Load balancing

Huy L. Nguyễn

1 Perfect hashing

Last time, we defined universal hashing and constructed a universal hash family. In this section, we will use such a hash family to construct a collision-free hash table for a *static* dataset i.e. no insertion and deletion. First, we need to determine how large the table size m needs to be to get no collision from n keys. Let $X_{i,j}$ be the indicator random variable of whether $h(i) = h(j)$. The expected number of collision is

$$\mathbb{E}[\sum_{i \neq j} X_{i,j}] \leq \binom{n}{2} \frac{1}{m}$$

For $m \geq n^2$, the expected number of collision is at most $1/2$. Thus, by Markov's inequality, the probability that we have a collision is at most $1/2$. However, this size is prohibitively large. Instead, we will use a two layer design to reduce the size.

First we have a hash table of size n. In each of the cell, instead of storing all elements hashed there in a linked list, we use a secondary hash table instead. If there are s_i elements with hash value i, we will use a hash table of size s_i^2 to store them. The total space of this design is $\sum_{i=1}^n s_i^2$. Note that $\sum_{i=1}^n s_i(s_i - 1)/2$ is the number of collisions that we just calculated. Therefore,

$$\mathbb{E}\left[\sum_{i=1}^n s_i^2\right] = \mathbb{E}\left[\sum_{i=1}^n s_i(s_i - 1) + \sum_{i=1}^n s_i\right] = \frac{n(n-1)}{n} + n = 2n - 1$$

We used the fact that $\sum_{i=1}^n s_i$ is simply the total number of keys, which is n.

2 Load balancing

Next we will consider a related problem to hashing, namely load balancing. We have n balls and n bins and we put the balls randomly into the bins. We would like to understand the maximum number of balls inside the same bin. Notice that this is the same as analyzing the number of elements with the same hash value when we hash n keys into a table of size n. For simplicity, we will only analyze the fully random assignment and not specific pseudo-random hash families.

2.1 Direct analysis

Question: What is the expected number of balls in bin 1?

Note that while the expectation is 1, the maximum over all bins can be a lot higher. Let's analyze the probability that bin 1 gets at least k balls. We will make use of a simple but useful technique: *the union bound*. The union bound states that for any two events E_1, E_2, we have

$$\Pr[E_1 \vee E_2] \leq \Pr[E_1] + \Pr[E_2]$$

We consider different choices for the first k balls that land in bin 1. There are $\binom{n}{k}$ choices. For a fixed set of k balls, the probability that they all land in bin 1 is $\frac{1}{n^k}$. Therefore, by the union bound,

$$\Pr[\text{at least } k \text{ balls land in bin } 1] \le \binom{n}{k} \cdot \frac{1}{n^k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \frac{1}{n^k} \le \frac{1}{k!}$$

A very useful approximation for the factorial function is Sterling's formula:

$$k! = \sqrt{2\pi k}\left(\frac{k}{e}\right)^k (1 + O(1/k))$$

Question: How do choose k (asymptotically) so that $\frac{1}{k!} \le \frac{1}{n^3}$?

If we choose $k = \Theta\left(\frac{\log n}{\log\log n}\right)$ then $\frac{1}{k!} \le \frac{1}{n^3}$.

Thus, the probability that bin 1 gets at least k balls is at most $\frac{1}{n^3}$. By the union bound over all bins, the probability that any bin gets at least k balls is at most $\frac{1}{n^2}$. We can conclude that with probability at least $1 - 1/n^2$, the maximum load is $O\left(\frac{\log n}{\log\log n}\right)$.

Question: What do you get from Markov's inequality?

2.2 Chebyshev's inequality

Let X be a random variable with expected value μ and variance σ^2. For any number $a > 0$, Chebyshev's inequality states that

$$\Pr[|X - \mu| \ge a] \le \frac{\sigma^2}{a^2}$$

Question: use Chebyshev's inequality to bound the probability that we have at least $1 + 2\sqrt{n}$ balls in bin 1.

Let X_i be the indicator variable of whether ball i lands in bin 1. We already have $\mathbb{E}[X_i] = \frac{1}{n}$.

Let's analyze the variance. By definition:

$$\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \frac{1}{n} - \frac{1}{n^2}$$

Let $X = X_1 + X_2 + \ldots + X_n$.

Recall the following fact: for any two independent random variables A, B, we have $\mathrm{Var}[A+B] = \mathrm{Var}[A] + \mathrm{Var}[B]$. Because $X_1, X_2, \ldots, X_n$ are independent, we have

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = 1 - \frac{1}{n}$$

Plugging this variance into Chebyshev's inequality, we obtain the answer to the question.