Hashing

Jonathan Windle

University of East Anglia

J.Windle@uea.ac.uk

June 3, 2017

Overview I

Intro

- Technique for performing insertions,deletions and finds in a dictionary in constant average time.
- **Hash table:**
  - An array, T of some fixed size is used to store the keys.
  - `size` referres to the size of T.
  - $S = \{0, 1, ..., size - 1\}$
- **Hashing function:**
  - $h : K \rightarrow S$.
  - Suppose K is the set of 6 digit non-negative integers, then a possible (but poor) choce for h is:

$$h(k) = k(mod 1000)$$

- **Collisions:**
  - A collision occurs when two keys hash to the same location in the hash table:
    $h(k) = h(k')$.
  - Want to choose the hash function to minimise the chance of collisions.
  - Need to decide how to handle collisions when they do occur.

Choosing a Hash Function

- A good hash function maps keys uniformly and randomly into the full range of possible locations.
- A good hash function should depend on all of the charactersf the characters in a key, but this is not a sufficient condition for a good hash function.
- Must not just depend on all of the characters in a key but must also distribute keys evenly over the table.
- The built in Java function `hashCode` returns an integer based on the objects reference unless the object is a string then it is based on the string itself.
- The Java class `HashTable` can be used with keys of any user-defined data type provided an instance method `hashCode` is defined.

Resolving Collisions

- Use some other location that is open in the table:
  - Open addressing
- Change the structure of the hash table so that each location can correspond to more than one value:
  - Chaining.
  - Buckets.

Chaining/Buckets

- Chaining:
  - For each location T, keep a list of allthe keys hashed to that location.
  - Each entry in T is thus a reference to a linked list of keys.
  - To form a search, just hash to find the list and then perform the appropriate operation.
- Buckets:
  - Each location in the hash table is a bucket.
  - A fixed number, b of locations to store the keys.
  - Total space available is thus:
    $size \times b$

Open Addressing

- If a collision occurs, alternative cells in T are tried until an empty cell is found.
- Locating an open loaction in te hash table is called probing
- May be necessary to try more than one alternqative location.
- The locations examined when a new key is inserted is called a probe sequence.
- Let $\langle S_j^k \rangle$ denote the probe sequence then:
  $s_0^k = h(k)$
  $s_j^k = (s_{j-1}^k + p(j,k))\%size, \quad j \geq 1.$
- Where $p(j,k)$ is called a probe increment.
- In the simplist scheme the probe increment is independant of both j and k. i.e. it is a constant p in particular linear probing, $p = 1$.

The End