e-mail

Betreuer: Prof. Dr. Christian Schindelhauer
Electronic Mail

- MUA - one of the most available application service
- Must provide, when remote destination temporary unreachable
- Uses independent addresses

local-part @ domain-name
Email Client

- Document Editor
- Address Book
- Permanent Storage
- Communications Module
Electronic Mail System

Figure 26.1 Conceptual components of an electronic mail system, Corner
EMS with Mail Forwarding

Figure 26.2 An extension of the mail system, Corner
Alias Expansion

- Replacement within a site

- Conflicting aliases
Possible Architecture

• no TCP/IP Connection needed
• Sufficient?

Eric
Barry
Irene

Company's Mail Server

Valid Format

- **RFC 2822** take place of the RFC 822
- Header
  - To:
  - From:
  - Reply-to:
  - Blank Line:
- Body
  - Not specified
Simple Mail Transfer Protocol (SMTP)

- Mail transfer from server to another server
- Communication in ASCII text
  - Abbreviated commands with 3-digit numbers
- Transport Layer Security (TLS) for encrypted session
SMPT Communication 2

S: 220 Beta.gov Simple Mail Transfer Service Ready
C: HELO Alpha.edu
S: 250 Beta.gov

C: MAIL FROM <Smith@Alpha.edu>
S: 250 OK

C: RCPT TO:<Jones@Beta.gov>
S: 250 OK

C: RCPT TO:<Green@Beta.gov>
S: 550 No such user here
C: RCPT TO:<Brown@Beta.gov>
S: 250 OK

C: DATA
S: 354 Start mail input; end with <CR><LF>.<CR><LF>
C: ...sends body of mail message...
C: ...continues for as many lines as message contains
C: <CR><LF>.<CR><LF>
S: 250 OK

C: QUIT
S: 221 Beta.gov Service closing transmission channel

Figure 26.3 [4, chap.26]
MX Record Email eXchanger

- DNS to decouple mail destination from the domain name assigned to machine than a ping request

- MX Record (Mail eXchanger)

```
e-mail.com IN MX 10 post.e-mail.com
```

DomainName

Priority

MailServer
E-mail Retrieval & Manipulation

• Post Office Protocol (POP)
  - POP3
  - POP3S

• Internet Message Access Protocol (IMAP)
  - IMAP4
  - allows Synchronisation
Multipurpose Internet Mail Extensions (MIME)

- Transmission of non-ASCII data through email
- 7-bit ASCII coding
- RFC 2822 Format
- MIME-Version
- Content-Type
- Content-Transfer-Encoding
- Base64 for sixty-four ASCII characters
Content Types

- **Text**
  - Textual document

- **Image**
  - Photograph or computer generated image

- **Audio**
  - Sound Recording

- **Video**
  - Video Recording with motion

- **Application**
  - Raw data for a program

- **Multipart**
  - Messages with separate content type and encoding

- **Message**
  - Forwarded an entire e-mail
MIME Multipart Messages

- Mixed
  - one message contain multiple, independent submessages with independent type and encoding

- Alternative
  - multiple representation of the same data

- Parallel
  - single message includes subparts to be viewed together (audio and video)

- Digest
  - single message contain a set of other messages
MIME Multipart Messages

From: bill@acollage.edu
To: john@example.com
MIME-Version: 1.0
Content-Type: Multipart/Mixed: Boundary=StartOfNextPart

—–StartOfNextPart
Content-Type: text/plain
Content-Transfer-Encoding: 7bit
John,
Here is the photo of our research lab that I promised to send you. You can see the equipment you donated.

Thanks again,
Bill

—–StartOfNextPart
Content-Type: image/gif
Content-Transfer-Encoding: base64
...data for the image...
Spam

- Unsolicited Bulk Mail (UBE)
- 94% of the e-mail
- 10 Billion € connection cost per year
- Waste of time and resources
- Decrease of trust in email communication
- 0.16% for lose of legitimate e-mail
Spam Filtering

• Cooperative
  - Content labeling
  - Recipient registration
  - Mail From

• Legal
  - Regulation
  - Contracts

• Heuristic
  - Origin filtering
  - Message filtering

Point of interest
Origin filtering

• Refusing IP connection from known UBE originators
• Refusing TCP connections from known UBE originators in the SMTP server
• Refusing SMTP messages from known UBE originators at the MAIL FROM command
• Refusing SMTP messages from originators whose domain name doesn't match their IP address
Message filtering

• In the message store
  - filter for particular key words, pattern, bag of words
  - with classifier
  - heuristic rules
  - risk of losing or mislabeling ham mail
  - delay

