

# Statistical Chinese Rules for SpamAssassin

Dr. Quang-Anh Tran

Network Research Center of Tsinghua University

CERNET Computer Emergency Response Team (CCERT)

Email: [qa@ccert.edu.cn](mailto:qa@ccert.edu.cn)

30 Oct. 2004

# Contents

- Quick review on anti-spam
- Our new approach
- Implementation
- Numeric results
- Conclusion

# Part I

## Quick Review on Anti-spam Techniques

# Spam

- Waste of time (money)
  - An Chinese user: 9.2 spam/day, 3 minutes/day, 18 hours/year, 18 USD/year. Total:  $18 * 87\text{M USD/year}$
- Denied of service
  - Hard to find an email among 100 spam
- Waste of computer resources
  - Bandwidth
  - Storage
  - CPU time

# Anti-spam

- Protection
  - Disable open relay
  - Blacklist, RBL, URI, white list
  - Challenge/response, Grey listing, Micro payment
  - Sender ID, Domain Keys
- Detection
  - Rule-based
  - Bayesian
- Response
  - Filter, mark
  - Trace back, Report, Law & Policy

# Rule-based

- Concept
  - Looking for spam-liked pattern, e.g. Subject contains “Free”
- Advantage
  - Rules can be shared, spam knowledge is popularized quickly
- Disadvantage
  - Rules are set up manually, it is hard to keep them up with the changes of spam

# Bayesian

- Concept
  - Train the classifier (detector) upon ham/spam
- Advantage
  - The detector are trained automatically, it is possible to keep them up with the change of spam
- Disadvantage
  - Cannot share the detector, it is hard to popularize spam knowledge

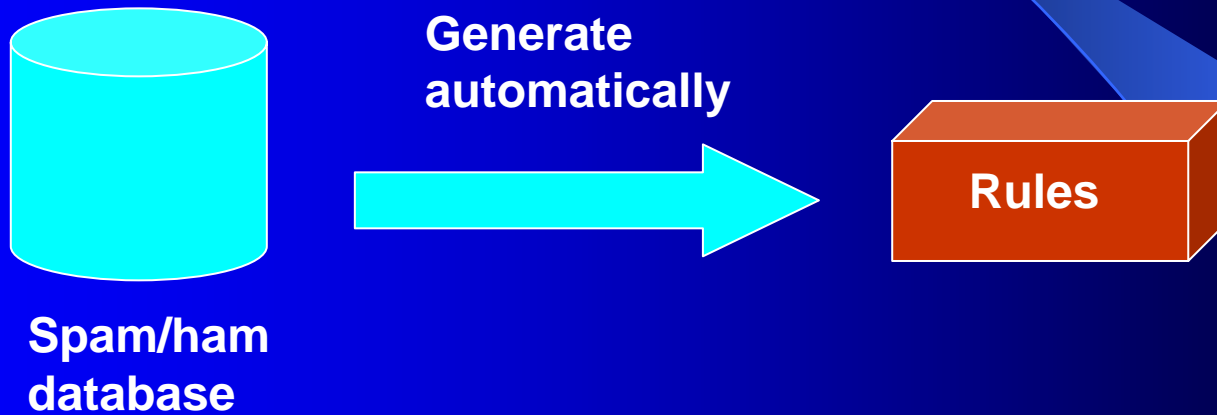
## Part II

# Our New Approach

# Statistical rules

- Concept:
  - Rules are built automatically by statistical method
- Advantage
  - It is possible to keep the rules up with the change of spam
  - It is possible to popularize spam knowledge by sharing the rules

# The framework



# Part III

# Implementation

# SpamAssassin

- Open source, Perl, widely used (Apache, RedHat)
- Rule-based approach

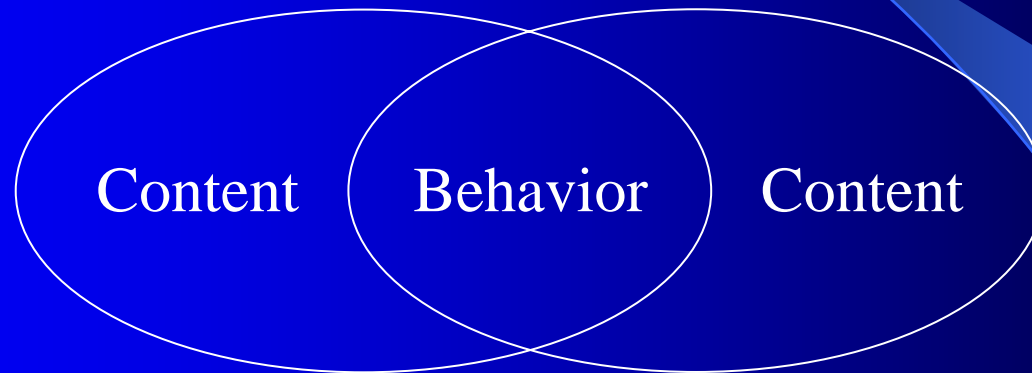
body	DEAR_FRIEND	/^\s*Dear Friend\b/i
describe	DEAR_FRIEND	Dear Friend? That's not very dear!
score	DEAR_FRIEND	0.542

