Internet Measurement and Data Analysis (12)

Kenjiro Cho

2011-12-14

review of previous class

Class 11 Measuring time series of the Internet

- Internet and time
- network time protocol
- time series analysis
- exercise: time series analysis

today's topics

Class 12 Measuring anomalies of the Internet

- anomaly detection
- spam filters
- Bayes' theorem
- exercise: anomaly detection

anomalies

- traffic problems
- routing problems, reachability problems
- DNS problems
- attacks, intrusions
- CPU load problems

causes of anomalies

- access concentration, congestion
- attacks: DoS, viruses/worms
- outages: equipment failures, circuit failures, accidents, power outages
- maintenance

anomaly detection

- avoid or reduce losses caused by service degradation or disruption
- monitoring individual items: post an alert when the monitored value exceeds the predefined threshold
  - passive monitoring
  - active monitoring
- signature based anomaly detection:
  - pattern matching with known anomalies
  - IDS: Intrusion Detection System
  - cannot detect unknown anomalies
  - need to keep the pattern database up-to-date
- anomaly detection by statistical methods:
  - detect discrepancies from normal states
  - in general, need to learn "normal" states

responses to anomalies

- report to system administrators
  - posting alert messages
- identifying types of anomalies
  - provide information to help operators to understand the cause of the problem
  - difficult to find causes, especially for statistical methods
- automated responses
  - automatically generating filtering rules, failover, etc

anomaly examples

- Flash Crowd
  - access concentration to specific services (news, events, etc)
- DoS/DDoS
  - send a large volume of traffic to a specific host
  - zombie PCs are often used as attackers
- scanning
  - for most cases, to find hosts having known security holes
- worms/viruses
  - many incidents (SQL Slammer, Code Red, etc)
- route hijacking
  - announcing someone else's prefixes (mostly by mis-configuration)

