问题与反馈

2014-6-19

8. When we hash n items into k locations, what is the probability that all n items hash to different locations? What is the probability that the ith item is the first collision? What is the expected number of items we must hash until the first collision? Use a computer program or spreadsheet to compute the expected number of items hashed into a hash table until the first collision with $k = 20$ and with $k = 100$.

What is the probability that all n items are mapped to different locations?

Obviously, for $k < n$ it's zero. For $k \geq n$, there are $k(k-1)\cdots(k-n-1)$ ways to map all the items to different locations, and k^n total possible mappings, so the probability is

$$\frac{k!}{(k-n)!k^n} = \prod_{j=1}^{n-1}\left(1 - \frac{j}{k}\right).$$

What is the probability that the i^{th} item gives the first collision?

Again, for $i > k$, it's obviously zero. For $i \leq k$, it's the probability that the first $i-1$ items are all mapped to different locations, and then the i^{th} item is mapped to one of the $i-1$ occupied locations, which is

$$\frac{(i-1)k!}{(k-i+1)!k^i} = \frac{i-1}{k} \prod_{j=1}^{i-2} \left(1 - \frac{j}{k}\right).$$

What is the expected number of items you hash until the first collision?

- **Let C be the number of the item that causes the first collision**

$$
\begin{aligned}
E[C] &= \sum_{c=1}^{k+1} P(C = c)c \\
&= \sum_{c=0}^{k} P(C > c) \\
&= \sum_{c=0}^{k} \frac{k!}{(k-c)!k^c} = 1 + \sum_{c=1}^{k} \prod_{j=0}^{c-1} \left(1 - \frac{j}{k}\right).
\end{aligned}
$$

The identity we used in going from the first line to the second (writing an expectation in terms of the cumulative probability distribution) can be useful more generally.

For large k, we can get a simple approximate expression by using that $1 - j/k \approx e^{-j/k}$:

$$
\begin{aligned}
E[C] &\approx 1 + \sum_{c=1}^{k} \prod_{j=0}^{c-1} \exp\left[-\frac{j}{k}\right] \\
&= 1 + \sum_{c=1}^{k} \exp\left[-\frac{1}{k} \sum_{j=0}^{c-1} j\right] \\
&= \sum_{c=0}^{k} \exp\left[-\frac{c(c-1)}{2k}\right] \\
&\approx \int_{0}^{\infty} dx \, \exp\left[-\frac{x^2}{2k}\right] \\
&= \sqrt{\pi k/2}.
\end{aligned}
$$

11.2-3

Professor Marley hypothesizes that he can obtain substantial performance gains by modifying the chaining scheme to keep each list in sorted order. How does the professor's modification affect the running time for successful searches, unsuccessful searches, insertions, and deletions?

Successful searches:

$\Theta(1 + \alpha)$, which is identical to the original running time. The element we search for is equally likely to be any of the elements in the hash table, and the proof of the running time for successful searches is similar to what we did in the lecture.

Unsuccessful searches:

1/2 of the original running time, but still $\Theta(1 + \alpha)$, if we simply assume that the probability that one element's value falls between two consecutive elements in the hash slot is uniformly distributed. This is because the value of the element we search for is equally likely to fall between any consecutive elements in the hash slot, and once we find a larger value, we can stop searching. Thus, the running time for unsuccessful searches is a half of the original running time. Its proof is similar to what we did in the lecture.

Insertions:

$\Theta(1 + \alpha)$, compared to the original running time of $\Theta(1)$. This is because we need to find the right location instead of the head to insert the element so that the list remains sorted. The operation of insertions is similar to the operation of unsuccessful searches in this case.

Deletions:

$\Theta(1 + \alpha)$, same as successful searches.

11-1 Longest-probe bound for hashing

Suppose that we use an open-addressed hash table of size m to store $n \leq m/2$ items.

a. Assuming uniform hashing, show that for $i = 1, 2, \ldots, n$, the probability is at most 2^{-k} that the ith insertion requires strictly more than k probes.

