CMSC 420: Hash Functions

Justin Wyss-Gallifent

January 6, 2023

1 Introduction

This is a brief overview of hash functions including definitions and an introduction to some of the issues.

2 The Problem

All of our data so far has used keys which are sortable in some way, whether they're integers or points, and which can be put in trees, for example, which can be quickly searched.

However what if the keys for our data are not so clean and tidy? For example keys might be strings or more abstract objects. In cases like this it's difficult to know how to begin.

We can of course use things like dictionaries in Python and assign:

```
city['justin'] = 'washington'
```

But this is obfuscating the problem of exactly how Python is storing this. After all, it's almost certainly not creating some sort of list with index `justin` and it's not even clear what that might mean.

The somewhat obvious approach would be to convert our keys into integers in a reliable manner but this presents its own problems. For example suppose for a string of characters $s = c_n...c_1$ with each $0 \leq c_i \leq 25$ (just the alphabet for now) we could create an index via the function:

$$f(s) = \sum_{i=0}^{n} c_i \cdot 26^i$$

But this rapidly gets out of control because the indices are so massive:

$$f(justin) = 9 \cdot 26^6 + 20 \cdot 26^5 + 18 \cdot 26^4 + 19 \cdot 26^3 + 8 \cdot 26^2 + 13 \cdot 26^1 = 3026434762$$

We need a better way.

3 Hash Functions and Hash Tables

Our solution is in fact to use a function to map our keys to integers as discussed above but in a more elegant way so that the range of the function is managable.

Definition 3.0.1. Given a set of keys K and a table T (think of a simple list with m entries) called a *hash table* A *hash function* is a function:

$$h : K \to \mathbb{Z}_m$$

The idea behind a hash function is that for a key $x \in K$ we somehow associate the key x and its data with the table entry $T[h(x)]$.

Note 3.0.1. A few things to note:

- In practice T should be small so that it is manageable.
- For various reasons (some of which we'll address) ideally we want our hash functions to be "as random as possible", meaning that if we don't know actually know the hash function (for example for an outside observer) then it should appear that each $h(x)$ is randomly assigned and it should appear unclear how $h(x)$ is calculated from x.
- There is no requirement that h be either one-to-one or onto.

Here are some examples which will immediately illustrate the idea although you will probably see the major problem.

Example 3.1. If the set of keys K is $\mathbb{Z}$ and T is a a list with 100 entries then we could define:

$$h : K \to \mathbb{Z}_{100}$$

By:

$$h(x) = x \mod 100$$

Then for example $h(7) = 7$ and $h(132343188) = 88$ and so the key $x = 7$ (and its data) would be stored in $T[7]$ and the key $x = 132343188$ (and its data) would be stored in $T[88]$.

Example 3.2. If the set of keys K is $\mathbb{Z}$ and T is a a list with 100 entries then we could pick an integers a and define:

$$h : K \to \mathbb{Z}_{100}$$

By:

$$h(x) = ax \mod 100$$

Here's a sneakier one:

Example 3.3. If the set of keys K is the set of all possible finite strings of bits then if we choose some fixed positive integers $j > 0$ and divide a key $x = b_n...b_0$ into substrings of length j, padding if necessary, then we can define $h(x)$ as the XOR of the substrings.

For example if $j = 5$ and $x = 100111110101000101$ then we do the following: We split up into groups of 5 bits and pad the right end:

$$10011 \Big| 11101 \Big| 01000 \Big| 101\underbrace{00}_{\text{pad}}$$

The we line them up and XOR the columns:

$$\begin{array}{ccccc} 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 \end{array}$$

So $h(100111110101000101) = 10010_2 = 18$.

Notice that the set of possible outputs is the set of all strings of 5 bits, meaning it consists of all integers between 0 and 31 inclusive. So then our T would a list of length 32 and the key 100111110101000101 and its data would be stored in $T[h(100111110101000101)] = T[18]$.

Our hash function is then:

$$h : \{\text{all finite binary strings}\} \longrightarrow \mathbb{Z}_{32}$$

The problem with both of these is that they have *collisions*, meaning two or more inputs can produce the same output.

Example 3.4. In our first example $h(7) = h(107) = h(207) = ...$

If we have $h(x_1) = h(x_2)$ with $x_1 \neq x_2$ then what exactly do we store in $h(x_1) = h(x_2)$?

The approach here is two-fold:

1. To try to engineer the hash function so as to have as few collisions as possible. This is partly accomplished by ensuring that the hash function "acts randomly".

2. To resolve the collisions after the hash function has done its job.

In addition this should all be as fast as possible.

4 Common Strategy Overview

Two common strategies for resolving collisions are:

1. *Closed addressing* aka *Open Hashing*: Instead of keeping the keys (and data) in the hash table we store them in an auxiliary structure typically pointed to by the entries in the hash table. We'll look at *separate chaining* as an example.

2. *Open addressing* aka *Closed Hashing*: We keep the keys (and data) in the hash table and avoid collisions another way. We'll look at *probing* as an example.

