Computer aided mail filtering using SVM

Lin Liao, Jochen Jaeger

Department of Computer Science & Engineering
University of Washington, Seattle
Introduction

What is SPAM?
- Electronic version of junk mail, bulk-mail, unwanted messages
- Unsolicited Commercial E-mail (UCE), Unsolicited Bulk Email (UBE)

Some impacts
- Annoying unsolicited advertisement or even harmful, offending content
- Waste of resources (time and bandwidth), ~1Mio per day
- Threat to the viability of e-mail and Internet commerce

Reasons for the proliferation of spam
- Efficiency and increasing popularity of e-mail, prone to misuse
- Very low cost for distribution with relatively high number of responses
- Trading of private e-mail addresses has become as business
Filtering E-mail

Typical counter-measures
• Either technical or regulatory – several bills and lawsuits had limited effect

Anti-spam filters
• Automatically identify/classify incoming messages
• Accept legitimate messages and reject spam

Several approaches
• Blacklist of spam sender-addresses – forged headers, faked addresses
• Simple rule-based solutions, hand-crafted key-word patterns
  • Perform poorly, require frequent updates and fine-tuning
• More sophisticated solutions take advantage of machine learning
  • Adaptive, user-specific, improves with experience
Mail Filtering Using SVM

Support Vector Machines (SVM) for mail filtering
- Outperforms other techniques: Rule-based learner, Decision Trees, Naïve Bayesian

Scenario
- Training mode: Analyze test set of marked (classified) mails
  - Extract features, build feature database with number of occurrences for each class
  - Feature Selection – select most decisive (unbalanced) features from database
  - Build feature vectors for each message – project message into feature space
  - Use SVM to learn feature characteristics and build support vectors
- Classify incoming mails using SVM
- Regularly update feature database with new messages
  - Repeat feature selection and apply SVM again
Analyzing Mail

Traditional text classification
• Each word is a feature, each document is a binary feature vector

Special text categorization suggests several extensions
• Integer value vectors for number of occurrences
• Different types of features, e.g.
  • Single words and pairs of words, characters – body content
  • Domain-specific, non-phrasal features, mostly header information
    • Addresses: Resolved familiar E-mail addresses, domain, domain type
    • Content-type, attachments, priority, time zone, receive path
    • Percentage of non-alphanumerical characters and capital letters

Feature data structure
• Type of the feature, string value – discretize non-string values
SVM classifier

- Find separating hyperplane with max distance to closest training example

- Advantage: avoids overfitting
- We used the libsvm implementation for Java
Result 1

Prediction accuracy for different feature selection methods

Number features

Accuracy

chi2

chi2n
Results, top 30 features for chi²

<table>
<thead>
<tr>
<th>From</th>
<th>Subject</th>
<th>To</th>
<th>Body</th>
</tr>
</thead>
<tbody>
<tr>
<td>T: 5</td>
<td>F: 295 : 0 398.59744 W: '.edu'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 50</td>
<td>F: 283 : 15 335.20468 W: 're'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 14</td>
<td>F: 245 : 4 317.8967 W: '.edu'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 13</td>
<td>F: 236 : 2 312.24432 W: 'cs.washington.edu'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 290 : 30 302.8129 W: 'jochen'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 10</td>
<td>F: 200 : 1 266.90472 W: 'jochen jaeger'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 12</td>
<td>F: 195 : 1 260.14935 W: '<a href="mailto:jj@cs.washington.edu">jj@cs.washington.edu</a>'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 56</td>
<td>F: 42 : 427 216.32468 W: '#'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 56</td>
<td>F: 29 : 383 212.34502 W: '&quot;'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 4</td>
<td>F: 151 : 0 204.02783 W: 'cs.washington.edu'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 56</td>
<td>F: 20 : 331 194.832 W: '%'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 35</td>
<td>F: 415 : 181 179.06517 W: 'text/plain'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 4 : 256 164.30302 W: 'click'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 56</td>
<td>F: 67 : 419 146.57521 W: '!'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 123 : 2 135.88109 W: 'washington'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 246 : 72 134.93947 W: 'be removed'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 55</td>
<td>F: 133 : 8 134.66705 W: 'offer'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 55</td>
<td>F: 0 : 177 127.22627 W: 'unsubscribe'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 147 : 24 126.156136 W: 'hi'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 247 : 93 126.041695 W: 'but'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T: 54</td>
<td>F: 182 : 48 126.041695 W: 'removed'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
From: klsd87@ktqblyq.aol.com
To: gacdo@tpfdans.tanet.net
Subject: In Debt :( We Can Help! :o)

This is NOT a loan and you will NOT be turned DOWN ....
DON'T GO FURTHER INTO DEBT BY BORROWING MORE MONEY ....
For more information please visit our site: ....
Or please fill out the attached form at the end ....
Are you thinking about borrowing money to consolidate your bills ....
Are you approaching or already past due on your credit cards ....
Do you usually only pay the monthly minimum on your credit cards ....
Are family disputes over money taking place on a daily basis ....
We can reduce your monthly payments by consolidating your unsecured debt ....
* credit cards * department store accounts * medical bills * collection accounts * unsecured loans ....
Save 20% - 60% on your debt by consolidating your unsecured into one low monthly payment ....
THIS IS NOT A LOAN ....
We reduce or eliminate the high interest that you are currently paying ....
As a non-profit organization, we believe in people, not the bottom line ....
Let us help you get out of debt ....
From: yuko@fiuts.washington.edu
To: jj@cs.washington.edu
Subject: FIUTS education outreach

I am the education outreach coordinator for FIUTS ....
Mara mentioned to me that you were very interested with our education outreach programs ....
I wanted to let you know that I have your name down on my list ....
The fall quarter was a very quiet quarter (education program wise), and we are in the midst of planning
for the winter and spring quartres ....
It is a professional development program for Seattle area K-12 teachers ....
The workshop involves a panel of international students, and I am looking for students that can help us out then ....
If you think you might be interested, please let me know ....
I can give you more information about it ....
Thanks for your willingness to participate in our education outreach programs ....
********************************************************************************
Yuko Mera Christianson Education Outreach Coordinator
Foundation for International Understanding Through Students (FIUTS) Box 352233, University of
Washington Seattle, Washington 98195 Tel: 206.543.0735 Fax: 206.685.8338 Email:
yuko@fiuts.washington.edu www.fiuts.washington.edu ********************************************************************************