CS 3110

Lecture 14: Hash tables

Prof. Clarkson

Spring 2015

Today's music: *Re-hash* by Gorillaz

Review

Recently:

- Data abstractions
  - Lists, stacks, queues, dictionaries, polynomials, ...
- Imperative features
  - Refs, arrays, mutable fields

Today:

- Hash tables

Question #1

How often do you dictionaries/maps/hash tables/ associative arrays/etc. in your own programming?

A. Never
B. Infrequently
C. Frequently
D. Nearly every program I write
E. I'm not awake yet

Maps*

```
module type MAP = sig
  type ('key, 'value) map
  exception NotFound
  val insert:
    'key -> 'value -> ('key, 'value) map
      -> ('key, 'value) map
  val find: 'key -> ('key, 'value) map
      -> 'value option
  val remove: 'key -> ('key, 'value) map
      -> ('key, 'value) map
  ...
end
```

*aka associative array, dictionary, symbol table

Map implementations

- Arrays
- Association lists
- Functions
- Balanced search trees
- Hash tables

Map implementations

For each implementation:

- What is the representation type?
- What is the abstraction function?
- What are the representation invariants (if any)?
- What is the efficiency of each operation?

Arrays

- Representation type:

```
type ('key, 'value) map = 'value option array
```

- Assume we can convert `'key` to `int` in constant time
  - Conversion must be *injective*: never maps two keys to the same integer
  - Then there is a unique *inverse* mapping integers to keys: inverse(i) = k
  - Easiest realization: restrict keys to be integers!

Arrays

- Abstraction function: An array `[|v1; v2; ...|]` represents the map `{k1=v1, k2=v2, ...}`, where `k1=inverse(1), k2=inverse(2), ...` If `vi = None`, then `ki` is not bound in the map.
- Aka *direct address table*
- Efficiency:
  - insert: O(1)
  - find: O(1)
  - remove: O(1)
  - wastes space, because some keys are unmapped

Association lists

- Representation type:

```
type ('key, 'value) map =
   ('key*'value) list
```

- Abstraction function:
  - A list `[(k1,v1); (k2,v2); ...]` represents the map `{k1=v1, k2=v2, ...}`.
  - If k occurs more than once in the list, then in the map it is bound to the left-most value in the list.
- Efficiency:
  - insert: O(1)
  - find: O(n)
  - remove: O(n)

Functions

- Representation type:

```
type ('key, 'value) map =
   'key -> 'value
```

- Abstraction function:
  - A function `fun k -> if k=k1 then v1 else (if k=k2 then v2 else ...)` represents the map `{k1=v1, k2=v2, ...}`
- Efficiency:
  - insert: O(1)
  - find: O(n)
  - remove: not supported.
    - Could introduce *negative entries* in function of the form `if k=k' then raise NotFound`
    - But then find is O(N) where N is the number of entries ever added to the map

Balanced search trees

Red-black trees:

- Representation type:

  ```
  type ('key,'value) map = ('key,'value) rbtree
  ```

- Abstraction function: a node with label **`(k,v)`** and subtrees **`left`** and **`right`** represents the smallest map containing the binding **`{k=v}`** unioned with the bindings of **`left`** and **`right`**
- Representation invariant:
  - none for the map itself, but note that the tree will have its own rep invariant, namely, *the red-black invariants*
- Efficiency:
  - insert: O(lg n)
  - find: O(lg n)
  - remove: O(lg n)
- OCaml's **`Map`** module uses a closely-related balanced search tree called *AVL tree*

Question #2

If you wanted to map office numbers (e.g., 461) to occupant names (e.g., "Clarkson"), which implementation would be most time efficient?

A. Association lists
B. Functions
C. Balanced search trees
D. Arrays

Question #2

If you wanted to map office numbers (e.g., 461) to occupant names (e.g., "Clarkson"), which implementation would be most time efficient?

A. Association lists

B. Functions

C. Balanced search trees

D. Arrays

Map implementations

	insert	find	remove
Arrays	O(1)	O(1)	O(1)
Association lists	O(1)	O(n)	O(n)
Functions	O(1)	O(n)	N/A
Balanced search trees	O(lg n)	O(lg n)	O(lg n)

- Arrays guarantee constant efficiency, but require injective conversion of keys to integers
- Balanced search trees guarantee logarithmic efficiency

  ...we'd like the best of both worlds:

  constant efficiency with arbitrary keys

Hash tables

Main idea: give up on injectivity

- Allow conversion from `'key` to `int` to map multiple keys to the same integer
- Conversion function called a *hash* function
- Location it maps to called a *bucket*
- When two keys map to the same bucket, called a *collision*

...how to handle collisions?

