Seoul
National
University

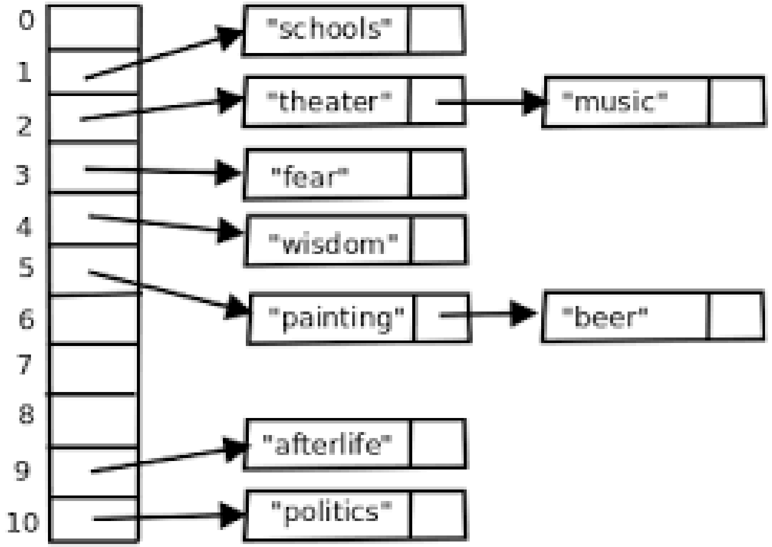
BB Model Examples

Name: Chong-kwon Kim

SCONE
Lab.

Hash Table

- How to handle if more than one objects hashed to the same slot?



- Chain Hashing
  - Use linked list
- Search time
  - Assume m objects into n slots
  - If search word is not in Hash table
    - ➔ Expected number of objects in a slot = m/n
  - If search word is in the table
    - ➔ Expected number of objects in a slot = 1+(m−1)/n
  - Worst case: if m=n, Max. Load $\geq \ln n / \ln \ln n$ with high probability

Set Membership Problem

- **Set membership problem**
  - Determine if an entity is a member of a set
- **Approximate set membership**
  - Allow wrong membership decisions if the probability of wrong is small
  - Examples:

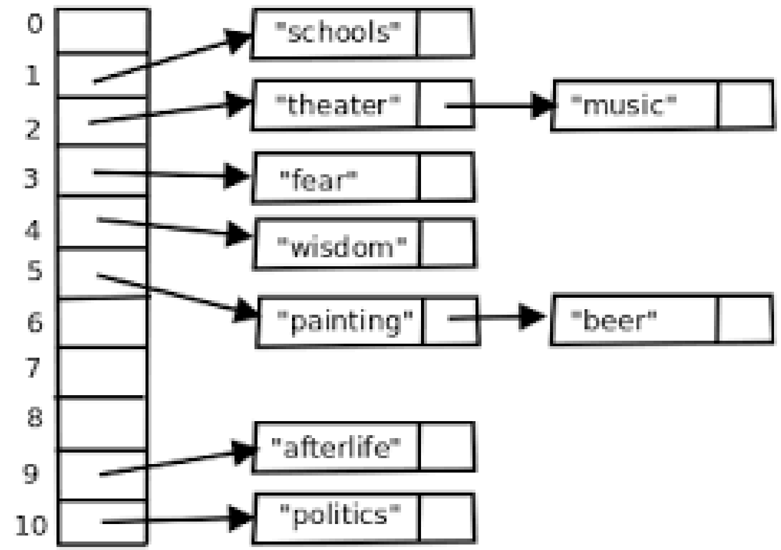
    Not a negative (i.e. positive), but judge as negative

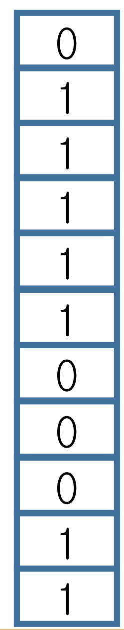
    - Spam email detection: Determine spam emails as normal (False negative)
    - Tumor detection: Determine normal clients as tumor patients (False positive)