For each of these strategies we'll make some comments on their performance. A full mathematical analysis of hash functions can be quite challenging in part because of the number of possibilities.

5 Separate Chaining

5.1 Overview and Insertion

Separate chaining is a form of *closed addressing*.

When using separate chaining the hash table doesn't contain the data but rather contains pointers to linked lists. That is, for any key x the entry $T[x]$ is a pointer to either *null* or the first node in a linked list. Each node in the linked list contains a key, its data, and a pointer to the next node.

Given a key x we calculate $h(x)$. If $T[h(x)] = null$ then there's nothing associated to $h(x)$ so we create a node with key x and associated data and we point $T[h(x)]$ to it. Later, if we have some x' with $h(x') = h(x)$ we simply add another node to the linked list.

The advantage to this is that we have solved the collision problem. The disadvantage of course is when we search for the data associated to a key x we potentially have to iterate (slowly) through the linked list pointed to by $T[h(x)]$. Moreover each linked list could potentially get rather long.

5.2 Search

To search for the data associated with a key x we find the linked list pointed to by $T[h(x)]$ and then we step through it as we would with any linked list until we find our key k and then its associated data. If the linked list is long this can slow the process down and make us wonder why we're using a hash in the first place.

5.3 Deletion

To delete a key x (and its data) we simply find it as with search and delete it as with a standard linked list.

6 Probing

6.1 Overview and Insertion

One idea behind probing is that when we have a collision we simply "probe" the table for an open spot and insert the key (and data) there.

Of course we have to do this in some reliable way which enables searching and deletion.

Here are some approaches to probing with a few comments:

1. *Linear Probing*: If $T[h(x)]$ is occupied we look at:

$$T[h(x)+1],\ T[h(x)+2],\ T[h(x)+3],\ ...$$

We do this until we find an open spot. Here all sums are mod m.

Linear probing seems pretty simple but has a major advantage in that the probing procedure is guaranteed to try every entry in T.

One major issue with linear probing is that as soon as adjacent entries of T start to fill up, the linear probing process ends up running over these strings of adjacent entries repeately and the probing chains can get long, which makes for slow operations. This causes an issue known as *primary clustering*.

The basic way to solve this is to try to use a probing function which "jumps around" in a more "random" manner.

2. *Quadratic Probing*: If $T[h(x)]$ is occupied we look at:

$$T[h(x)+1],\ T[h(x)+4],\ T[h(x)+9],\ ...$$

We do this until we find an open spot. Here all sums are mod m.

Quadratic probing can appear more "random" than linear probing which suggests it might help solve the primary clustering issue, however unlike linear probing it turns out that quadratic probing is not guaranteed to try every entry in T. This can be shown using fairly simple examples.

However we can be somewhat careful:

Theorem 6.1.1. If we use quadratic probing and if m is prime then the first $\lfloor m/2 \rfloor$ probes are guaranteed to be distinct.

Proof. By way of contradiction assume that $h(x)+i^2 = h(x)+j^2 \mod m$ where m is an odd prime and $0 \leq i < j \leq \lfloor m/2 \rfloor$.

Then $m \big| (j^2 - i^2)$ so $m \big| (j-i)(j+i)$. Since m is prime it follows that $m \big| (j-i)$ or $m \big| (j+i)$. However since $0 \leq i < j \leq \lfloor m/2 \rfloor < m/2$, with this latter strict inequality because m is odd, both $j-i$ and $j+i$ are less than m and so neither of these divisibility requirements can hold and we have a contradiction. $\mathcal{QED}$

3. A generalization of the above would be to pick a function $p : \mathbb{Z} \to \mathbb{Z}_m$ where m is the size of T and then if $h(x)$ is occupied we look at:

$$T[h(x) + p(1))], T[h(x) + p(2)], T[h(x) + p(3)], ...$$

We do this until we find an open spot. Here all sums are mod m.

Some medium-level number theory can be used to show that judicious choices of p and m can result in the probing process trying every entry.

4. *Double-Hashing*: We define $g(x)$ to be another hash function and then if $h(x)$ is occupied we look at:

$$T[h(x) + g(x)], T[h(x) + 2g(x)], T[h(x) + 3g(x)], ...$$

We do this until we find an open spot. Here all sums are mod m.

Notice that this is different from before because our probing function $g(x)$ takes the keys as input whereas the probing function h took integers as input.

Analyis of double-hashing is challenging but it can be shown that double-hashing is very efficient.

6.2 Search

To search for the data associated with a key k we use the same approach as insertion. That is, we apply $h(x)$ and then probe as needed until we find our target key.

6.3 Deletions

Deletion is tricky business when using probing. The reason for this is that if we find the entry we wish to delete and simply empty out the table entry (to *null*) the probing function will not be able to "see past it" for any probes which need do.

The typically way to manage this is that when we delete a key (and its data) we put a special entry in the hash table which says something like *deleted*. Then when we are searching such an entry will indicate that the probing should keep going and when are probing for insertion such an entry will indicate that this table entry is available for insert.

