COMP 6730 Advanced Database Systems HW2

October 2016

1 Problem 1

Exercise 4.3.1: For the situation of our running example (8 billion bits, 1 billion members of the set S), calculate the false-positive rate if we use three hash functions? What if we use four hash functions?

Answer: The false positive for this problem in general is: $(1 - e^{-km/n})^k$, where $m = |S|$, $n = |B|$, k is the number of hash function.

(1) If there're three hash functions, then:

$$\begin{aligned} false-positive \quad rate &= (1 - e^{-km/n})^k \\ &= (1 - e^{-3/8})^3 \\ &\approx 0.031 \\ &= 3.1\% \end{aligned}$$

(2) If there're four hash functions, then:

$$\begin{aligned} false-positive \quad rate &= (1 - e^{-km/n})^k \\ &= (1 - e^{-4/8})^4 \\ &\approx 0.024 \\ &= 2.4\% \end{aligned}$$

2 Problem 2

Exercise 4.4.1: Suppose our stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Our hash functions will all be of the form h(x)=ax+b mod 32 for some a and b. You should treat the result as a 5-bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is:

(a) h(x)=2x+1 mod 32.
(b) h(x)=3x+7 mod 32.
(c) h(x) = 4x mod 32.

Answer:

(a) when $h(x) = (2x + 1) \quad mod \quad 32$, for the integers $\{3, 1, 4, 1, 5, 9, 2, 6, 5\}$, the value of the hash function is $\{7, 3, 9, 3, 11, 19, 5, 13, 11\}$, for each result, we need the binary representation and count the trailing zeros. For example, $7 = 0111$, the trailing zero is 0. So we can get the trailing zeros for this stream is $\{0, 0, 0, 0, 0, 0, 0, 0, 0\}$, so the $r(a) = 0$, and the $R = max\{r(a)\} = 0$, and estimated distinct elements = $2^R = \mathbf{1}$.

(b) the hash function value for the stream is $\{16, 10, 19, 10, 22, 2, 13, 25, 22\}$, and the trailing zeros is $\{4, 1, 0, 1, 1, 1, 0, 0, 1\}$, so the max value is $R = maxr(a) = 4$, and the estimated distinct elements = $2^R = 2^4$=**16**.

(c) the hash function value for the stream is $\{12, 4, 16, 4, 20, 4, 8, 24, 20\}$, and the trailing zeros is $\{2, 2, 4, 2, 2, 2, 3, 3, 2\}$, so the max value is $R = maxr(a) = 4$, and the estimated distinct elements = $2^R = 2^4$=**16**

3 Problem 3

Exercise 4.6.1: Suppose the window is as shown in Fig. 4.2. Estimate the number of 1s the the last k positions, for k = (a) 5 (b) 15. In each case, how far off the correct value is your estimate?

Answer:

(a) For k = 5, first we need to decide how many windows it covered, the sequence is {...10110}, so it covers two windows of size 1, and covers part of one windows with size 2, so the estimated count should be $2 + \frac{1}{2} * 2 = \mathbf{3}$, the real 1s in the last k bits are **3**, so there is no difference between the estimation and the correct value.

(b) For k = 15, we also need to count how many windows it covered, the sequence is {...101110110010110}, so it covers two windows of size 1, one window of size 2, one window of size 4, and part of the window of size 8, the partial cover including the zeros preceding has nearly covered the half of size 4, so it should be $\frac{1}{2} * 4 = 2$. So the estimation should be $1+1+2+4+2 = \mathbf{10}$, and the actual 1s in the last k bits are **9**, so there's a distance of **1** between my estimate and the correct value.