Collision resolution strategies

1. Store multiple key-value pairs in a collection at a bucket; usually the collection is a list
   – called *open hashing, closed addressing, separate chaining*
   – this is what OCaml's **`Hashtbl`** does
2. Store only one key-value pair at a bucket; if bucket is already full, find another bucket to use
   – called *closed hashing, open addressing*

Hash table implementation

- Representation type:

```
 type ('key, 'value) map =
   ('key*'value) list array
```

- Abstraction function: An array

```
[|[(k11,v11); (k12,v12);...];
  [(k21,v21); (k22,v22);...]; ...|]
```

  represents the map `{k11=v11, k12=v12, ...}`.
  - If k occurs more than once in a bucket, then in the map it is bound to the left-most value in the bucket.
- Representation invariant:
  – A key **`k`** appears in array index **`b`** iff **`hash(k)=b`**
- Efficiency: ???
  – have to search through list to find key
  – no longer constant time

Question #3

Why does the representation type need to contain the **`'key`**?

```
type ('key, 'value) map =
    ('key*'value) list array
```

A. The type system requires it
B. A given bucket might contain many keys
C. To support an inverse operation
D. The hash table representation invariant requires it
E. None of the above

Question #3

Why does the representation type need to contain the **`'key`**?

```
type ('key, 'value) map =
    ('key*'value) list array
```

A. The type system requires it
B. A given bucket might contain many keys
C. To support an inverse operation
D. The hash table representation invariant requires it
E. None of the above

Efficiency of hash table

- Terrible hash function: `hash(k) = 42`
  - All keys collide; stored in single bucket
  - Degenerates to an association list in that bucket
    - insert: O(1)
    - find & remove: O(n)
- Perfect hash function: injective
  - Each key in its own bucket
  - Degenerates to array implementation
    - insert, find & remove: O(1)
  - Surprisingly, possible to design
    - if you know the set of all keys that will ever be bound in advance
    - size of array is the size of that set
    - so you want the size of the set to be much smaller than the size of the universe of possible keys
- Middleground? Compromise?

Efficiency of hash table

- New goal: constant-time efficiency on average
  - Desired property of hash function: distribute keys randomly among buckets to keep average bucket length small
  - If expected length is on average L:
    - insert: O(1)
    - find & remove: O(L)
- Two new problems to solve:
  1. How to make L a constant that doesn't depend on number of bindings in table?
  2. How to design hash function that distributes keys randomly?

Independence from # bindings

Let's think about the *load factor*...

= average number of bindings in a bucket = expected bucket length

= n/m, where n=# bindings in hash table, m=# buckets in array

- *e.g.*, 10 bindings, 10 buckets, load factor = 1.0
- *e.g.*, 20 bindings, 10 buckets, load factor = 2.0
- *e.g.*, 5 bindings, 10 buckets, load factor = 0.5

- Both OCaml **`Hashtbl`** and **`java.util.HashMap`** provide functionality to find out current load factor
- Implementor of hash table can't prevent client from adding or removing bindings
  - so n isn't under control
- But can *resize* array to be bigger or smaller
  - so m can be controlled
  - hence load factor can be controlled
  - hence expected bucket length can be controlled

Control the load factor

- If load factor gets too high, make the array bigger, thus reducing load factor
  - OCaml **`Hashtbl`** and **`java.util.HashMap`**: if load factor > 2.0 then double array size, bringing load factor back to around 1.0
  - Rehash elements into new buckets
  - Efficiency on average:
    - insert: O(1)
    - find & remove: O(2), which is still constant time
    - rehashing: will return to this next week (spoiler: constant time!)
- If load factor gets too small (hence memory is being wasted), could shrink the array, thus increasing load factor
  - Neither OCaml nor Java does this

Question #4

How would you resize this representation type?

```
type ('key, 'value) map =
  ('key*'value) list array
```

A. Mutate the array elements
B. Mutate the array itself
C. Neither of the above

Question #4

How would you resize this representation type?

```
  type ('key, 'value) map =
    ('key*'value) list array
```

A. Mutate the array elements

B. Mutate the array itself *(can't—it's immutable)*

C. Neither of the above

Resizing the array

Requires a new representation type:

```
type ('key, 'value) map =
  ('key*'value) list array ref
```