Since we assume uniform hashing, we can use the same observation as is used in Corollary 11.7: that inserting a key entails an unsuccessful search followed by placing the key into the first empty slot found. As in the proof of Theorem 11.6, if we let X be the random variable denoting the number of probes in an unsuccessful search, then $\Pr\{X \geq i\} \leq \alpha^{i-1}$. Since $n \leq m/2$, we have $\alpha \leq 1/2$. Letting $i = k + 1$, we have $\Pr\{X > k\} = \Pr\{X \geq k + 1\} \leq (1/2)^{(k+1)-1} = 2^{-k}$.

b. Show that for $i = 1, 2, \ldots, n$, the probability is $O(1/n^2)$ that the ith insertion requires more than $2 \lg n$ probes.

Substituting $k = 2 \lg n$ into the statement of part (a) yields that the probability that the ith insertion requires more than $k = 2 \lg n$ probes is at most $2^{-2 \lg n} = (2^{\lg n})^{-2} = n^{-2} = 1/n^2$.

Let the random variable X_i denote the number of probes required by the ith insertion. You have shown in part (b) that $\Pr\{X_i > 2\lg n\} = O(1/n^2)$. Let the random variable $X = \max_{1 \le i \le n} X_i$ denote the maximum number of probes required by any of the n insertions.

c. Show that $\Pr\{X > 2\lg n\} = O(1/n)$.

Let the event A be $X > 2\lg n$, and for $i = 1, 2, \ldots, n$, let the event A_i be $X_i > 2\lg n$. In part (b), we showed that $\Pr\{A_i\} \le 1/n^2$ for $i = 1, 2, \ldots, n$. From how we defined these events, $A = A_1 \cup A_2 \cup \cdots \cup A_n$. Using Boole's inequality, (C.18), we have

$$\begin{aligned}
\Pr\{A\} &\le \Pr\{A_1\} + \Pr\{A_2\} + \cdots + \Pr\{A_n\} \\
&\le n \cdot \frac{1}{n^2} \\
&= 1/n \,.
\end{aligned}$$

d. Show that the expected length $\mathrm{E}[X]$ of the longest probe sequence is $O(\lg n)$.

We use the definition of expectation and break the sum into two parts:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{k=1}^{n} k \cdot \Pr\{X = k\} \\
&= \sum_{k=1}^{\lceil 2\lg n \rceil} k \cdot \Pr\{X = k\} + \sum_{k=\lceil 2\lg n \rceil + 1}^{n} k \cdot \Pr\{X = k\} \\
&\le \sum_{k=1}^{\lceil 2\lg n \rceil} \lceil 2\lg n \rceil \cdot \Pr\{X = k\} + \sum_{k=\lceil 2\lg n \rceil + 1}^{n} n \cdot \Pr\{X = k\} \\
&= \lceil 2\lg n \rceil \sum_{k=1}^{\lceil 2\lg n \rceil} \Pr\{X = k\} + n \sum_{k=\lceil 2\lg n \rceil + 1}^{n} \Pr\{X = k\} \; .
\end{aligned}$$

Since X takes on exactly one value, we have that $\sum_{k=1}^{\lceil 2\lg n \rceil} \Pr\{X = k\} = \Pr\{X \le \lceil 2\lg n \rceil\} \le 1$ and $\sum_{k=\lceil 2\lg n \rceil + 1}^{n} \Pr\{X = k\} \le \Pr\{X > 2\lg n\} \le 1/n$, by part (c). Therefore,

$$\begin{aligned}
\mathrm{E}[X] &\le \lceil 2\lg n \rceil \cdot 1 + n \cdot (1/n) \\
&= \lceil 2\lg n \rceil + 1 \\
&= O(\lg n) \; .
\end{aligned}$$

11-2 *Slot-size bound for chaining*

Suppose that we have a hash table with n slots, with collisions resolved by chaining, and suppose that n keys are inserted into the table. Each key is equally likely to be hashed to each slot. Let M be the maximum number of keys in any slot after all the keys have been inserted. Your mission is to prove an $O(\lg n / \lg \lg n)$ upper bound on $\mathrm{E}[M]$, the expected value of M.