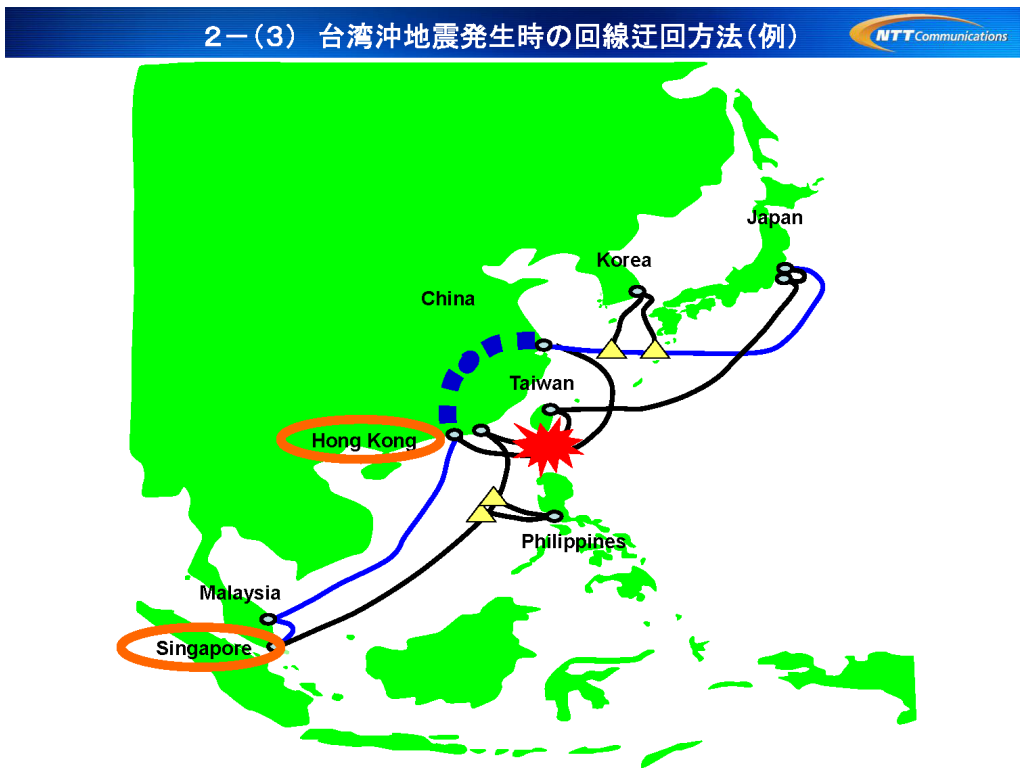
YouTube hijacked

- 2008-02-24: worldwide traffic to YouTube was redirected to Pakistan
- cause
  - by the order of Pakistan government, Pakistan Telecom announced a false prefix on BGP in order to block domestic access to YouTube
  - a large ISP, PCCW, leaked the announce to the global Internet
  - as a result, worldwide traffic to YouTube was redirected to Pakistan by the false route announcement

reference:
`http://www.renesys.com/blog/2008/02/pakistan_hijacks_youtube_1.shtmly`

communication service disruption by Taiwan earthquake

- 2006-12-26: M7.1 earthquake occurred off the coast of Taiwan
- submarine cables were damaged, communication services to/from Asia were affected
- Indonesia's international link capacity became less than 20%
- ISPs restored services by rerouting



source: JANOG26

`http://www.janog.gr.jp/meeting/janog26/doc/post-cable.pdf`

disconnection between ISPs

- a case of a dispute of 2 Tier-1 ISPs over connection fees
- in 2005, Level 3 asked Cogent to switch from non-paid peering to paid connection because of the increase in traffic
- other cases
  - in 2008, Cogent and Telia stopped peering
  - in 2008, Level 3 and Cogent stopped peering
  - in 2010, Level 3 and Comcast dispute

references:
`http://www.renesys.com/blog/2006/11/sprint-and-cogent-peer.shtml`
`http://wirelesswire.jp/Watching_World/201012011624.html`

anomaly detection by statistical methods

- time-series
- correlation
- PCA
- clustering
- entropy

identifying and filtering SPAM email

SPAM: unsolicited bulk messages

SPAM test methods

- tests by senders
  - white lists
  - black lists
  - gray listing
- tests by content
  - bayesian spam filter: widely used
  - learns frequencies of words from SPAM and HAM email, calculate a probability for an email to be SPAM
  - the accuracy improves as it is used

conditional probability

Question:

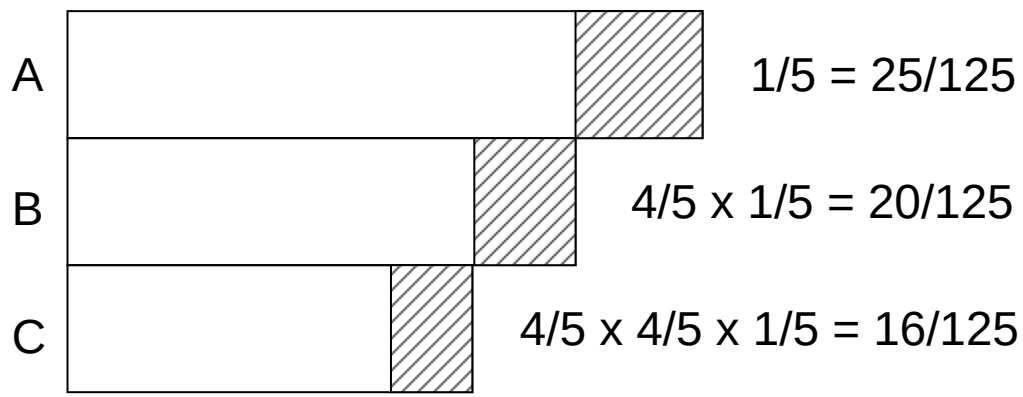
- Student K leaves behind his cap once every 5 times. He visited 3 friends, A, B and C in this order and when he came home he found his cap was left behind. What is the probability that K left his cap at B's house? (1976, Waseda University, entrance exam)

conditional probability

Question:

