	0.071131
planing	0	0	0.400000
please	11780	2108	0.589846
pleasure	117	13	0.698442
possession	10	9	0.222359
preasure	0	0	0.400000
princess	0	0	0.400000
princessbuma	0	0	0.400000
pro	1388	102	0.777873
prominent	6	0	0.990000
reason	552	487	0.225823
receive	8509	348	0.862871
received	19967	10164	0.499875
relation	20	3	0.631763
relationship	133	69	0.331570
republic	34	16	0.353529
response	645	311	0.347992
response.our	0	0	0.400000
reveal	29	3	0.713276
rivers	0	0	0.400000
royal	168	16	0.729885
safekeeping	10	0	0.990000
sani	0	0	0.400000
saro	0	0	0.400000
saro-wiwa	0	0	0.400000
sarowiwa	0	0	0.400000
school	313	68	0.542239
since	299	854	0.082654
sincere	22	0	0.990000
single	229	372	0.136755
smtp	2374	1702	0.264140
some	1981	2262	0.183924
someone	728	517	0.265988
spam	1167	956	0.239049
squander	0	0	0.400000
state	929	467	0.338597
stay	453	201	0.367084
stories	112	44	0.395793
strong	10357	154	0.945377
subject	22169	10497	0.500000
such	1026	848	0.237435

Token	Spam Frequency	Ham Frequency	Spamicity
sun	2608	1611	0.294089
taken	382	122	0.446225
tells	11	29	0.088933
text	19009	4012	0.549410
that	10559	9075	0.345789
the	34475	16621	0.500000
therefore	117	122	0.197946
they	2319	2640	0.184376
three	607	245	0.389346
through	4241	758	0.590138
tired	227	128	0.313369
tittle	0	0	0.400000
two	775	940	0.175036
unknown	2667	695	0.496866
urgent	93	31	0.435678
urgently.my	0	0	0.400000
us-ascii	665	1891	0.082989
venice.example.com	0	0	0.400000
very	1173	980	0.235490
virtually	136	18	0.660371
was	3573	4367	0.173933
what	3050	3548	0.181150
when	2404	2614	0.191378
which	1200	2132	0.126521
who	2041	1183	0.307476
will	9749	4255	0.370922
with	39458	15761	0.500000
wiwa	0	0	0.400000
would	6023	3296	0.319851
write	903	329	0.413948
www.incamail.com	0	0	0.400000
x-priority	11524	852	0.776826
x-suffix	0	0	0.400000
yahoo.com.au	0	0	0.400000
year	1096	421	0.401182
years	1397	503	0.416820
you	40273	9606	0.500000
younde	0	0	0.400000
younger	250	4	0.941466
your	31926	4534	0.531370
yours	682	75	0.700611

Now that the filter has calculated the spamicity value for each token in the message, it needs to choose 15 tokens that will be plugged into the Bayesian formula to calculate the message's overall spamicity. Using a subset of the tokens in the message enhances the Bayesian filter's performance, especially when dealing with large messages.

Early implementations of Bayesian filters chose the 15 tokens that had the most extreme values (i.e. the 15 tokens whose value was furthest from the neutral value of 0.5). Spammers have started including words that they're fairly sure will have a low spamicity, such as "congresswoman" and "umbrella", in their messages in an attempt to circumvent this system. As a result, the Bayesian filter included in PreciseMail uses a sampling algorithm based on standards of deviation to choose the 15 tokens fed to the Bayesian formula.

For our sample message, the 15 tokens chosen by the Bayesian filter are:

Token	Spamicity
account	0.210984
after	0.197740
crude	0.990000
faithfully	0.990000
good	0.173185
inherited	0.010000
invest	0.836338
investment	0.845059
let	0.207959
overload	0.010000
prominent	0.990000
receive	0.862871
safekeeping	0.990000
sincere	0.990000
therefore	0.197946

Once the Bayesian filter has selected 15 tokens, it plugs their spamicity values into Bayes' formula, as shown below. (With 15 different values, this gets a little bit messy on paper.) For our sample message, the probability of the message being spam is:

$$\frac{\begin{array}{c}(0.210984)(0.197740)(0.990000)(0.990000)(0.173185)(0.010000)(0.836338)(0.845059)\\(0.207959)(0.010000)(0.990000)(0.862871)(0.990000)(0.990000)(0.197946)\end{array}}{\begin{array}{c}(0.210984)(0.197740)(0.990000)(0.990000)(0.173185)(0.010000)(0.836338)(0.845059)\\(0.207959)(0.010000)(0.990000)(0.862871)(0.990000)(0.990000)(0.197946)+\\(1-0.210984)(1-0.197740)(1-0.990000)(1-0.990000)(1-0.173185)\\(1-0.010000)(1-0.836338)(1-0.845059)(1-0.207959)(1-0.010000)\\(1-0.990000)(1-0.862871)(1-0.990000)(1-0.990000)(1-0.197946)\end{array}}$$

This equation simplifies to:

$$\frac{0.00000001724922088357441036105371521631 8}{0.0000000172493341952014463710 86}$$

Solving this equation yields a probability of 0.999993, or a 99.9993% chance that the message is spam. If this message was sent to an email server protected by PreciseMail, it would be quarantined, discarded, or tagged as spam based on the options chosen by the systems administrator.

About PreciseMail Anti-Spam Gateway

PreciseMail Anti-Spam Gateway is an enterprise software solution that eliminates spam, phishing and virus threats at the Internet gateway or mail server. It has a proven 98% spam detection accuracy rate out-of-the-box without filtering legitimate messages. PreciseMail Anti-Spam Gateway has a highly sophisticated filtering engine is based on a combination of proven heuristic, DNS blacklisting, and Bayesian artificial intelligence technologies, which automatically learn how to separate spam messages from legitimate email. As a result, PreciseMail Anti-Spam Gateway can determine whether email is spam instead of passively reacting to known spammers by creating rules that block them after a spam attack occurs.

About Process Software

Process Software has been a premier supplier of communications software solutions to mission critical environments for twenty years. We were early innovators of email software and anti-spam technology. Process Software has a proven track record of success with thousands of customers, including many Global 2000 and Fortune 1000 companies.

PROCESS™ SOFTWARE

U.S.A.: (800) 722-7770 • International: (508 879-6994 • Fax: (508) 879-0042
E-mail: info@process.com • Web: http://www.process.com/