第 12 讲: Hashing 方法

姓名： 林凡琪　　**学号：** 211240042

评分： _____　　**评阅：** _____

2022 年 5 月 11 日

请独立完成作业，不得抄袭。
若得到他人帮助, 请致谢。
若参考了其它资料，请给出引用。
鼓励讨论，但需独立书写解题过程。

1　作业 (必做部分)

题目 1 (CS 5.5-8 (a, b, c))
Suppose you hash n items into k locations.
a. What is the probability that all n items hash to different locations?
b. What is the probability that the ith item is the first collision?
c. What is the expected number of items you hash until the first collision?

解答：
(a) 如果 k<n, 那么概率为 0
否则如果 k>=n, 那么 $P = \frac{k^n}{k^n}$
(b) 如果 $i - 1 > k$ 那么概率为 0;
否则如果 $i - 1 \leq k$, 那么 P[i-1]=$\frac{k^{i-1}}{k^{i-1}}$
所以 P[i]=$\frac{k^{i-1}}{k^{i-1}} \cdot \frac{i-1}{k}$
(c)$E[X] = \sum_{i=1}^{n}(i-1) \cdot \frac{k^{i-1}}{k^{i-1}} \cdot \frac{i-1}{k}$

题目 2 (TC 11.2-3)

解答：
成功搜索: 没有区别, $\Theta(1+\alpha)(1+)$.
不成功搜索: 更快但仍然是 $\Theta(1+\alpha)(1+)$.
插入: 和成功搜索一样, $\Theta(1+\alpha)(1+)$.
删除: 如果我们使用双向链表，则与以前相同, $\Theta(1)(1)$.

题目 3 (TC 11.3-3)

解答:
我们将证明每个字符串散列到它的数字之和 $\mod 2^p - 1$.
我们将通过对字符串长度的归纳来做到这一点。
Base case
假设字符串是单个字符，那么该字符的值就是 k 的值，然后取 mod m。
Inductive step
令 $w = w_1 w_2$ 其中 $|w_1| \geq 1$ 和 $|w_2| = 1$。假设 $h(w_1) = k_1$。那么，$h(w) = h(w_1)2^p + h(w_2) \mod 2^p - 1 = h(w_1) + h(w_2) \mod 2^p - 1$。所以，由于 $h(w_1)$ 是除了最后一个数字 $\mod m$ 之外的所有数字的总和，我们添加最后一个数字 $\mod m$，就能得到想要的结论。

题目 4 (TC 11.4-3)

解答:
$\alpha = 3/4$,
不成功: $\frac{1}{1-\frac{3}{4}} = 4$ probes, 成功: $\frac{1}{\frac{3}{4}} \ln \frac{1}{1-\frac{3}{4}} \approx 1.848$ probes.
$\alpha = 7/8$
不成功: $\frac{1}{1-\frac{7}{8}} = 8$ probes,
成功: $\frac{1}{\frac{7}{8}} \ln \frac{1}{1-\frac{7}{8}} \approx 2.377$ probes.

题目 5 (TC Problem 11-1)

解答:
(a)X_i:probe 的数量
我们想要计算 $Pr(X_i > k)$
A_j 是第 j 个 probe 发生在已占用的 slot
$X_i > k = A_1 \cap A_2 \cap ... \cap A_k$

$$\begin{aligned} Pr(X_i > k) &= Pr(A_1 \cap A_2 \cap ... \cap A_k) \\ &= Pr(A_1) \cdot Pr(A_2|A_1) \cdot Pr(A_3|A_1 \cap A_2)...Pr(A_k|A_1 \cap A_2 \cap ... \cap A_{k-1}) \\ &= \frac{n}{m} \cdot \frac{n-1}{m-1} ... \frac{n-k+1}{m-k+1} \leq (\frac{n}{m})^k = \alpha^k \leq 2^{-k} \end{aligned}$$

(b)$(X > 2\lg n) = \bigvee_{i=1}^{n} (X_i > 2\lg n)$
$Pr(X > 2\lg n) = Pr(\bigvee_{i=1}^{n} (X_i > 2\lg n)) \leq n \cdot O(\frac{1}{n^2}) = O(1/n)$
(c)

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} i \cdot Pr(X = i) \\ &= \sum_{i=1}^{\lceil 2\lg n \rceil} i \cdot Pr(X = i) + \sum_{i=\lceil 2\lg n \rceil + 1}^{n} Pr(X = i) \\ &\leq \lceil 2\lg n \rceil \sum_{i=1}^{\lceil 2\lg n \rceil} i \cdot Pr(X = i) + n \sum_{i=\lceil 2\lg n \rceil + 1}^{n} Pr(X = i) \\ &= \lceil 2\lg n \rceil Pr[X \leq \lceil 2\lg n \rceil] + nPr(X > \lceil 2\lg n \rceil) \\ &\leq \lceil 2\lg n \rceil + nO(1/n) \\ &= O(\lg n) \end{aligned}$$