- Student K leaves behind his cap once every 5 times. He visited 3 friends, A, B and C in this order and when he came home he found his cap was left behind. What is the probability that K left his cap at B's house? (1976, Waseda University, entrance exam)

Answer:



the prob. of the cap left at B / the prob. of the cap left at either house $= 20/61$

Bayes' theorem

conditional probability

- the probability of B when A is known to occur: $P(B|A)$
  - the sample space is restricted to event A, within which the area $(A \cap B)$ is of interest

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- independence: when A and B are independent

$$P(B|A) = P(B) and P(A|B) = P(A)$$

Bayes' theorem

- posterior probability: when A causes B, the probability of event A occurring given that event B has occurred: $P(A|B)$
  - $P(A)$: the probability of A to occur (prior probability)
  - $P(A|B)$: the probability of A occurring given that B has occurred (posterior probability)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

applications of bayes' theorem

based on the observations, inferring the probability of a cause: many engineering applications

- communications: based on received signal with noise, extract original signal
- medical tests: based on a medical test result, find the probability of a person actually having the disease
- spam tests: based on the content of email, find the probability of an email being spam

example: disease test

Question:

- the population ratio having a certain disease is 50/1000. a test for the disease is known to have positive for 90% of people having the disease but also have positive for 10% of people not having the disease.
  when a person get positive by this test, what is the probability of the person actually having the disease?

example: disease test

Question:

- the population ratio having a certain disease is 50/1000. a test for the disease is known to have positive for 90% of people having the disease but also have positive for 10% of people not having the disease.
  when a person get positive by this test, what is the probability of the person actually having the disease?

Answer: the probability of the person having the disease:
$P(D) = 50/1000 = 0.05$
the probability of a result to be positive: $P(R) = P(D \cap R) + P(\bar{D} \cap R)$
when the result is positive, the posterior probability that the person has the disease

$$\begin{aligned} P(D|R) &= \frac{P(D \cap R)}{P(R)} \\ &= (0.05 \times 0.9)/(0.05 \times 0.9 + 0.95 \times 0.1) = 0.321 \end{aligned}$$

spam email tests

- for training, prepare spam messages (SPAM) and non-spam messages (HAM)
- for words often included in SPAM, compute
  - the conditional probability that SPAM include a word
  - the conditional probability that HAM include a word
- then, compute the posterior probability of an unknown message being SPAM

example: for word A, assume $P(A|S) = 0.3$, $P(A|H) = 0.01$, $\frac{P(H)}{P(S)} = 2$. then, compute $P(S|A)$.

$$\begin{aligned} P(S|A) &= \frac{P(S)P(A|S)}{P(S)P(A|S) + P(H)P(A|H)} \\ &= \frac{P(A|S)}{P(A|S) + P(A|H)P(H)/P(S)} \\ &= \frac{0.3}{0.3 + 0.01 \times 2} = 0.94 \end{aligned}$$

naive Bayesian classifier

- in practice, multiple tokens are used
  - combinations of tokens require huge data
- naive Bayesian classifier: assumes tokens are independent
  - tokens are not independent, but it works most of the cases
  - training step:
    - using classified training samples, compute the conditional probabilities of tokens being included in SPAM
  - prediction step:
    - for unknown messages, compute the posterior probabilities of tokens included in a message to decide whether the message is SPAM or HAM
- in the training step, the conditional probability of each token can be independently computed
- use Bayesian joint probability to compute the joint probability for SPAM testing from individual token's SPAM probability

naive Bayesian classifier (details)

let tokens be $x_1, x_2, \ldots, x_n$. when these tokens are observed, the posterior probability of a message being SPAM is:

