Design and Analysis of Algorithms I

Data Structures

Bloom Filters

Bloom Filters: Supported Operations

Raison d'être: fast Inserts and Lookups.

Comparison to HashTables:

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

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 $i = 1, 2, \ldots, k$
set $A[h_i(x)] = 1$ (whether or not bit already set to 1)

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

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.

Tim Roughgarden

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

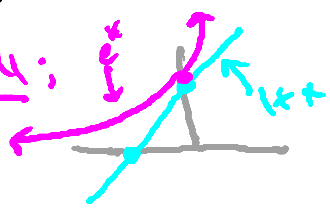
Setup: n bits, insert data set S into bloom filter.

Note: for each bit of A, the probability it's been set to 1 is (under above assumption):

$$1 - \underbrace{(1 - 1/n)^{k|S|}}_{} \le 1 - e^{-\frac{k|S|}{n}} = 1 - e^{-k/b}$$

← b = # of bits per object ($n/|S|$)

Recall:



Heuristic Analysis (con'd)

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

So: under assumption, for $x \notin S$, false positive probability is $\leq \boxed{[1 - e^{-k/b}]^k}$, where b = # of bits per object.

↳ error rate ε

How to set k?: For fixed b, ε is minimized by setting $k \approx (\ln 2) \cdot b$

$\ln 2 \approx .693$

Plugging back in: $\varepsilon \approx \left(\frac{1}{2}\right)^{(\ln 2) b}$ (exponentially small in b)

or $b \approx 1.44 \log_2 \frac{1}{\varepsilon}$

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