    Not a positive (i.e. negative), but judge as positive
- Consider problems
  - Controlled false positives are allowed
  - Save space (Memory)

Set Membership Problem

- Two methods
  - Hash table
  - Bloom filter
- Hash table
  - Instead of storing entities in a hashed position, only mark the position





- Membership of target x
  - If hash position of x is marked, ➔ Yes

    If not, ➔ No
- False positive prob.
  - Probability that a slot is marked

Hash Table

- Given a set $S = \{s_1, s_2, \cdots, s_m\}$ of m elements, is x an element of S?
- Hash each element s_i with b bit long index and mark the hash table (bit map)
  - m balls into 2^b bins

Let $b = 2log_2 m$, $1 - \left(1 - \frac{1}{m^2}\right)^m < \frac{1}{m}$

- Pr(False positive) = Pr(marked bin) = $1 - \left(1 - \frac{1}{2^b}\right)^m$

$$\approx 1 - e^{-m/2^b}$$

- To make the false positive probability $\leq$ c

$$e^{-m/2^b} \geq 1 - c,$$

$$b \geq \log_2 \frac{m}{\ln(1/(1-c))}$$

b = Ω(log m)

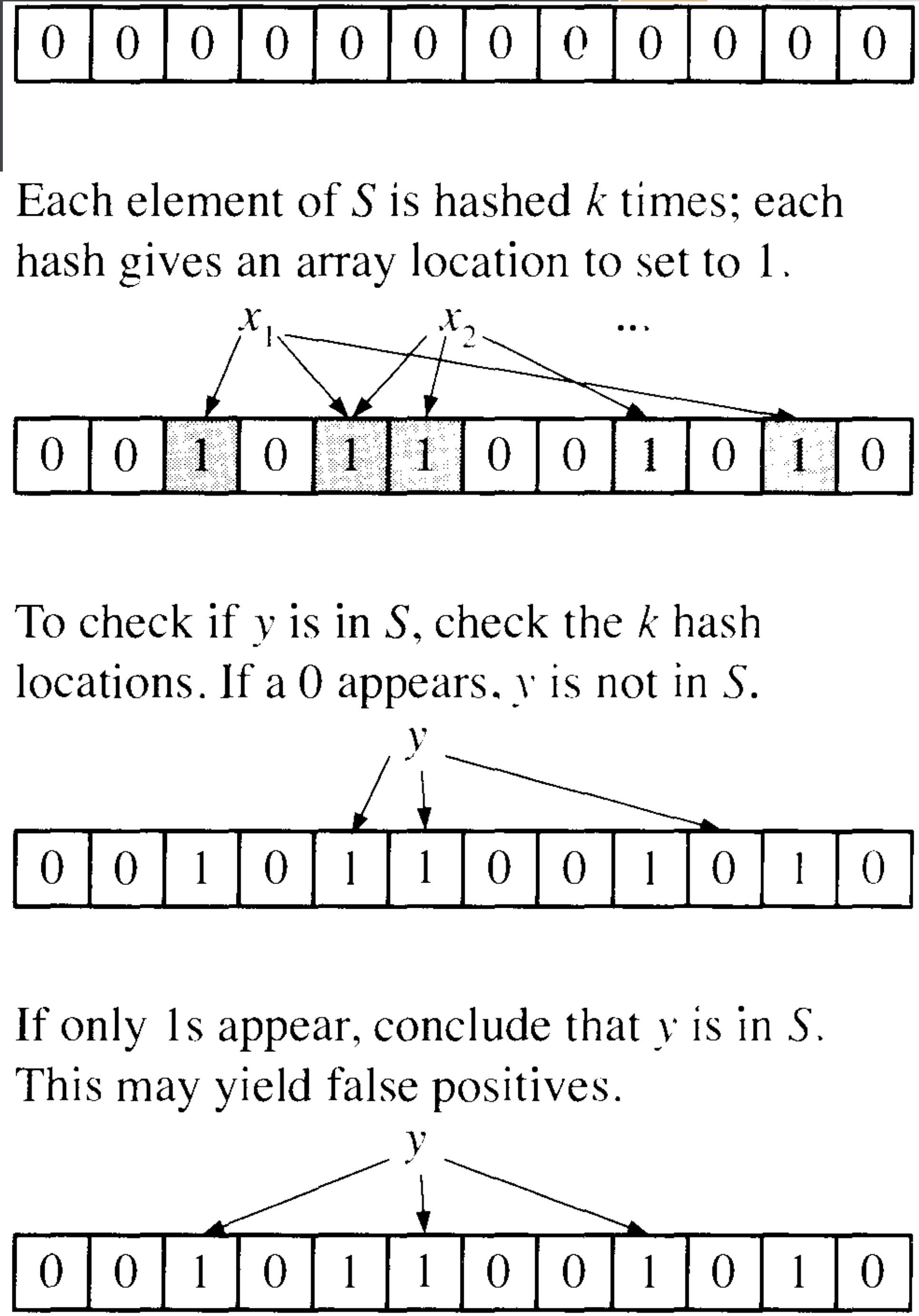
Announcements

- We will jump to Chapter 7
  - Return to Chapter 6 if time allows
- Supp. Class on this Friday

Bloom Filter

- Any better solutions for *approximate set membership* problem?

- How about applying several different hash functions to an element?



Bloom Filter

There are trade-offs between Bloom filter size, # hash functions, and false probability

- Assume m elements are stored in a Bloom filter
  - Each to k slots
- Optimal number of hash functions, k ?
- Large k
  - More chances to meet 0 bit slots
  - More 1 slots
- Small k
  - Less chances to meet 0 bit slots