$$P(S|x_1, \ldots, x_n) = \frac{P(S)P(x_1, \ldots, x_n|S)}{P(x_1, \ldots, x_n)}$$

the numerator shows the joint probability of the token to be observed and the message is SPAM, and thus, can be written as follows. by applying the definition of conditional probability:

$$\begin{aligned} P(S, x_1, \ldots, x_n) &= P(S)P(x_1, \ldots, x_n|S) \\ &= P(S)P(x_1|S)P(x_2, \ldots, x_n|S, x_1) \\ &= P(S)P(x_1|S)P(x_2|S, x_1)P(x_3, \ldots, x_n|S, x_1, x_2) \end{aligned}$$

assume each token is conditionally independent from other tokens

$$P(x_i|S, x_j) = P(x_i|S)$$

then, the above joint probability becomes

$$P(S, x_1, \ldots, x_n) = P(S)P(x_1|S)P(x_2|S)\cdots P(x_n|S) = P(S)\prod_{i=1}^{n} P(x_i|S)$$

thus, assuming tokens are independent, the posterior probability of the message being SPAM is

$$P(S|x_1, \ldots, x_n) = \frac{P(S)\prod_{i=1}^{n} P(x_i|S)}{P(S)\prod_{i=1}^{n} P(x_i|S) + P(H)\prod_{i=1}^{n} P(x_i|H)}$$

previous exercise: autocorrelation

- compute autocorrelation using traffic data for 1 week

```
# ruby autocorr.rb autocorr_5min_data.txt > autocorr.txt
# head -10 autocorr_5min_data.txt
2011-02-28T00:00 247 6954152
2011-02-28T00:05 420 49037677
2011-02-28T00:10 231 4741972
2011-02-28T00:15 159 1879326
2011-02-28T00:20 290 39202691
2011-02-28T00:25 249 39809905
2011-02-28T00:30 188 37954270
2011-02-28T00:35 192 7613788
2011-02-28T00:40 102 2182421
2011-02-28T00:45 172 1511718
# head -10 autocorr.txt
0 1.0
1 0.860100559860259
2 0.859909329457425
3 0.8568488888567
4 0.856910911636432
5 0.853982084154458
6 0.850511942135165
7 0.848741549347501
8 0.845725096810473
9 0.840762312233673
```

