CSE332: Data Abstractions

Lecture 10:Hashing

Ruth Anderson
Winter 2011

Announcements

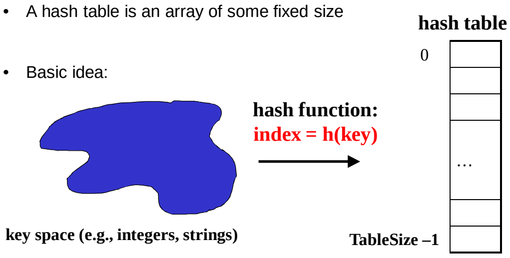
- **Project 2** – posted!
- **Homework 3**– due Friday Jan 28st at the BEGINNING of lecture

Today

- Dictionaries
  - Hashing

Hash Tables

- Aim for constant-time (i.e., *O*(1)) **`find`**, **`insert`**, and **`delete`**
  - "On average" under some reasonable assumptions
- A hash table is an array of some fixed size
- Basic idea:



Hash tables

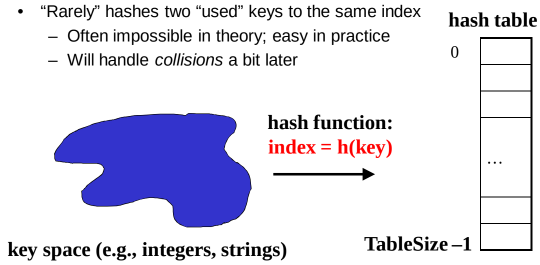
- There are *m* possible keys (*m* typically large, even infinite) but we expect our table to have only *n* items where *n* is much less than *m* (often written *n* << *m*)

Many dictionaries have this property

- Compiler: All possible identifiers allowed by the language vs. those used in some file of one program
- Database: All possible student names vs. students enrolled
- AI: All possible chess-board configurations vs. those considered by the current player
- …

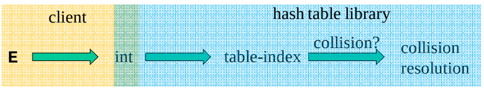
Hash functions

An ideal hash function:
- Is fast to compute
- "Rarely" hashes two "used" keys to the same index
  - Often impossible in theory; easy in practice
  - Will handle *collisions* a bit later



Who hashes what?

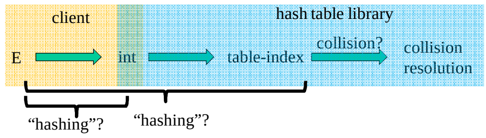
- Hash tables can be generic
  - To store elements of type **E**, we just need **E** to be:
    1. Comparable: order any two **E** (like with all dictionaries)
    2. Hashable: convert any **E** to an **int**
- When hash tables are a reusable library, the division of responsibility generally breaks down into two roles:



- We will learn both roles, but most programmers "in the real world" spend more time as clients while understanding the library

More on roles

Some ambiguity in terminology on which parts are "hashing"



Two roles must both contribute to minimizing collisions (heuristically)

- Client should aim for different ints for expected items
  - Avoid "wasting" any part of **E** or the 32 bits of the **int**
- Library should aim for putting "similar" **int**s in different indices
  - conversion to index is almost always "mod table-size"
  - using prime numbers for table-size is common

What to hash?

In lecture we will consider the two most common things to hash: integers and strings

- If you have objects with several fields, it is usually best to have most of the "identifying fields" contribute to the hash to avoid collisions
- Example:
  ```
  class Person {
     String first; String middle; String last;
     int age;
  }
  ```
- An inherent trade-off: hashing-time vs. collision-avoidance
  - Bad idea(?): Only use first name
  - Good idea(?): Only use middle initial
  - Admittedly, what-to-hash is often an unprincipled guess ☹

Hashing integers

- key space = integers
- Simple hash function:
  h(key) = key % TableSize
  - Client: **f(x) = x**
  - Library **g(x) = x % TableSize**
  - Fairly fast and natural
- Example:
  - TableSize = 10
  - Insert 7, 18, 41, 34, 10
  - (As usual, ignoring data "along for the ride")

0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

Hashing integers (Soln)