a. Argue that the probability Q_k that exactly k keys hash to a particular slot is given by

$$Q_k = \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \binom{n}{k} .$$

A particular key is hashed to a particular slot with probability $1/n$. Suppose we select a specific set of k keys. The probability that these k keys are inserted into the slot in question and that all other keys are inserted elsewhere is

$$\left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} .$$

Since there are $\binom{n}{k}$ ways to choose our k keys, we get

$$Q_k = \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \binom{n}{k} .$$

b. Let P_k be the probability that $M = k$, that is, the probability that the slot containing the most keys contains k keys. Show that $P_k \leq nQ_k$.

For $i = 1, 2, \ldots, n$, let X_i be a random variable denoting the number of keys that hash to slot i, and let A_i be the event that $X_i = k$, i.e., that exactly k keys hash to slot i. From part (a), we have $\Pr\{A\} = Q_k$. Then,

$$\begin{aligned}
P_k &= \Pr\{M = k\} \\
&= \Pr\left\{\left(\max_{1 \leq i \leq n} X_i\right) = k\right\} \\
&= \Pr\{\text{there exists } i \text{ such that } X_i = k \text{ and that } X_i \leq k \text{ for } i = 1, 2, \ldots, n\} \\
&\leq \Pr\{\text{there exists } i \text{ such that } X_i = k\} \\
&= \Pr\{A_1 \cup A_2 \cup \cdots \cup A_n\} \\
&\leq \Pr\{A_1\} + \Pr\{A_2\} + \cdots + \Pr\{A_n\} \qquad \text{(by inequality (C.18))} \\
&= nQ_k \,.
\end{aligned}$$

c. Use Stirling's approximation, equation (3.18), to show that $Q_k < e^k / k^k$.

We start by showing two facts. First, $1 - 1/n < 1$, which implies $(1 - 1/n)^{n-k} < 1$. Second, $n!/(n-k)! = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) < n^k$. Using these facts, along with the simplification $k! > (k/e)^k$ of equation (3.17), we have

$$
\begin{aligned}
Q_k &= \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \frac{n!}{k!(n-k)!} \\
&< \frac{n!}{n^k k!(n-k)!} && ((1 - 1/n)^{n-k} < 1) \\
&< \frac{1}{k!} && (n!/(n-k)! < n^k) \\
&< \frac{e^k}{k^k} && (k! > (k/e)^k) \ .
\end{aligned}
$$

d. Show that there exists a constant $c > 1$ such that $Q_{k_0} < 1/n^3$ for $k_0 = c \lg n / \lg \lg n$. Conclude that $P_k < 1/n^2$ for $k \geq k_0 = c \lg n / \lg \lg n$.

Notice that when $n = 2$, $\lg \lg n = 0$, so to be precise, we need to assume that $n \geq 3$.

In part (c), we showed that $Q_k < e^k / k^k$ for any k; in particular, this inequality holds for k_0. Thus, it suffices to show that $e^{k_0} / {k_0}^{k_0} < 1/n^3$ or, equivalently, that $n^3 < {k_0}^{k_0} / e^{k_0}$.

Taking logarithms of both sides gives an equivalent condition:

$$\begin{aligned} 3 \lg n &< k_0 (\lg k_0 - \lg e) \\ &= \frac{c \lg n}{\lg \lg n} (\lg c + \lg \lg n - \lg \lg \lg n - \lg e) \ . \end{aligned}$$

Dividing both sides by $\lg n$ gives the condition

$$\begin{aligned} 3 &< \frac{c}{\lg \lg n} (\lg c + \lg \lg n - \lg \lg \lg n - \lg e) \\ &= c \left(1 + \frac{\lg c - \lg e}{\lg \lg n} - \frac{\lg \lg \lg n}{\lg \lg n} \right) \ . \end{aligned}$$

Let x be the last expression in parentheses:

$$x = \left(1 + \frac{\lg c - \lg e}{\lg\lg n} - \frac{\lg\lg\lg n}{\lg\lg n}\right) .$$

We need to show that there exists a constant $c > 1$ such that $3 < cx$.

