Spam detection using IP geolocation

O-talk

Andriy Stetsko

Outline

Possible sources of geolocation data

Overview of available geolocation databases

Results of tests of selected geoip databases

- performance test
- coverage & accuracy tests

Results of tests of Bayes classifier detection accuracy when

- geolocation information IS NOT used
- geolocation information IS used

Sources of geolocation data (1)

Public WHOIS database

- maps IPs & real-world entries
- organization address ≠ geolocation of IP

Users

- answer to location questions on web-site entries
- can be wrong

Applications

- HTTP Accept-Charset header sent by browser
- not always available
- can be falsified

Sources of geolocation data (2)

Round trip time to landmark

- ICMP echo message
- not all target hosts respond to ICMP echo
- target host may consider it as attack
- poor with regard to hosts with high latency connections
- target host can delay its replies

ISP

- obtain (purchase) network topology

Overview of geolocation databases (1)

IP2Location

- 18 databases

MaxMind

- GeoIP Country, GeoIP City
- **GeoLite Country**, GeoLite City

Quova

Digital-element

IPligence

- Lite, Basic, Max
- **Lite Free**

Overview of geolocation databases (2)

Geobytes

IPInfoDB (compiled from GeoLite City)

- **Country**
- City
  - 3 IP digits precision
  - 4 IP digits precision

Software77

IP::Country::Fast

IP::Country::DB_File

- built from publicly available statistics files of Regional Internet Registries

Databases to test

Performance test

Test measures time needed to process 1000000 requests

Test was repeated 10 times for each database

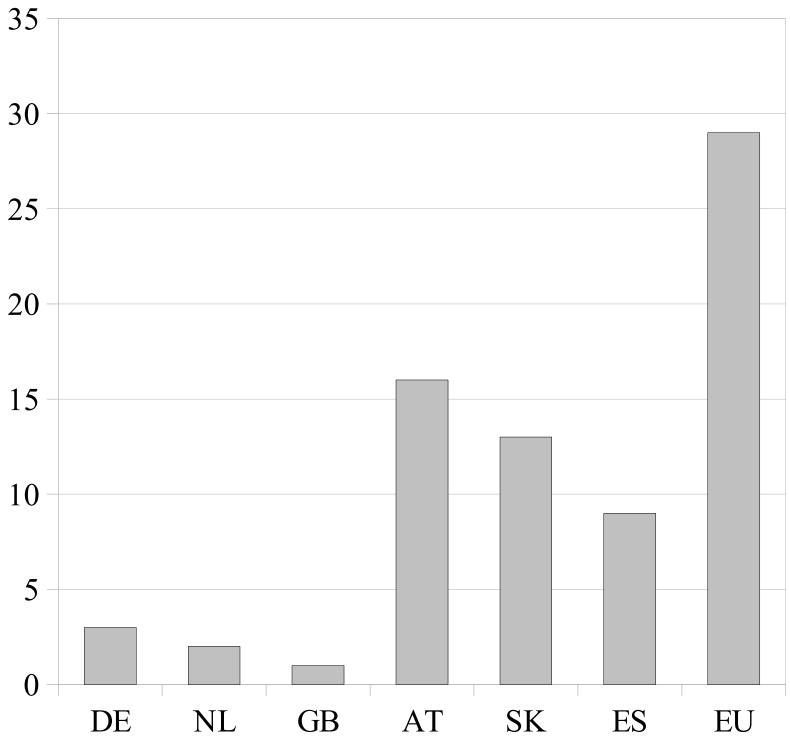
Coverage & accuracy tests (1)

Test was done for 840 IP addresses

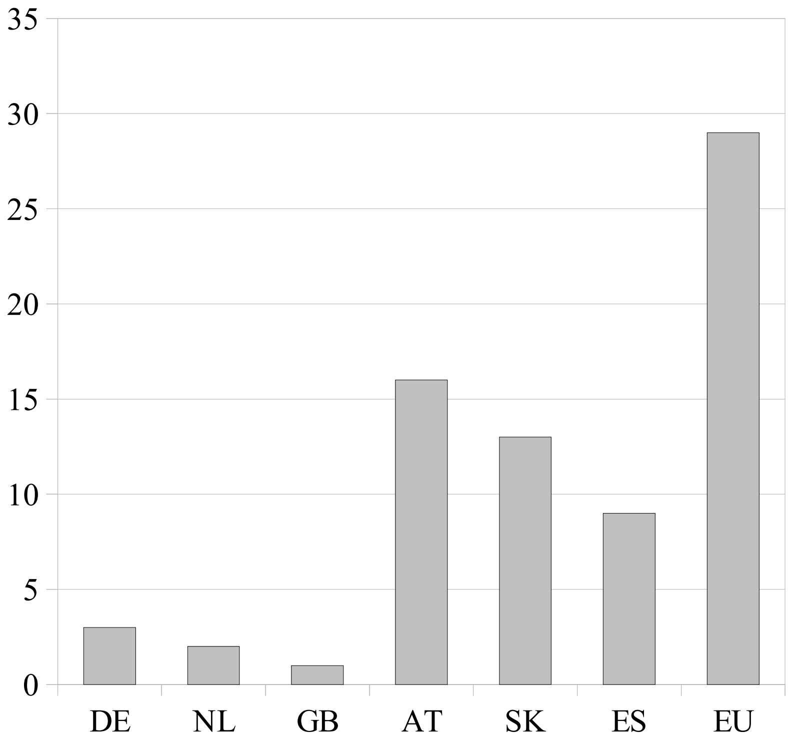
Coverage = (840 - #Uncovered) / 840

Accuracy = (840 - #Incorrect) / 840

Coverage & accuracy tests (2)



Database #3



Database #4

Bayes detection accuracy (1)

SpamAssassin v. 3.3.1 running on Perl v. 5.10.1

RelayCountry plugin

- analyses “Received” headers
- adds “X-Relay-Countries” header

Database of e-mails

- 10762 spam & 9072 ham
- e-mails contain headers added by spam detection software

Bayes detection accuracy (2)

Test no.	Train (%)	Test (%)	Auto-learning (E/D)	Auto-expiration (E/D)
1	50	50	D	E
2	70	30	D	E
3	70	30	D	D
4	70	30	D	D

Test #4:

No difference between detection accuracy of Bayes classifier when RelayCountry plugin is used and when it is not used

Bayes detection accuracy (3)

Test #1 and #2:

- introduction of geolocation info increased detection accuracy but not so much
- ham recognition accuracy was the same in both cases
- detection accuracy in Test#2 was worse than in Test#1

Test #3:

- detection accuracy was higher than in Tests #1 and #2

Conclusion & future work

Geolocation info increases detection accuracy of SpamAssassin Bayes classifier

- increase weights of tokens containing geolocation info

Token expiration has great impact on detection accuracy of Bayes classifier

- propose more effective expiration policy

Explore correlation between e-mail charset and country code (returned by RelayCountry) of e-mail sender

Explore correlation between TLD of sender e-mail and country code (returned by RelayCountry module) of e-mail sender

Configuration tool for RelayCountry plugin

EVROPSKÁ UNIE

esf

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

2007–13

OP Vzdělávání
pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Thank you for your attention!