Design and Analysis of Algorithms I

Data Structures

Bloom Filters

Bloom Filters: Supported Operations

Raison Dêtre: fast Inserts and Lookups.

Comparison to Hash Tables:

Pros: more space efficient.

Cons:

1) can't store an associated object
2) No deletions
3) Small false positive probability
   (i.e., might say x has been inserted even though it hasn't been)

Bloom Filters: Applications

Original: early spellcheckers.

Canonical: list of forbidden passwords

Modern: network routers.
- Limited memory, need to be super-fast

Tim Roughgarden

Bloom Filter: Under the Hood

Ingredients: 1) array of n bits ($So \frac{n}{|S|} = \# \ of \ bits \ per \ object \ in \ data \ set \ S$)

2) k hash functions $h_1, \ldots, h_k$ (k = small constant)

Insert(x): for i = 1,2,...,k (whether or not bit already set ot 1)

set $A[h_i(x)]=1$

Lookup(x): return TRUE $\Leftrightarrow$ $A[h_i(x)] = 1$ for every I = 1,2,....,k.

Note: no false negatives. (if x was inserted, Lookup (x) guaranteed to succeed)

But: false positive if all k $h_i(x)$'s already set to 1 by other insertions.

Heuristic Analysis

Intuition: should be a trade-off between space and error (false positive) probability.

Assume: [not justified] all $h_i(x)$'s uniformly random and independent (across different i's and x's).

Setup: n bits, insert data set S into bloom filter.

Note: for each bit of A, the probability it's been set to I is (under above assumption):

Under the heuristic assumption, what is the probability that a given bit of the bloom filter (the first bit, say) has been set to 1 after the data set S has been inserted?

- ○ $(1 - {}^{1}/_{n})^{k|S|}$ — prob 1^{st} bit = 0
- ○ $1 - (1 - {}^{1}/_{n})^{k|S|}$ — prob 1^{st} bit = 1
- ○ $({}^{1}/_{n})^{|S|}$
- ○ $(1 - {}^{1}/_{n})^{|S|}$

Heuristic Analysis

Intuition: should be a trade-off between space and error (false positive) probability.

Assume: [not justified] all $h_i(x)$'s uniformly random and independent (across different i's and x's).

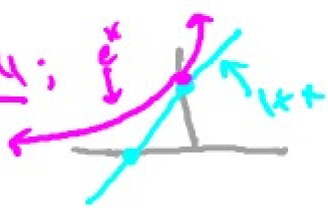
Setup: n bits, insert data set S into bloom filter.

Note: for each bit of A, the probability it's been set to I is (under above assumption):

$$1 - (1 - \frac{1}{n})^{k|S|} \leq 1 - e^{-\frac{k|S|}{n}} = 1 - e^{-\frac{k}{b}}$$

b = # of bits per object ($n/|S|$)

Recall:



Heuristic Analysis (con'd)

Story so far: probability a given bit is 1 is $\leq 1 - e^{\frac{-k}{b}}$

So: under assumption, for x not in S, false positive probability is $\boxed{\leq [1 - e^{\frac{-k}{b}}]^k}$

where b = # of bits per object.

error rank ϵ

How to set k?: for fixed b, ϵ is minimized by setting

Plugging back in: $\epsilon \approx (\frac{1}{2})^{(ln2)b}$ or $b \approx 1.44 log_2 \frac{1}{\epsilon}$

(exponentially small in b)

$k \approx (ln2).b \approx 0.693$

Ex: with b = 8, choose k = 5 or 6 , error probability only approximately 2%.