- Allow user-defined rules
- Integrated with Sendmail, Qmail, Postfix, Exim
- Run on Unix, Windows

# SpamAssassin

- Scores of SA default rules are not suitable for Chinese spam
- No special rules for Chinese spam

# Chinese rules vs. English rules



English rules

Chinese rules

# Statistical spam/ham-liked words

- Spam-liked words
  - Most frequent words in Spam's Subject
  - Most frequent words in Spam's Body
- Ham-liked words
  - Most frequent words in Ham's Subject
  - Most frequent words in Ham's Body

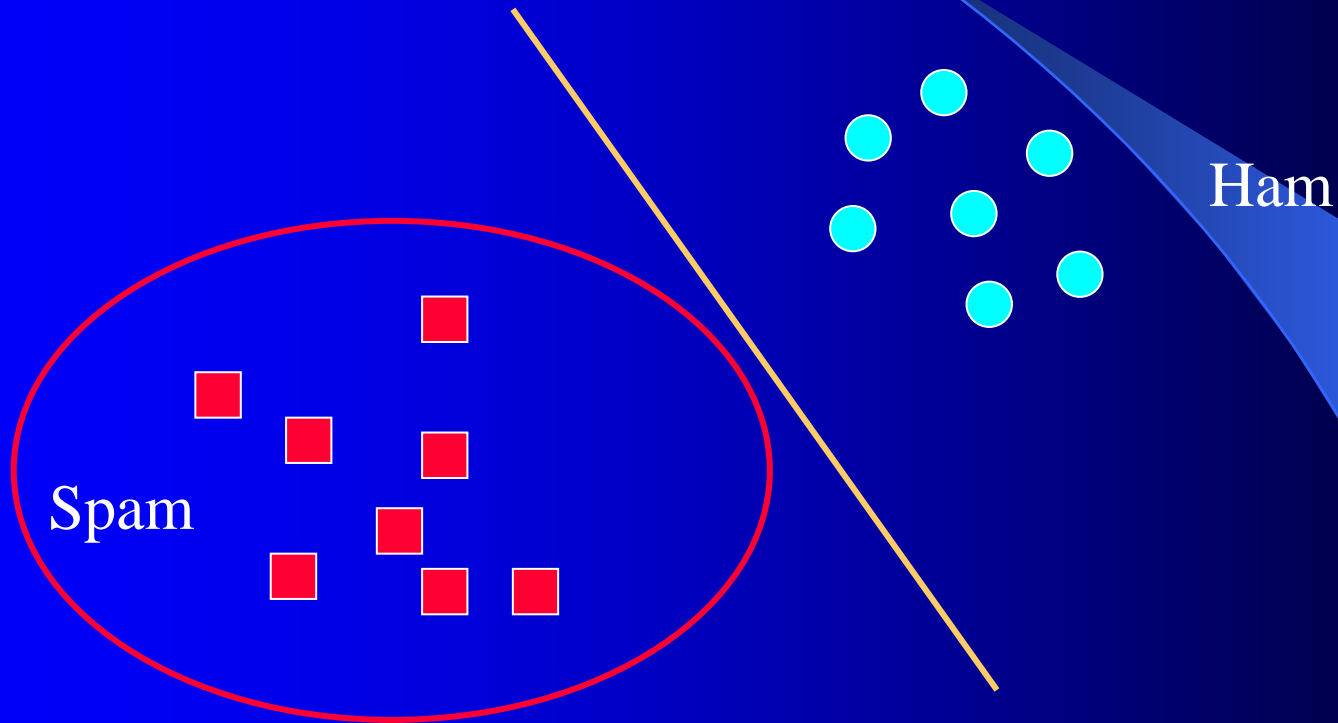
# Rule creation

- Follow the format of SpamAssassin's rule
- Subject rules
- Body rules

header	CN_SUBJECT_1	Subject =~ /免费/
describe	CN_SUBJECT_1	Subject contains "免费"
score	CN_SUBJECT_1	0.749434

body	CN_BODY_4	/企业/
describe	CN_BODY_4	Body contains "企业"
score	CN_BODY_4	0.739289

# The role of ham-like words



Separate plane with  
ham-like words removed

Separate plane without  
removing ham-like words

# Score setting

- Genetic Algorithms
- Chromosome
- Objective
  - Recall, Precision, Error
- Crossover, mutation and selection
- Stop criteria