- Mutate an array element to **`insert`** or **`remove`**
- Mutate array ref to resize

Good hash functions

Three steps to transform key to bucket index:

1. **Serialize** key into a stream of bytes
   – should be injective
2. **Diffuse** bytes into a single large integer
   – small change to key should cause large, unpredictable change in integer
   – might lose injectivity here, but good diffusion into an int64 is likely to still be injective
3. **Compress** the integer to be within range of bucket indices
   – dependence on number of buckets: need to map from key to [0..m-1]
   – definitely lose injectivity

Responsibility for each step is typically divided between client and implementer...

Responsibilities

OCaml **`Hashtbl`**:

- function **`Hashtbl.hash : 'a -> int`** does serialization and diffusion in native C code, based on MurmurHash
- function **`Hashtbl.key_index`** does compression
- so implementer is responsible for everything
- great for client... until client wants a relaxed notion of equality on keys
  - e.g., keys are case-insensitive strings

Responsibilities

OCaml **`Hashtbl.Make`**:

- functor with input signature **`Hashtbl.HashedType`**, with functions
  - **`equal : t -> t -> bool`** and
  - **`hash : t -> int`**
- client provides **`equal`** and **`hash`** to do serialization and diffusion
  - must guarantee that if two keys are equal they have the same hash
- so implementer is responsible only for compression

Responsibilities

`java.util.HashMap:`

- method **`Object.hashCode()`** does serialization and diffusion
  - typical default implementation is to return address of object as an integer; not much diffusion there
  - client may override, must guarantee that if two keys are equal then they have the same hash
- method **`HashMap.hash()`** does further diffusion
  - implementer doesn't trust client!
- method **`HashMap.indexFor()`** does compression
- so implementer splits responsibilities with client

Designing your own hash function

- Compression:
  - Both Java and OCaml make the number **`m`** of buckets a power of two, and compress by computing **`mod m`**
- Serialization:
  - Both Java and OCaml support serialization; in OCaml it's in the **`Marshal`** module
- Diffusion:
  - Various techniques, including modular hashing, multiplicative hashing, universal hashing, cryptographic hashing...
  - If you don't achieve good diffusion, you lose constant-time performance!
  - If your hash function isn't constant time, you lose constant-time performance!
  - If you don't obey **`equals`** invariant, you lose correctness!
  - Designing a good hash function is hard

`Hashtbl` representation type

```
type ('a, 'b) t =
  { mutable size: int;
    mutable data: ('a, 'b) bucketlist array;
    ... }

and ('a, 'b) bucketlist =
    Empty
  | Cons of 'a * 'b * ('a, 'b) bucketlist
```

Why not use **`list`**? *Probably to save on one indirection.*

`Hashtbl` hash function

```
(* key_index : ('a, 'b) t -> 'c -> int *)
let key_index h key =
  ...
  (seeded_hash_param 10 100 h.seed key)
    land (Array.length h.data - 1)
  (* first line is serialization and diffusion,
   * second line is compression *)

external seeded_hash_param :
  int -> int -> int -> 'a -> int =
    "caml_hash" "noalloc"
(* caml_hash : 300 lines of C *)
(* hard to write good hash functions! *)
```

Hashtbl insert

```
(* add :  ('a, 'b) t -> 'a -> 'b -> unit *)
let add (h: ('a,'b) t) (key: 'a) info =
  let i = key_index h key in
  let bucket =
    Cons(key, info, h.data.(i)) in
  h.data.(i) <- bucket; (* mutation! *)
  h.size <- h.size + 1;
  if h.size >
    Array.length h.data lsl 1
    (* i.e. #buckets * 2 *)
  then resize key_index h
```

`Hashtbl` resize

```
let resize indexfun h =
  let odata = h.data in
  let osize = Array.length odata in
  let nsize = osize * 2 in (* double # buckets! *)
  if nsize < Sys.max_array_length then begin
    let ndata = Array.make nsize Empty in
    h.data <- ndata; (* mutation! *)
    let rec insert_bucket = function
        Empty -> ()
      | Cons(key, data, rest) ->
          insert_bucket rest;
          let nidx = indexfun h key in (* rehash! *)
          ndata.(nidx) <- Cons(key, data, ndata.(nidx)) in
    for i = 0 to osize - 1 do
      insert_bucket odata.(i)
    done
  end
```