computing autocorrelation functions

autocorrelation function for time lag k

$$R(k) = \frac{1}{n} \sum_{i=1}^{n} x_i x_{i+k}$$

normalize by $R(k)/R(0)$, as when $k = 0$, $R(k) = R(0)$

$$R(0) = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

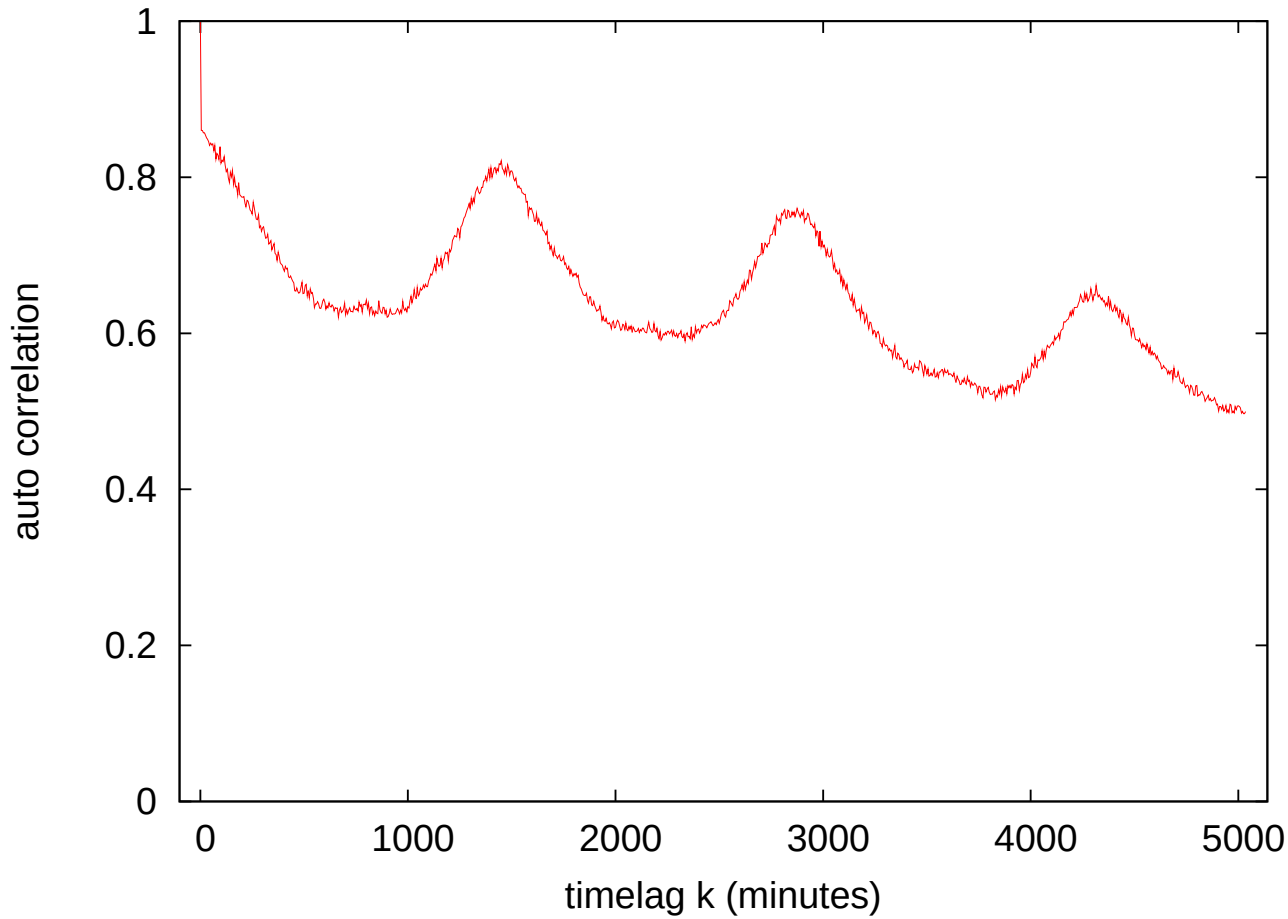
need 2n data to compute $k = n$

autocorrelation computation code

```
# regular expression for matching 5-min timeseries
re = /^(\d{4}-\d{2}-\d{2})T(\d{2}:\d{2})\s+(\d+)\s+(\d+)/

v = Array.new() # array for timeseries
ARGF.each_line do |line|
  if re.match(line)
    v.push $3.to_f
  end
end

n = v.length # n: number of samples
h = n / 2 - 1 # (half of n) - 1

r = Array.new(n/2) # array for auto correlation
for k in 0 .. h # for different timelag
  s = 0
  for i in 0 .. h
    s += v[i] * v[i + k]
  end
  r[k] = Float(s)
end

# normalize by dividing by r0
if r[0] != 0.0
  r0 = r[0]
  for k in 0 .. h
    r[k] = r[k] / r0
    puts "#{k} #{r[k]}"
  end
end
```

autocorrelation plot

```
set xlabel "timelag k (minutes)"
set ylabel "auto correlation"
set xrange [-100:5140]
set yrange [0:1]
plot "autocorr.txt" using ($1*5):2 notitle with lines
```



summary

Class 12 Measuring anomalies of the Internet

- anomaly detection
- spam filters
- Bayes' theorem
- exercise: anomaly detection

next class

Class 13 Data mining (12/21)

- pattern extraction
- classification
- clustering
- exercise: clustering