- key space = integers
- Simple hash function:
  h(key) = key % TableSize
  - Client: **f(x) = x**
  - Library **g(x) = x % TableSize**
  - Fairly fast and natural
- Example:
  - TableSize = 10
  - Insert 7, 18, 41, 34, 10
  - (As usual, ignoring data "along for the ride")

0	10
1	41
2	
3	
4	34
5	
6	
7	7
8	18
9	

Collision-avoidance

- With "**x % TableSize**" the number of collisions depends on
  - the ints inserted (obviously)
  - **TableSize**
- Larger table-size tends to help, but not always
  - Example: 7, 18, 41, 34, 10 with **TableSize** = 10 and **TableSize** = 7
- Technique: Pick table size to be prime. Why?
  - Real-life data tends to have a pattern, and "multiples of 61" are probably less likely than "multiples of 60"
  - Later we'll see that one collision-handling strategy does provably better with prime table size

More arguments for a prime table size

If **`TableSize`** is 60 and…
- Lots of data items are multiples of 5, wasting 80% of table
- Lots of data items are multiples of 10, wasting 90% of table
- Lots of data items are multiples of 2, wasting 50% of table

If **`TableSize`** is 61…
- Collisions can still happen, but 5, 10, 15, 20, … will fill table
- Collisions can still happen but 10, 20, 30, 40, … will fill table
- Collisions can still happen but 2, 4, 6, 8, … will fill table

In general, if **`x`** and **`y`** are "co-prime" (means **`gcd(x,y)==1`**), then
`(a * x) % y == (b * x) % y` if and only if **`a % y == b % y`**
- So good to have a **`TableSize`** that has not common factors with any "likely pattern" **`x`**

What if the key is not an int?

- If keys aren't **`int`**s, the client must convert to an **`int`**
  - Trade-off: speed and distinct keys hashing to distinct **`int`**s
- Very important example: Strings
  - Key space K = $s_0 s_1 s_2 \ldots s_{m-1}$
    - (where s_i are chars: $s_i \in [0,52]$ or $s_i \in [0,256]$ or $s_i \in [0,2^{16}]$)
  - Some choices: Which avoid collisions best?

1. h(K) = s_0 % TableSize
2. h(K) = $\left(\sum_{i=0}^{m-1} s_i\right)$ % TableSize
3. h(K) = $\left(\sum_{i=0}^{m-1} s_i \cdot 52^i\right)$ % TableSize

Specializing hash functions

How might you hash differently if all your strings were web addresses (URLs)?

Additional operations

- How would we do the following in a hashtable?
  - findMin()
  - findMax()
  - predecessor(key)
- Hashtables really not set up for these; need to search everything, O(n) time
- Could try a hack:
  - Separately store max & min values; update on insert & delete
  - What about '2nd to max value', predecessor, in-order traversal, etc; those are fast in an AVL tree

Hash Tables vs. Balanced Trees

- In terms of a Dictionary ADT for just **`insert`**, **`find`**, **`delete`**, hash tables and balanced trees are just different data structures
  - Hash tables *O*(1) on average (*assuming* few collisions)
  - Balanced trees *O*(**`log`** *n*) worst-case
- Constant-time is better, right?
  - Yes, but you need "hashing to behave" (avoid collisions)
  - Yes, but **`findMin`**, **`findMax`**, **`predecessor`**, and **`successor`** go from *O*(**`log`** *n*) to *O*(*n*)
- **Moral:** If you need to use operations like **`findMin`**, **`findMax`**, **`printSorted`**, **`predecessor`**, and **`successor`** often then you may prefer a balanced BST instead.

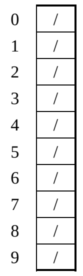
Collision resolution

Collision:
When two keys map to the same location in the hash table

We try to avoid it, but number-of-keys exceeds table size

So hash tables should support collision resolution
- Ideas?