# Part IV

## Numeric Results

# Numeric results

- Size of training and testing dataset
  - Spam: 20,000
  - Ham: 100,000
- Number of rules
  - Subject rules: 550
  - Body rules: 50

# Error rate vs. threshold

Threshold	Spam recall (%)	Ham error (%)
0.5	82.6	14.1
1	73.4	5.6
1.5	65.4	2.8
2	60.9	2.0
2.5	56.5	1.2
3	51.4	0.8
3.5	46.9	0.7

# Performance

- It takes 0.06 seconds to scan an email with size 3519.3 bytes (P4-2.0G CPU, 512M)
- 1,440,000 emails / day

- Outlook 快捷方式
- 日历
- 联系人
- 收件箱 (482)
- dhx@cernet (2)
- 802.1X (229)
- 发件箱
- 已发送的邮件
- 已删除的邮件 (405)
- 自定义快捷方式
- 其他快捷方式

发件人	主题	接收时间	大小
support@ca...	*****SPAM***** 抱谷戾岫岫伏企值厝厝厝搦榭榭倂倂倂倂倂...	2004-9-21 (星期二) 9:39	22 KB
华夏数据网	*****SPAM***** 十一期间推出注册域名赠送精美网站活动	2004-9-21 (星期二) 9:39	13 KB
fdf	*****SPAM***** www.zhihuidi.com,网尽天下,股雄网,...	2004-9-21 (星期二) 9:39	12 KB
hello	*****SPAM***** dhx 2004版--715万企业名录大全!	2004-9-21 (星期二) 9:39	11 KB
asdf@asdf.com	*****SPAM***** 嗨,是我!	2004-9-21 (星期二) 9:38	7 KB
ksbnsz	*****SPAM***** 帮您建立购物商城	2004-9-20 (星期一) 23:09	11 KB
kunxlv	*****SPAM***** 帮您建立购物商城	2004-9-20 (星期一) 23:09	10 KB
limol278@c...	*****SPAM***** 04~n09XCp19	2004-9-20 (星期一) 23:08	31 KB
cttra	*****SPAM***** 帮您建立购物商城	2004-9-20 (星期一) 23:08	11 KB
zijingzhyf	*****SPAM***** 帮您建立购物商城	2004-9-20 (星期一) 23:08	10 KB
中国人	*****SPAM***** 勿忘历史,...	2004-9-20 (星期一) 23:08	12 KB

已删除该邮件多余的换行符。要恢复,请单击此处。

发件人: zijingzhyf [dmmhuyams@sina.com]  
 主题: \*\*\*\*\*SPAM\*\*\*\*\* 帮您建立购物商城  
 附件: 帮您建立购物商城 (1.36 KB) (4 KB)

SA detects email with score higher than 4.5 as spam, score of this email is 10.3

Content analysis details: (10.3 points, 4.5 required)

pts	rule name	description
1.0	CN_SUBJECT_133	Most frequent word in Chinese spam subject
0.0	CN_SUBJECT_274	Most frequent word in Chinese spam subject
1.0	CN_SUBJECT_564	Most frequent word in Chinese spam subject
0.3	CN_SUBJECT_156	Most frequent word in Chinese spam subject
1.0	CN_BODY_122	BODY: Most frequent word in Chinese spam body
1.0	CN_BODY_37	BODY: Most frequent word in Chinese spam body
1.0	CN_BODY_44	BODY: Most frequent word in Chinese spam body
1.0	CN_BODY_433	BODY: Most frequent word in Chinese spam body
1.0	CN_BODY_59	BODY: Most frequent word in Chinese spam body
2.3	RCVD_IN_BL_SPAMCOP_NET	RBL: Received via a relay in bl.spamcop.net [Blocked - see < <a href="http://www.spamcop.net/bl.shtml?222.94.170.102">http://www.spamcop.net/bl.shtml?222.94.170.102</a> >]
0.8	MSGID_FROM_MTA_HEADER	Message-Id was added by a relay

Without the Chinese rules, the score of this email is 3.1, SA detects it as ham

# Chinese rules releases

- SpamAssassin website
  - <http://wiki.apache.org/spamassassin/CustomRulesets>
  - <http://www.exit0.us>
- CCERT website
  - [http://www.ccert.edu.cn/spam/sa/Chinese\\_rules.htm](http://www.ccert.edu.cn/spam/sa/Chinese_rules.htm)
- Google search

# Conclusion

- User information feedback
- Online training
- Interference between rule sets
- URI rules
- User and server group
- Integrated with other approaches

# References

- Alan Schwartz, SpamAssassin, O'Reilly Media, Inc. 2004
- SpamAssassin, URL:  
<http://www.spamassassin.org>

# Thank you!

Quang-Anh Tran  
30 Oct. 2004