Noting that $\lim_{n\to\infty} x = 1$, we see that there exists n_0 such that $x \geq 1/2$ for all $n \geq n_0$. Thus, any constant $c > 6$ works for $n \geq n_0$.

We handle smaller values of n—in particular, $3 \leq n < n_0$—as follows. Since n is constrained to be an integer, there are a finite number of n in the range $3 \leq n < n_0$. We can evaluate the expression x for each such value of n and determine a value of c for which $3 < cx$ for all values of n. The final value of c that we use is the larger of

- 6, which works for all $n \geq n_0$, and
- $\max_{3\leq n<n_0} \{c : 3 < cx\}$, i.e., the largest value of c that we chose for the range $3 \leq n < n_0$.

Thus, we have shown that $Q_{k_0} < 1/n^3$, as desired.

To see that $P_k < 1/n^2$ for $k \geq k_0$, we observe that by part (b), $P_k \leq n\,Q_k$ for all k. Choosing $k = k_0$ gives $P_{k_0} \leq n\,Q_{k_0} < n \cdot (1/n^3) = 1/n^2$. For $k > k_0$, we will show that we can pick the constant c such that $Q_k < 1/n^3$ for all $k \geq k_0$, and thus conclude that $P_k < 1/n^2$ for all $k \geq k_0$.

To pick c as required, we let c be large enough that $k_0 > 3 > e$. Then $e/k < 1$ for all $k \geq k_0$, and so e^k/k^k decreases as k increases. Thus,

$$\begin{aligned} Q_k &< e^k/k^k \\ &\leq e^{k_0}/k^{k_0} \\ &< 1/n^3 \end{aligned}$$

for $k \geq k_0$.

e. Argue that

$$\mathrm{E}[M] \le \Pr\left\{M > \frac{c \lg n}{\lg \lg n}\right\} \cdot n + \Pr\left\{M \le \frac{c \lg n}{\lg \lg n}\right\} \cdot \frac{c \lg n}{\lg \lg n} .$$

Conclude that $\mathrm{E}[M] = O(\lg n / \lg \lg n)$.

The expectation of M is

$$\begin{aligned}
\mathrm{E}[M] &= \sum_{k=0}^{n} k \cdot \Pr\{M = k\} \\
&= \sum_{k=0}^{k_0} k \cdot \Pr\{M = k\} + \sum_{k=k_0+1}^{n} k \cdot \Pr\{M = k\} \\
&\le \sum_{k=0}^{k_0} k_0 \cdot \Pr\{M = k\} + \sum_{k=k_0+1}^{n} n \cdot \Pr\{M = k\} \\
&\le k_0 \sum_{k=0}^{k_0} \Pr\{M = k\} + n \sum_{k=k_0+1}^{n} \Pr\{M = k\} \\
&= k_0 \cdot \Pr\{M \le k_0\} + n \cdot \Pr\{M > k_0\} ,
\end{aligned}$$

which is what we needed to show, since $k_0 = c \lg n / \lg \lg n$.

To show that $\mathrm{E}[M] = O(\lg n / \lg\lg n)$, note that $\Pr\{M \le k_0\} \le 1$ and

$$\begin{aligned}
\Pr\{M > k_0\} &= \sum_{k=k_0+1}^{n} \Pr\{M = k\} \\
&= \sum_{k=k_0+1}^{n} P_k \\
&< \sum_{k=k_0+1}^{n} 1/n^2 \qquad \text{(by part (d))} \\
&< n \cdot (1/n^2) \\
&= 1/n \, .
\end{aligned}$$

We conclude that

$$\begin{aligned}
\mathrm{E}[M] &\le k_0 \cdot 1 + n \cdot (1/n) \\
&= k_0 + 1 \\
&= O(\lg n / \lg\lg n) \, .
\end{aligned}$$