CS112 Lab 09, April 4, 2010

http://cs-people.bu.edu/deht/cs112_spring11/lab09/

Diane H. Theriault

deht@cs.bu.edu

http://cs-people.bu.edu/deht/

Hash Tables

- Primary advantage?

- Potential Tradeoffs?

Hash Tables

- Primary advantage?
  - Speed. O(1) vs O(log n) in a tree or O(n) in a list

- Potential Tradeoffs?
  - Space for Time
  - Trickiness

Hash Tables

- How do we get O(1) insertion/retrieval?

Hash Tables

- How do we get O(1) insertion/retrieval?
  - Insertion / lookup into an array is constant time!
  - Somehow, convert our key into an array index

Hash Tables

- What are the issues that come up?

Hash Tables

- What are the issues that come up?
  - Not all data is integers
    - Inside a computer it is!
  - Possible values of data may be VERY LARGE
    - How big of an array would you need to have room for all of the words in the English language?
  - Need some sort of compression of keys:
    - **Hash functions**

Properties of Hash Functions

- Map from data to an array index
- Many-to-one relationship
  - Not invertible!
  - Repeatable (Deterministic)
- Should evenly distribute data among all possible array indexes
- Cheap to compute

Properties of Hash Functions

- What are some really bad hash functions?

Properties of Hash Functions

- What are some really bad hash functions?
  - Hash(KeyType Key) = Rand();
  - Hash(KeyType Key) = 5;
  - Hash(KeyType Key) = RecursiveFibonacci(Key.toInt())

Example Hash Functions

- Hash(Integer X)

- Hash(String S)

- Hash(Image)

Example Hash Functions

- Hash(Integer X)
  - X mod ArraySize
- Hash(String S)
  - Sum of integer values of all characters (mod ArraySize)
  - Treat string as huge base 16 integer (mod ArraySize)
- Hash(Image)
  - (Open research problem)

Using Hash Functions

- What do you do with the hash value once you have it?
    - Duh, insert your data into your array

Collisions

- What happens when two (or more) items have the same hash value?

Collisions

- What happens when two (or more) items have the same hash value?
  - One strategy: "separate chaining"
  - Store multiple items at the same location.
  - How?

Collisions

- What happens when two (or more) items have the same hash value?
  - One strategy: “separate chaining”
  - Store multiple items at the same location.
  - How?
    - Your array is an array of data structures that can store multiple items
      (e.g. linked list, search tree, symbol table)

Hash Table Miscellany

- Best when amount of data is small with respect to possible values of data.
  - E.g. 1,000,000,000 possible social security numbers, but only 10,000 customers
- Use prime array sizes
- Don't use hash tables when you'll want to do range queries
- Make sure your hash function actually does a good job of evenly distributing your data.

Hash Tables Applied to Calculator

Hash Tables Applied to Movie Database