题目 6 (TC Problem 11-2)

解答:

(a) 一个特定的密钥以概率 $\frac{1}{n}$ 散列到一个特定的槽，假设我们选择一组特定的 k 个密钥。这些 k 个密钥被插入到相关槽中的概率以及所有其他密钥的概率插入其他地方是 $(\frac{1}{n})^k(\frac{1}{n-1})^{n-k}$ 因为有 $(\frac{n}{k})$ 种方法来选择我们的 k 个密钥，所以我们得到 $Q_k = (\frac{1}{n})^k(\frac{1}{n-1})^{n-k}(\frac{n}{k})$

b. 对于 i = 1, 2, ...,n，令 X_i 是一个随机变量，表示散列到插槽 i 的键的数量，令 A_i 是 X_i = k 的事件，即 k 密钥哈希到插槽 i。我们知道 Pr{A} = Q_k,∴.

P_k = Pr{M = k}

= Pr{$(\frac{max}{1\leq i\leq n}X_i)$ = k}

= Pr{there exists i such that X_i = k and that X_i≤k for i = 1,2,⋯,n}

≤Pr{there exists i such that X_i = k}

= Pr{$A_1 \bigcup A_2 \bigcup \ldots \bigcup A_n$}

≤Pr{A_1} + Pr{A_2} + ⋯+ Pr{A_n}

= nQ_k c. 首先，1-$\frac{1}{n}$ < 1，这意味着 (1 - $\frac{1}{n}$)$^{n-k}$ < 1。其次， $fracn!(n-k)!$ = n(n - 1)(n - 2)...(n - k + 1) < n^k. 使用这些事实和 k! > $(\frac{k}{e})^k$, 我们有

Q_k < n^kk!(n - k)!((1 - $\frac{1}{n}$)$^{n-k}$ < 1)

<k^k(k! > $(\frac{k}{e})^k$)

d. 当 n = 2 时，lglgn = 0，所以准确地说，我们需要假设 n≥3

对于任何 k，我们都有 Q_k <$\frac{e^k}{k^k}$，特别是，这个不等式对 k_0 成立。因此足以证明 $\frac{e^{k_0}}{k_0^{k_0}}$, 或 n^3 < $\frac{k_0^{k_0}}{e_0^{k_0}}$ ∴3lgn = lglgn(lgc + glgn - lglglgn - lge)

3 < lglgn(lgc + glgn - lglglgn - lge)

= c$(\frac{lgc-lgelglglgn}{1+lglgn-lglgn})$

所以 x = c$(\frac{lgc-lgelglglgn}{1+lglgn-lglgn})$ 我们需要证明存在一个常数 c > 1 使得 3 <cx

注意到 $\lim_{n\to\infty}$x = 1，我们看到存在 n_0 使得 x≥$\frac{1}{2}$ 对于所有 n ≥n_0, 因此任何常数 c > 6 适用于 n ≥n_0

我们处理较小的 n 值，特别是 3≤n < n_0-，如下所示。由于 n 被约束为整数，因此在 3 ≤n < n_0。我们可以为每个这样的 n 值计算表达式 x，并确定 c 的值，对于所有 n 值，3 < cx。我们使用的 c 的最终值是 6 中的较大者，它适用于所有 n > n_0，我们为范围 3 ≤n < n_0 选择的 c 的最大值

因此，我们知道 Q_{k_0} < $\frac{1}{n^3}$

选择 k = k_0 给出 P_{k_0}≤nQ_{k_0} < n, $(\frac{1}{n^3})$ = $\frac{1}{n^2}$。

∴Q_k < $\frac{e^k}{k^k}$

≤$\frac{e^{k_0}}{k^{k_0}}$

< $\frac{1}{n^3}$ for k ≥k_0

e.E[M] = ∑k * Pr{M = k}

≤∑k_0Pr{M = k} + ∑n * Pr{M = k}

≤k_0∑Pr{M = k} + n∑Pr{M = k}

= k_0 * Pr{M≤k_0} + n * Pr{M > k_0}

因为 k_0 =$\frac{clgn}{lglgn}$

Pr{M > k_0} = ∑Pr{M = k}

=∑P_k

<∑$\frac{1}{n^2}$

< n * $(\frac{1}{n^2})$

= $\frac{1}{n}$

∴E[M]≤k_0 * 1 + n $(\frac{1}{n})$

= k_0 + 1

$= O(\frac{lgn}{lglgn})$

---

2 作业 (选做部分)

题目 1 (TC 11.2-6)

解答:

---

3 Open Topics

Open Topics 1 (Perfect Hashing)
介绍 Perfect hashing。
参考资料:

- Section 11.5 of CLRS

Open Topics 2 (Bloom filter)
介绍 Bloom filter。
参考资料:

- Bloom filter @ wiki

4 反馈