  - Less 1 slots
- Probability of false positive ≡ probability that all k slots are 1
  - First, Probability that a slot is 0 = $\left(1-\frac{1}{n}\right)^{km} \approx e^{-km/n}$ $\equiv p$

n: # slots
m: # elements
k: # hash functions
In terms of BB model,
throw k·m balls into n bins

Bloom Filter

- Probability of false positive

$$\left(1-\left(1-\frac{1}{n}\right)^{km}\right)^k \approx (1-e^{-km/n})^k = (1-p)^k$$

- Let f(k) = $(1-e^{-km/n})^k = (1-p)^k$
- Let g(k) = ln f(k)

$$\frac{dg}{dk} = \ln(1-e^{-km/n}) + \frac{km}{n}\frac{e^{-km/n}}{1-e^{-km/n}}$$

g() is minimum at k = (n/m)*ln 2

- Note that p = $e^{-k\left(\frac{m}{n}\right)} = \frac{1}{2}$, $at\ k =$(n/m)*ln 2

f($k = \ln 2 \cdot (\frac{n}{m})$) = $(\frac{1}{2})^k$ = $(0.6185)^{n/m}$

Hash Table VS Bloom Filter

- For the approximate set membership problem, which one is better?

- Hash Table
  - Require $\Omega(\log m)$ bits per element to achieve constant false positive error probability

- Bloom Filter
  - Require $\Omega(1)$ bits per item
  - Example
    - When $n/m = 8$, k is 5 or 6

      ➔ False positive probability ≈ 0.02

Bloom Filter

- Prior explanations are based on the assumption that

  Fraction of 0 slots is p $= e^{-k\left(\frac{m}{n}\right)}$

- Actual case
  - Fraction of empty bins after throwing km balls into n bins

- Questions

  1. E[# entries with 0 balls]
  2. How close is E[# entries with 0 balls] to np

Bloom Filter

- Let $X_j = \begin{cases} 1, & if\ bin\ j\ is\ empty \\ 0, & o.w \end{cases}$
- Let X $= X_1 + X_2 + \cdots + X_n$

  ➔ E[# entries with 0 balls] = E[X]