There are of course other ways to manage deletion such as shifting other entries on the probe path.

7 Load Management

7.1 Approach

Note 7.1.1. In what follows, we'll use m to denote the size of the hash table and n to denote the number of keys which have been inserted into the hash

table at any given instant.

In both cases there is the notion that a hash table could get "too full". With separate chaining we'd like to keep our linked lists short on average and with probing we may simply fill up the hash table completely.

One way to solve this issue is to monitor the situation and rebuild the hash table if needed.

We do this as follows:

1. Suppose the hash table has m entries and at any instant we have inserted n keys into it. Define $\lambda = n/m$ to be the *load factor* of the hash table. We'd like to keep λ low, ideally at most 1. Observe that if $\lambda \leq 1$ then we are averaging fewer than one key per hash table entry.
2. We set an acceptable range $[\lambda_{min}, \lambda_{max}]$ of load factors and we require that we keep:
$$0 \leq \lambda_{min} \leq \lambda \leq \lambda_{max} \leq 1$$
3. If we insert a key and find $\lambda > \lambda_{max}$ we rebuild the entire hash table from scratch with a larger m and new hash function.
4. If we delete a key and find $\lambda < \lambda_{min}$ we rebuild the entire hash table from scratch with a smaller m and new hash function.

When we rebuild the hash table from scratch we do this as follows. Assume we have n keys in the hash table.

1. Create a new hash table T' with size:
$$m' = \left\lceil \frac{2n}{\lambda_{max} + \lambda_{min}} \right\rceil$$
Why this value? In such a case the new load factor of the hash table will satisfy:
$$\lambda \approx \frac{n}{2n/(\lambda_{max} + \lambda_{min})} = \frac{n}{2n/(\lambda_{max} + \lambda_{min})} = \frac{\lambda_{max} + \lambda_{min}}{2}$$
This is essentially the average of the minimum and maximum load factors. This give us some leeway for future insertions and deletions.
2. Create a new hash function $h' : K \to \mathbb{Z}_{m'}$.
3. Insert all the data from the old table into the new table.

Example 7.1. Suppose our set of keys K consists of all positive integers. We start with a list T indexed from 0 to 3 (a list of length 4) and $h : K \to \mathbb{Z}_4$ by $h(x) = 3x \mod 4$ using linear probing. Note that $m = 4$.
We set $\lambda_{min} = 0.1$ and $\lambda_{max} = 0.6$ and so we must have $0.1 \leq \lambda \leq 0.6$.

We start inserting keys into the hash. For each insert we get the following result:

- Insert $x = 1$: $h(1) = 3(1) = 3$ so $T = [null, null, null, 1]$. Now $n = 1$ and so $\lambda = 1/4 = 0.25$ which is fine.
- Insert $x = 10$: $h(10) = 3(10) = 2$ so $T = [null, null, 10, 1]$. Now $n = 2$ and so $\lambda = 2/4 = 0.5$ which is fine.
- Insert $x = 42$: $h(32) = 3(42) = 2$ but $T[2] \neq null$ so we probe linearly and find $T[0] = null$ so $T = [42, null, 10, 1]$. Now $n = 3$ and so $\lambda = 3/4 = 0.75$ which is too large.

We rebuild:

We create a new hash table T' with size:

$$m' = \left\lceil \frac{2(3)}{0.2 + 0.6} = \frac{6}{0.8} \right\rceil = 8$$

Thus our new T' is indexed from 0 to 7.

We create a new hash function $h' : K \to \mathbb{Z}_8$ by $h'(x) = 5x \mod 8$ using linear probing. Note that $m = 4$.

We then re-insert our values:

- Insert $x = 1$: $h(1) = 5(1) = 5$ so $T = [null, null, null, null, null, 1, null, null]$. Now $n = 1$ and so $\lambda = 1/8 = 0.125$ which is fine.
- Insert $x = 10$: $h(10) = 5(10) = 2$ so $T = [null, null, 10, null, null, 1, null, null]$. Now $n = 2$ and so $\lambda = 2/8 = 0.25$ which is fine.
- Insert $x = 42$: $h(42) = 5(42) = 2$ but $T(2) \neq null$ so we probe linearly and find $T[3] = null$ so $T = [null, null, 10, 42, null, 1, null, null]$. Now $n = 3$ and so $\lambda = 3/8 = 0.3125$ which is fine.

Note that we still have space for more growth.

7.2 Time Complexity

Our approach to load balancing suggests that amortized analysis might be the correct approach to determing the cost of a sequence of dictionary operations in a hash table.

This is true in fact and we have:

Theorem 7.2.1. Suppose each individual hashing operation is $\mathcal{O}(1)$. If we start with an empty hash table then the amortized cost of hashing is at most:

$$1 + \frac{\lambda_{max}}{\lambda_{max} - \lambda_{min}}$$

Proof. Omitted for now. There is a proof of this in David Mount's notes. $\mathcal{QED}$