COMP90056 Stream Computing and Applications

Assignment 1, 2019 Sem2

Xiuge Chen 961392

Question a

To store a subset S, with $|S| = m$, of a universe $U = [n]$, initialise a bitmap of width r and choose a single hash function drawn from a 2-universal hash family. To achieve false positive rate $FP \leq \epsilon$, where $\epsilon > 0$, the width r of the Bloom filter bitmap should be initialised as at least $-\frac{m}{ln(1-\epsilon)}$

Pf:

Since the hash function h used here are drawn from a 2-universal hash family, thus $Pr(h(i) = h(j)) \leq \frac{1}{r}, \forall i \neq j \in U$. We all know that Bloom filter only makes false positive error. That is if $x \in S$, Bloom filter guarantees to report yes on the query of x, however, if $x \notin S$, Bloom filter might report yes on the query of x if $\exists y \in S, h(y) = h(x)$. Therefore, to analysis the error probability of the above Bloom filter, it is suffice to consider only false positive error, the situation when query on $x \notin S$ but report yes. Let p_{ij} be the probability that $h(i) = h(j)$, since for each pair of i, j, event $(h(i) = h(j))$ are independent, it is easy to show that for $x \notin S$, the probability of correctly report no is:

$$\begin{aligned} \mathcal{P}(report\ no\ on\ x \notin S) &= \mathcal{P}((h(x) \neq h(s_1)) \wedge (h(x) \neq h(s_2)) \cdots \wedge (h(x) \neq h(s_m))) \\ &= \mathcal{P}(h(x) \neq h(s_1)) \wedge \mathcal{P}(h(x) \neq h(s_2)) \cdots \wedge \mathcal{P}(h(x) \neq h(s_m)) \\ &= (1 - p_{xs_1}) \times (1 - p_{xs_2}) \times \cdots \times (1 - p_{xs_m}) \\ &= \Pi_{i=1}^{m}(1 - p_{xs_i}) \approx \Pi_{i=1}^{m}(e^{-p_{xs_i}}) \end{aligned} \tag{1}$$

Thus the probability ϵ of mistakenly report yes (false positive error) on $x \notin S$ is:

$$\begin{aligned} \epsilon &= \mathcal{P}(report\ yes\ on\ x \notin S) = 1 - \mathcal{P}(report\ no\ on\ x \notin S) \\ &\approx 1 - \Pi_{i=1}^{m}(e^{-p_{xs_i}}) \end{aligned} \tag{2}$$

Since $p_{xs_i} \leq \frac{1}{r}$ holds for all p_{xs_i}, we have:

$$\begin{gathered} \epsilon \approx 1 - \Pi_{i=1}^{m}(e^{-p_{xs_i}}) \leq 1 - \Pi_{i=1}^{m}(e^{-\frac{1}{r}}) = 1 - e^{-\frac{m}{r}} \\ e^{-\frac{m}{r}} \leq 1 - \epsilon \\ ln(e^{-\frac{m}{r}}) \leq ln(1 - \epsilon) \\ -\frac{m}{r} \leq ln(1 - \epsilon) \\ r \geq -\frac{m}{ln(1-\epsilon)} \end{gathered}$$

Therefore, in order to keep false positive rate $FP \leq \epsilon$, where $\epsilon > 0$, the width r of the Bloom filter bitmap should be initialised as at least $-\frac{m}{ln(1-\epsilon)}$.