• At the mail client
Classification

- **tp** - ham correctly predicted
- **tn** - spam correctly predicted
- **fp** - spam misclassified as ham
- **fn** - ham misclassified as spam

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Category</th>
<th>ham(+)</th>
<th>spam(-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ham(+)</td>
<td></td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>spam(-)</td>
<td></td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

[5] Table 2 SpamAssassin
Statistical Approach: Naive Bayes

- Divide emails in spam and ham
- Assign probabilities to frequently occurring good and bad words
- Use words with probabilities far from 0.5
- Calculate probabilities of 15 interesting bad or good words
- New word with 0.4 of spam probability as neutral

Prior doesn’t have to be the same for different users

\[
Pr(\text{spam} | \text{words}) = \frac{Pr(\text{words} | \text{spam}) \cdot Pr(\text{spam})}{Pr(\text{words})}
\]
## Spam vs. Ham

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>madam</td>
<td>0.99</td>
<td>continuation</td>
<td>0.01</td>
</tr>
<tr>
<td>promotion</td>
<td>0.99</td>
<td>describe</td>
<td>0.01</td>
</tr>
<tr>
<td>republic</td>
<td>0.99</td>
<td>continuations</td>
<td>0.01</td>
</tr>
<tr>
<td>shortest</td>
<td>0.047225013</td>
<td>example</td>
<td>0.033600237</td>
</tr>
<tr>
<td>mandatory</td>
<td>0.047225013</td>
<td>programming</td>
<td>0.05214485</td>
</tr>
<tr>
<td>standardization</td>
<td>0.07347802</td>
<td>i'm</td>
<td>0.055427782</td>
</tr>
<tr>
<td>sorry</td>
<td>0.08221981</td>
<td>examples</td>
<td>0.07972858</td>
</tr>
<tr>
<td>supported</td>
<td>0.09019077</td>
<td>color</td>
<td>0.9189189</td>
</tr>
<tr>
<td>people's</td>
<td>0.09019077</td>
<td>localhost</td>
<td>0.09883721</td>
</tr>
<tr>
<td>enter</td>
<td>0.9075001</td>
<td>hi</td>
<td>0.116539136</td>
</tr>
<tr>
<td>quality</td>
<td>0.8921298</td>
<td>california</td>
<td>0.84421706</td>
</tr>
<tr>
<td>organization</td>
<td>0.12454646</td>
<td>same</td>
<td>0.15981844</td>
</tr>
<tr>
<td>investment</td>
<td>0.8568143</td>
<td>spot</td>
<td>0.1654587</td>
</tr>
<tr>
<td>very</td>
<td>0.14758544</td>
<td>us-ascii</td>
<td>0.16804294</td>
</tr>
<tr>
<td>valuable</td>
<td>0.82347786</td>
<td>what</td>
<td>0.19212411</td>
</tr>
</tbody>
</table>

SpamAssassin

• Open-source hybrid spam filter of:
  - bayesian learner
  - set of 500+ heuristics
  - each heuristic has a weight score
  - each rule represented as binary attribute
Total Cost Ratio (TCR)

- Sensitive personal messages = 1000
- Business related messages = 500
- E-commerce related message = 100
- Mailing lists / discussion forums = 50
- Promotional offers = 25
- Cost of misclassifying spam = 1

\[
TCR = \frac{fp + tn}{\sum_{x \in fn} C(x) + fp} < 1
\]
Error Rates

- Estimated probability of misclassification from ham to spam

\[ \text{ham}_{e} = \frac{fn}{fn + tp} \]

- Estimated probability of misclassification from spam to ham

\[ \text{spam}_{e} = \frac{fp}{fp + tn} \]
Test

- hx – 475 ham mailbox
- ix – 2163 ham mailbox
- ux – 363 ham mailbox
- * - no true ham was misclassified

\[ hame = \frac{fn}{fn+tp} \]

<table>
<thead>
<tr>
<th></th>
<th>as_train</th>
<th>as_test</th>
<th>hx</th>
<th>ix</th>
<th>ux</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>36.9%</td>
<td>58.5%</td>
<td>0.2%</td>
<td>0.3%</td>
<td>0.8%</td>
</tr>
<tr>
<td>SAnb</td>
<td>*7.79%</td>
<td>*8.6%</td>
<td>0.4%</td>
<td>0.1%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Log</td>
<td>0.65%</td>
<td>*1.7%</td>
<td>2.7%</td>
<td>1.8%</td>
<td>1.1%</td>
</tr>
<tr>
<td>SMO</td>
<td>0.68%</td>
<td>*1.2%</td>
<td>2.7%</td>
<td>0.9%</td>
<td>1.1%</td>
</tr>
<tr>
<td>MLR</td>
<td>*0.65%</td>
<td>*1.9%</td>
<td>2.7%</td>
<td>1.3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>J48</td>
<td>0.61%</td>
<td>*1.6%</td>
<td>2.9%</td>
<td>0.9%</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

