Universal Hashing

Lecturer: Phil Klein, Scribe: Rob Hunter

Lecture Date: September 9, 2002

1 Hashing

The goal of hashing is to create a data structure that implements a function f. The function f maps keys (from a set S) to values. We use s to denote $|S|$. For any x such that $f(x) = v$, we would like our implementation to be able to compute v quickly (given x).

1.1 A Simple Implementation

We define a function h_1. The function h_1 maps keys to indices into an array of size r. This array contains a list of key-value pairs. Thus, to extract $f(key)$ (our goal) from h_1, we search through the list returned by $h_1(key)$ until we find the pair (key, v), and then we return v. One problem we may encounter is that the lists may be too long. This will obviously depend on where we insert key-value pairs into the array.

2 Universal Families

We first give the intuition for the journey we are about to embark upon. Instead of using a fixed hash function like h_1, we pull a random hash function out of some (very large) *family* of hash functions.

We define a measure of collision, δ, as follows:

Definition 1 *Given sets X_1 and X_2 and set of functions H, $\delta(X_1, X_2, H)$ is defined to be* $|\{x_1, x_2, h) \in X_1 \times X_2 \times H \ : \ h(x_1) = h(x_2) \text{ and } x_1 \neq x_2\}|$

The following definition (of individual collision) will also be helpful:

Definition 2 *$C^h_{x_1,x_2}$ is defined to be* $\begin{cases} 1 & \text{if } h(x_1) = h(x_2) \\ 0 & \text{otherwise} \end{cases}$

Definition 3 *H is universal means that for every $x_1, x_2 \in S, \delta(\{x_1\}, \{x_2\}, H) \leq \frac{|H|}{r}$, where r is the size of our table.*

Theorem 1 *Given any key x_1, and a universal family of hash functions H, the expected number of collisions with x_1 is less than or equal to $\frac{s-1}{r}$.*

Proof: Let E be the expected number of collisions. Thus,
$E = \sum_{h \in H} [Pr[h \text{ chosen}] \times (\text{number of collisions if } h \text{ is used})]$

$$= \sum_{h \in H} \left[\frac{1}{|H|} \sum_{x_2 \in S, x_2 \neq x_1} C^h_{x_1,x_2} \right] \tag{1}$$

$$= \sum_{x_2 \in S; x_2 \neq x_1} \left[\frac{1}{|H|} \sum_{h \in H} C^h_{x_1,x_2} \right] \tag{2}$$

$$\leq \sum_{x_2 \in S; x_2 \neq x_1} \frac{1}{|H|} \left(\frac{|H|}{r} \right) \tag{3}$$

$$= \frac{s-1}{r} \tag{4}$$

where the inequality follows from the definition of universality. ■

We see immediately that if $s - 1 < r$, then the expected number of collisions is less than 1.

3 An Example of a Universal Family

Chose a prime $p > m$, where m is the size of the total universe of keys. Let $g(x) = x \mod r$.

Definition 4 *For $a, b \in \mathbb{Z}_p$ such that $a \neq 0$, $f_{a,b}$ is defined to be $ax + b \pmod p$ and $h_{a,b}$ is defined to be $g(f_{a,b}(x))$*

We can now define our family of hash functions:

Definition 5 *H is defined to be $\{h_{a,b} : a \neq 0\}$*

Theorem 2 *H is universal.*

Proof: Fix x_1 and x_2 arbitrarily. Count the number of h's such that $h_{a,b}(x_1) = h_{a,b}(x_2)$. To do this, we first examine possibilities for $f_{a,b}$. Let $i, j \in \mathbb{Z}_p$. How many (a, b)'s are there such that $f_{a,b}(x_1) = i$ and $f_{a,b}(x_2) = j$? We can answer this question by solving the two equations simultaneously as follows:

$$ax_1 + b \ (\mod p) = i \tag{5}$$

$$ax_2 + b \ (\mod p) = j \tag{6}$$

Note that the two "variables" here are a and b (all other quantities are fixed). For $i = j$ there are no solutions, and for $i \neq j$ there is exactly one solution—this is because $\mathbb{Z}_p$ is a field.

We now ask the following question: How many pairs (i, j) satisfy the following?

$$i \equiv j \ (\mod r), i \neq j \tag{7}$$

We know that for each pair $(i, j), i \neq j$ that satisfies this equation, there is only one f that yields i with x_1 as an argument and j with x_2 as an argument. Thus we only need to determine the number of such pairs, and this will be the number of (a, b) pairs such that $h_{a,b}(x_1) = h_{a,b}(x_2)$.

A moment's reflection should convince you that the number of satisfying pairs of (7) is $\delta(\mathbb{Z}_p, \mathbb{Z}_p, \{g\})$. The following lemma establishes that this number is upperbounded by $\frac{p(p-1)}{r}$. Since $|H| = p(p-1)$, $\delta(\{x_1\}, \{x_2\}, H) \leq \frac{p(p-1)}{r} = \frac{|H|}{r} \leq \frac{|H|}{s}$. ■

Lemma 1 $\delta(\mathbb{Z}_p, \mathbb{Z}_p, \{g\}) \leq p\left(\frac{p-1}{r}\right)$

Proof: We have p choices for the first variable. Let $k \in \mathbb{Z}_r$. Then $k, k + r, k + 2r, k + 3r, ..., k + \alpha r$ in $\mathbb{Z}_p$, for some α, is the complete list of elements in $\mathbb{Z}_p$ that map to k under g. Assume we select $x \in \mathbb{Z}_p$, and that $g(x) = k$. We now have $\alpha - 1$ choices for the second variable. $\alpha \leq \frac{p-1-k}{r} \leq \frac{p-1}{r}$. ■