April 20, 2009

Data Streams

Let's look at a data streams. A data stream consists of elements $a_1, a_2, \ldots, a_n$ where n is the length of the string, and each element a_i is from the alphabet {1, 2, …, m}. The ones we want to think about are very long data streams consisting of maybe trillions of elements.

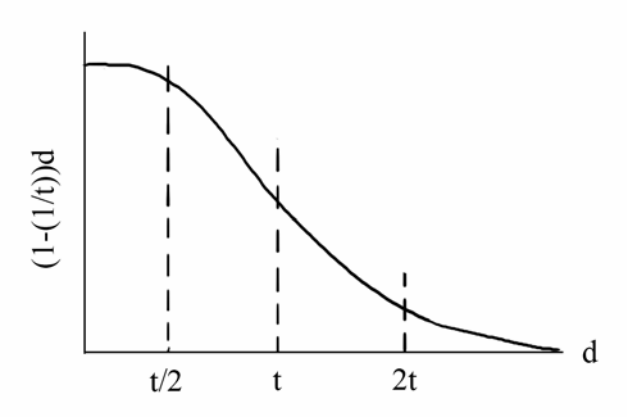
We want to find how many distinct elements d exist in the stream, but this would take log(n) bits, which for a very long data stream may be too many, so we just want to come up with an approximation that uses less space, but is provably within a certain range of the exact answer.

First Approximation Algorithm

1. hash elements of the stream
   $h_i\{1,2,\ldots,m\} \rightarrow \{1,2,\ldots,t\}$
2. test if there exists an element a_i such that $h(a_i) = 1$. If so then assume we have seen approximately t or more distinct elements.

Now we would like to prove that with high probability this algorithm yields the correct answer.

What is the probability that something gets mapped to 1?
(Note that (1-(1/t))^d is the probability of d elements selected uniformly at random all **not** being mapped to 1.)



$$\left(1-\frac{1}{t}\right)^{\frac{1}{2}t} = \frac{1}{\sqrt{e}}$$

$$\left(1-\frac{1}{t}\right)^{t} = \frac{1}{e}$$

$$\left(1-\frac{1}{t}\right)^{2t} = \frac{1}{e^2}$$

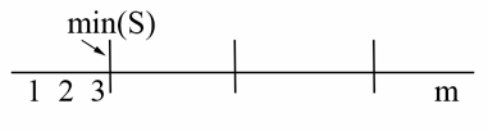
The top line is the probability that none of t/2 elements (chosen as stated above) map to 1. The second line is the probability that none of t elements map to 1. The third line is the probability that none of 2t elements map to 1.

Second Approximation Algorithm

Let S be a subset of elements drawn uniformly at random from the set {1, 2, …, m}

(Note: |S| << m, to prevent hash collisions)

We expect the elements of S to lie approximately evenly spaced on the number line:



|S| = the number of lines (each representing an element of S), each with spacing $\frac{m}{|S|-1}$.

We can then see that min(S) ≈ $\frac{m}{|S|-1}$, so based on the minimum value we draw, we should be able to tell approximately how many values were drawn.

We encounter 2 problems with this model:

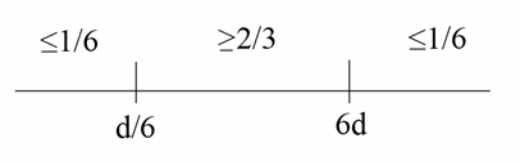
Problem 1: what if the elements are not drawn uniformly at random?

We solve this problem by using a hash function.

Problem 2: How far off are we with the selected positions, i.e. how far off is the min?

To solve problem 2, we introduced a lemma:

Lemma: with probability of at least 2/3, $\frac{d}{6} \le \frac{M}{\min(S)} \le 6d$ (M is used here instead of m) where d is the number of elements



Proof:

First, we want to show that $prob\left[\frac{M}{\min(S)} \le 6d\right] \ge \frac{1}{6}$

$$prob\left[\frac{M}{\min(S)} \le 6d\right] = prob\left[\min(S) \ge \frac{M}{6d}\right] = prob\left[\forall a_i \quad h(a_i) \ge \frac{M}{6d}\right]$$

We also define an indicator variable: z_i = { 1 if $h(a_i) < m/6d$, or 0 otherwise

$Z = \Sigma z_i$

$$prob(z_i = 1) = \frac{1}{6d}$$

$$E(z_i) = \frac{1}{6d}$$

$$E(z) = \frac{1}{6}$$

We know that z is at least 1, therefore:

$$prob\left[\frac{M}{\min S} \geq 6d\right] = prob\left[\min S \leq \frac{M}{6d}\right] = prob\left[\exists a_i \quad h(a_i) \leq \frac{M}{6d}\right]$$

$$\leq prob(z \geq 1) \leq prob(z \geq 6E(z)) \leq \frac{1}{6}$$

(by Markov inequality)

Secondly, we want to prove that: $prob\left[\frac{M}{\min S} \leq \frac{d}{6}\right] \geq \frac{1}{6}$

We define $y_i = 1$ if $h(a_i) \leq \frac{6M}{d} = \frac{M}{d/6}$ or 0 otherwise

$Y = \Sigma y_i$