Analys of the insertion time in Cuckoo Hashing

Based on the notes for undergraduates by R. PAGH

Class of ADVANCED ALGORITHMS

UNIVERSITY OF PISA – Roberto Grossi

Academic year 2016-2017

Preliminaries

We choose $h_1(x), h_2(x)$ independently and uniformly at random from the universal family

$$\{((ax+b) \bmod p) \bmod m : p > m,\ a \in \mathbb{Z}_p^*,\ b \in \mathbb{Z}_p\}$$

In particular, given $x \in U$ and $i, j \in [m]$, we have

$$\Pr(h_1(x) = i \wedge h_2(x) = j) \leq \frac{\lceil p/m \rceil^2}{(p-1)p} \sim \frac{1}{m^2} \qquad (1)$$

This has a similar proof to that seen for 2-wise independence

Analysis

Consider the set S of keys and the hash functions h_1, h_2

Conceptually build an undirected graph $G = (V, E)$ where $|V| = m$ and $|E| = n$, and

- vertices are in $V = \{0, 1, \dots, m-1\}$ and represent the table positions
- edges are random: $E = \{\{h_1(x), h_2(x)\} : x \in S\}$ ← G is actually a multigraph as two vertices can be connected by multiple edges

Consider the insertion of a new key x, which corresponds to a new edge
$e = \{h_1(x), h_2(x)\}$ in G

One of Three situations may happen:
(i) one of the table positions $h_1(x)$ or $h_2(x)$ is free and x is accomodated there
↳ cost is $O(1)$ time

(ii) the positions are taken, so the insertion follows a path in G throwing out keys till a free position is found: let i be the starting position and j be the ending position in the path [note: i is either $h_1(x)$ or $h_2(x)$]

(iii) as in (ii) but a cycle is traversed in G ⇒ REHASHING

We need to analyze (ii) and (iii) ⟶ $O(1)$ expected time

we assume that $m > 2cn$ for a constant $c > 2$

(2)

Case (ii):

Consider the length of the path from i to j in G

By induction on $\ell \geq 1$, the probability that the path length is ℓ is $\leq \frac{1}{c^{\ell} m}$ (3)

Since the insertion cost is $O(1+\ell)$, we obtain an expected cost of

$$O\Big(1+\underbrace{\sum_{\ell=1}^{n} \ell \cdot \frac{1}{c^{\ell} m}}\Big) = O\Big(1+\frac{1}{m}\Big) = O(1) \text{ time}$$

↳ we use Knuth, Concrete Math, p. 33

$$\sum_{\ell=1}^{n} \ell \left(\frac{1}{c}\right)^{\ell} = \frac{\frac{1}{c} - \frac{n+1}{c^{n+1}} + \frac{n}{c^{n+2}}}{\left(1-\frac{1}{c}\right)^2} < \frac{1}{c} \Big/ \left(1-\frac{1}{c}\right)^2 = \frac{c}{(c-1)^2}, \quad c \neq 1$$

We now prove (3) $\longrightarrow$

BASE CASE: $\ell = 1$

Here $\{i,j\}$ is an edge, thus

$$\Pr(\exists \text{ edge } \{i,j\}) = \sum_{x \in S} \Big[\Pr\Big((h_1(x) = i \wedge h_2(x) = j) \vee (h_1(x) = j \wedge h_2(x) = i) \Big) \Big]$$

$$\leq n \cdot 2 \cdot \frac{1}{m^2} \qquad \text{by } (1)$$

$$\leq \frac{1}{cm} \qquad \text{by } (2)$$

INDUCTIVE CASE ($\ell > 1$):

Here we have a path that goes through a vertex $k \neq i,j$, such that the successor of k in the path is j, i.e. $\{k,j\}$ is an edge:

$$\underbrace{i \rightsquigarrow}_{\text{path of length } \ell-1} \overbrace{k - j}^{\text{edge}}$$

path of length $\ell-1$

on which we apply the inductive hp.

$$\Pr(\exists \text{ path } i \overset{\ell>1}{\rightsquigarrow} j) \le \sum_{k \in V-\{i,j\}} \Pr(\overbrace{\exists \text{ path } i \overset{\ell-1}{\rightsquigarrow} k}^{B} \wedge \overbrace{\exists \text{ edge } \{k,j\}}^{A}) =$$

$\Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B)$

$$= \sum_{k \in V-\{i,j\}} \underbrace{\Pr(\exists \text{ path } i \overset{\ell-1}{\rightsquigarrow} k)}_{\le \frac{1}{c^{\ell-1} m}} \cdot \underbrace{\Pr(\exists \text{ edge } \{k,j\} \mid \exists \text{ path } i \overset{\ell-1}{\rightsquigarrow} k)}_{\le \frac{1}{cm}}$$

by inductive hp.

As in the base case $\ell = 1$, replacing $\sum_{x \in S}$ with $\sum_{x \in S - \{\text{keys in path } i \rightsquigarrow k\}}$

$$< m \frac{1}{c^\ell m^2} = \frac{1}{c^\ell m} \qquad \square$$

CASE (iii)

A cycle appears and thus a REHASHING is performed in $O(n)$ time

$$\Pr(\exists \text{ cycle in } G) = \sum_{i=0}^{m-1} \Pr(\exists \text{ path from } i \text{ to } j=i)$$

$$= \sum_{i=0}^{m-1} \sum_{\ell=1}^{n} \Pr(\exists \text{ path } i \xrightarrow{\ell} j=i)$$ ← it's $\leq \frac{1}{c^{\ell} m}$ as in case (ii)

$$\leq \sum_{i=0}^{m-1} \sum_{\ell=1}^{n} \frac{1}{c^{\ell} m} = \sum_{\ell=1}^{n} \frac{1}{c^{\ell}} = \frac{\frac{1}{c^{n+1}} - 1}{\frac{1}{c} - 1} - 1 < \boxed{\frac{1}{c-1}} = p$$

$$\sum_{\ell=0}^{n} x^{\ell} \underset{x \neq 1}{=} \frac{x^{n+1} - 1}{x - 1}$$

Q.: How many REHASHINGs can occur?

1 with pr $= p$
2 " " p^2
...
i " " p^i

} special case of binomial distribution

Expected number of REHASHINGs is

$$\sum_{i=1}^{\infty} i p^i = O(1)$$ as $p < 1$ with $c > 2$

↑ see case (ii)

Turning back to the insertion: if there is a cycle starting at position i,
$O(1)$ REHASHINGs are performed on average

$$\Pr(\exists \text{ cycle starting from position } i) \leq \Pr(\exists \text{ path from } i \text{ to } j = i)$$

$$= \sum_{\ell=1}^{n} \frac{1}{c^\ell m} < \frac{p}{m}$$

Thus we pay $O(n)$ with probability $< \frac{p}{m}$, giving $O\left(\frac{np}{m}\right) = O(1)$ expected cost.

Note the above analysis is a bit sloppy for educational purposes.
Also the time is amortized expected $O(1)$.

$\square$