- E[X] = $\sum_i E[X_i] = n \cdot (1 - 1/n)^{km}$

Bloom Filter

2. How close is E[# entries with 0 balls] to np

- Let $p' = (1 - 1/n)^{km}$ and r = km
- Pr(|X − n$\cdot p'$|$\geq \varepsilon n$ in Exact Case(EC))

  $\leq$ e$\sqrt{r}$ ·Pr(|X − n$\cdot p'$|$\geq \varepsilon n$ in Poisson Case(PC))

- Consider Poisson Case

  ➔Xj's are independent and each of them has probability p' to be 1

  ➔ X is sum of n independent Bernoulli trials each with probability p' of success

  ➔ Bin(n, p')

Bloom Filter

- Pr(|X − n· p'|≥ εn) in Exact Case(EC))

  ≤ e$\sqrt{r}$ ·Pr(|X − n· p'|≥ εn) in Poisson Case(PC))

  = e$\sqrt{r}$ ·Pr(|Bin(n, p') − n· p'|≥ εn) in Poisson Case(PC))

  ≤ e$\sqrt{r}$ ·$(2e^{-n\varepsilon^2/3p'})$ Apply Chernoff Bound

  ≤ 0.00001 When n is large,

Coupon Collection Problem

- Previously, we've showed that

  Expected # coupons required to collect all n types is $n \cdot H_n \approx n \cdot \ln n$ (Section 2.4.1)

Section 3.3.1

- Also, after collecting $n \cdot \ln n + cn$ coupons,
- Pr(i–th type is not collected) = $(1-1/n)^{n \cdot \ln n + cn}$

$$\le e^{-\frac{n \cdot \ln n + cn}{n}} = e^{-c}/n$$

- Pr(Any missing types)

  $\le \sum \Pr($i–th type is not collected$) = e^{-c}$

➔ Pr(No missing type) $\ge 1 - e^{-c}$

Coupon Collection Problem

- Theorem: Let X be # coupons collected before obtaining all n types of coupons. Then for any constant c

$$\lim_{n\to\infty} \Pr(X > n \cdot \ln n + cn) = 1 - e^{-e^{-c}}$$

- Sharp threshold:
  - Distribution is highly concentrated around the mean
  - For large n,

    When c=−4, $1 - e^{-e^{-c}} \approx 1$

    When c=4, $1 - e^{-e^{-c}} \approx 0.02$

    ➔ # coupons between $[n \cdot \ln n - 4n,\ n \cdot \ln n + 4n]$ is 98%

Coupon Collection Problem

- Note that Coupon Collection Problem is the same as the Balls into Bins model
  - m balls = m coupons
  - n types = n bins

- Again, we approximate with much easier PC and then apply the bounds

- With the Poisson approximation
  - # balls in a bin is Poisson with mean ln n + c

    ➔ Expected total # balls , m = $n \cdot \ln n + cn$
  - Pr(i–th bin is empty)=Pr$(Y_i = 0)$ = e^{-c}/n
  - Pr(No empty bin) = $(1 - e^{-c}/n)^n$ = $e^{-e^{-c}}$

Coupon Collection Problem

- Let ε be the event that no empty bin
- Let Y be # balls thrown in the Poisson case
- Let r = $\sqrt{2m \cdot ln\, m}$
- For large n

Note: m = $\boldsymbol{n \cdot ln\, n + cn}$

$$\Pr(\varepsilon) = \Pr(\varepsilon \mid |Y-m| \le r) \cdot \Pr(|Y-m| \le r) + \Pr(\varepsilon \mid |Y-m| > r) \cdot \Pr(|Y-m| > r)$$

$$\approx \Pr(\varepsilon \mid |Y-m| \le r) \cdot \Pr(|Y-m| \le r)$$

$$\approx \Pr(\varepsilon \mid |Y-m| = 0) \cdot \Pr(|Y-m| = 0)$$

We need to prove that
$\Pr(|Y-m| > r) \approx \mathbf{0}$
$\Pr(|Y-m| \le r) \approx \Pr(|Y-m|=0) \approx 1$