[5] Table 2 SpamAssassin
Test 2

Standard deviation for ham error rates

• V0 to V6 - various approaches for testing
• Ham split into F1 and F2
• Two-fold crossvalidation (CV)

<table>
<thead>
<tr>
<th></th>
<th>hx</th>
<th>ix</th>
<th>μx</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0</td>
<td>2.53%</td>
<td>1.20%</td>
<td>1.10%</td>
</tr>
<tr>
<td>V1</td>
<td>4.63±2.38%</td>
<td>0.79±0.07%</td>
<td>1.65±2.34%</td>
</tr>
<tr>
<td>V2</td>
<td>2.53±0.60%</td>
<td>0.79±0.07%</td>
<td>0.55±0.78%</td>
</tr>
<tr>
<td>V3</td>
<td>5.68±2.08%</td>
<td>1.16±0.33%</td>
<td>2.20±0.78%</td>
</tr>
<tr>
<td>V4</td>
<td>0.63±0.30%</td>
<td>0.14±0.07%</td>
<td>0.55±0.78%</td>
</tr>
<tr>
<td>V5</td>
<td>0.21±0.30%</td>
<td>0.32±0.07%</td>
<td>0.55±0.78%</td>
</tr>
<tr>
<td>V6</td>
<td>0.21±0.30%</td>
<td>0.23±0.07%</td>
<td>1.10±0.00%</td>
</tr>
</tbody>
</table>

Like human error rate of 0.16%
Best Models

Spam error rates

- Good performance of SAnb
- Improvement of factor four is possible
- Spam error rate from as_test

\[
spam_e = \frac{fp}{fp + tn}
\]

<table>
<thead>
<tr>
<th></th>
<th>hx</th>
<th>ix</th>
<th>ux</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td></td>
<td>64.7%</td>
<td></td>
</tr>
<tr>
<td>SAnb</td>
<td></td>
<td>9.5%</td>
<td></td>
</tr>
<tr>
<td>V4</td>
<td>7.3%</td>
<td>45.5%</td>
<td>2.1%</td>
</tr>
<tr>
<td>V5</td>
<td>3.2%</td>
<td>4.9%</td>
<td>2.1%</td>
</tr>
<tr>
<td>V6</td>
<td>2.1%</td>
<td>2.7%</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

threshold value of 1.0 from default value 0.5

[5] Table 4 SpamAssassin
Total Cost Ratio

- TCR in each case better as manual deletion
- TCR 36.42 vs. 1

\[ TCR = \frac{fp + tn}{\sum_{x \in fn} C(x) + fp} \]

<table>
<thead>
<tr>
<th></th>
<th>hx</th>
<th>ix</th>
<th>ux</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>1.53</td>
<td>1.39</td>
<td>1.12</td>
</tr>
<tr>
<td>SAnb</td>
<td>9.31</td>
<td>7.29</td>
<td>2.02</td>
</tr>
<tr>
<td>V4</td>
<td>10.92</td>
<td>2.07</td>
<td>5.46</td>
</tr>
<tr>
<td>V5</td>
<td>26.18</td>
<td>10.50</td>
<td>5.46</td>
</tr>
<tr>
<td>V6</td>
<td>36.42</td>
<td>17.25</td>
<td>2.92</td>
</tr>
</tbody>
</table>

[5]Table 5 SpamAssassin
Conclusions

- A well-trained Bayesian model as core of a good spam filter
- Heuristics alone in any case insufficient
- No direct content features are available:
  - such as phrase occurrence
  - filtering for word pairs, or even triples words
- Example:
  - Word "offers" has a probability of 0.96
  - "special offers" and "valuable offers" o 0.99
  - "approach offers" (as "this approach offers") 0.1 or less
References

1. Comer, Chap. 26, Electronic Mail (SMTP, POP, IMAP, MIME)
2. P. Resnick, RFC 2822, 2001
3. Alexander K. Seewald, Combining Bayesian and Rule Score Learning: Automated Tuning for SpamAssassin
6. Paul Hoffman and Dave Crocker, Unsolicited Bulk Email: Mechanisms for Control, Internet Mail Consortium IMCR-005, October 13, 1997
8. Wikipedia