Separate Chaining

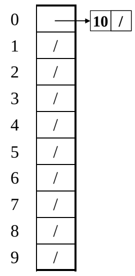


Chaining: All keys that map to the same table location are kept in a list (a.k.a. a "chain" or "bucket")

As easy as it sounds

Example: insert 10, 22, 107, 12, 42 with mod hashing and **`TableSize`** = 10

Separate Chaining

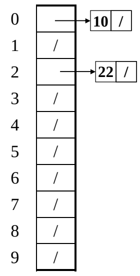


Chaining: All keys that map to the same table location are kept in a list (a.k.a. a "chain" or "bucket")

As easy as it sounds

Example: insert 10, 22, 107, 12, 42 with mod hashing and **`TableSize`** = 10

Separate Chaining

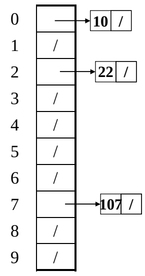


Chaining: All keys that map to the same table location are kept in a list (a.k.a. a "chain" or "bucket")

As easy as it sounds

Example: insert 10, 22, 107, 12, 42 with mod hashing and **`TableSize`** = 10

Separate Chaining

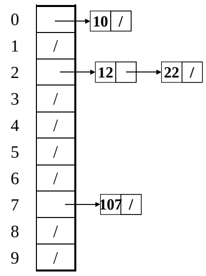


Chaining: All keys that map to the same table location are kept in a list (a.k.a. a "chain" or "bucket")

As easy as it sounds

Example: insert 10, 22, 107, 12, 42 with mod hashing and **`TableSize`** = 10

Separate Chaining

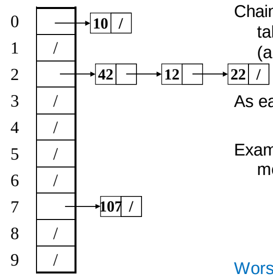


Chaining: All keys that map to the same table location are kept in a list (a.k.a. a "chain" or "bucket")

As easy as it sounds

Example: insert 10, 22, 107, 12, 42 with mod hashing and **`TableSize`** = 10

Separate Chaining



Chaining: All keys that map to the same table location are kept in a list (a.k.a. a "chain" or "bucket")

As easy as it sounds

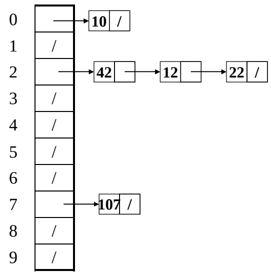
Example: insert 10, 22, 107, 12, 42 with mod hashing and **`TableSize`** = 10

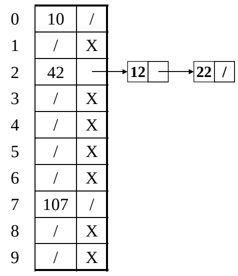
Worst case time for find?

Thoughts on separate chaining

- Worst-case time for **`find`**?
  - Linear
  - But only with really bad luck or bad hash function
  - So not worth avoiding (e.g., with balanced trees at each bucket)
    - Keep # of items in each bucket small
    - Overhead of AVL tree, etc. not worth it for small n
- Beyond asymptotic complexity, some "data-structure engineering" may be warranted
  - Linked list vs. array or a hybrid of the two
  - Move-to-front (part of Project 2)
  - Leave room for 1 element (or 2?) in the table itself, to optimize constant factors for the common case
    - A time-space trade-off…

Time vs. space (constant factors only here)





More rigorous separate chaining analysis

Definition: The load factor, λ, of a hash table is

$$\lambda = \frac{N}{\text{TableSize}} \quad \leftarrow \textbf{number of elements}$$

Under chaining, the average number of elements per bucket is ___

So if some inserts are followed by *random* finds, then on average:

- Each unsuccessful **`find`** compares against ____ items
- Each successful **`find`** compares against _____ items

- How big should TableSize be??

More rigorous separate chaining analysis

Definition: The load factor, λ, of a hash table is

$$\lambda = \frac{N}{\text{TableSize}} \quad \leftarrow \textbf{number of elements}$$

Under chaining, the average number of elements per bucket is λ

So if some inserts are followed by *random* finds, then on average:

- Each unsuccessful **`find`** compares against λ items
- Each successful **`find`** compares against $\lambda / 2$ items
- If λ is low, find & insert likely to be O(1)
- We like to keep λ around 1 for separate chaining

Separate Chaining Deletion?

Separate Chaining Deletion

- Not too bad
  - Find in table
  - Delete from bucket
- Say, delete 12
- Similar run-time as insert