COMP20007 Design of Algorithms

Hashing

Daniel Beck

Lecture 15

Semester 1, 2020

Dictionaries - Recap

- Abstract Data Structure: collection of *(key, value)* pairs.

Dictionaries - Recap

- Abstract Data Structure: collection of *(key, value)* pairs.
- Required operations: Search, Insert, Delete

Dictionaries - Recap

- Abstract Data Structure: collection of *(key, value)* pairs.
- Required operations: Search, Insert, Delete
- Last lecture: Binary Search Trees (and extensions)

Dictionaries - Recap

- Abstract Data Structure: collection of *(key, value)* pairs.
- Required operations: Search, Insert, Delete
- Last lecture: Binary Search Trees (and extensions)
- This lecture: Hash Tables.

Hash Tables

- A hash table is a continuous data structure with m preallocated entries.

like array

Hash Tables

- A hash table is a continuous data structure with m preallocated entries.
- Average case performance for Search, Insert and Delete: $\Theta(1)$

Hash Tables

- A hash table is a continuous data structure with m preallocated entries.
- Average case performance for Search, Insert and Delete: $\Theta(1)$
- Requires a *hash function*: $h(K) \to i \in [0, m-1]$.

key

in example
output: color

in real hash table,
the output is a position

Hash Tables

- A hash table is a continuous data structure with m preallocated entries.
- Average case performance for Search, Insert and Delete: $\Theta(1)$
- Requires a *hash function*: $h(K) \to i \in [0, m-1]$.
- A hash function should:
  - Be efficient ($\Theta(1)$).
  - Distribute keys evenly (uniformly) along the table.

want to avoid collison

Identity Hash Function

Question: if keys are integers, why do I need a hash function? I could just use the key as the index, no?

Identity Hash Function

Question: if keys are integers, why do I need a hash function? I could just use the key as the index, no?

- This is the *identity* hash function: $h(K) = K$.

Identity Hash Function

Question: if keys are integers, why do I need a hash function? I could just use the key as the index, no?

- This is the *identity* hash function: $h(K) = K$.
- Note that $K \in [0, m-1]$. In other words *we need to know the maximum number of keys in advance.*

should be bounded

Identity Hash Function

Question: if keys are integers, why do I need a hash function? I could just use the key as the index, no?

- This is the *identity* hash function: $h(K) = K$.
- Note that $K \in [0, m-1]$. In other words *we need to know the maximum number of keys in advance*.
- Sometimes this is possible: postcodes, for example.

Identity Hash Function

Question: if keys are integers, why do I need a hash function? I could just use the key as the index, no?

- This is the *identity* hash function: $h(K) = K$.
- Note that $K \in [0, m-1]$. In other words *we need to know the maximum number of keys in advance*.
- Sometimes this is possible: postcodes, for example.
- Many times it is not:
  - *m is too large* (need to preallocate) → eg. for postcode, you need to preallocate 10000 element
  - Unbounded integers (student IDs)
  - Non-integer keys (games)

Hashing Integers

- For large/unbounded integers, an alternative function is $h(K) = K \mod m$

Hashing Integers

- For large/unbounded integers, an alternative function is $h(K) = K \mod m$
- Allow us to set the size m.

Hashing Integers

- For large/unbounded integers, an alternative function is $h(K) = K \mod m$
- Allow us to set the size m.
- Small m results in lots of collisions, large m takes excessive memory. Best m will vary.

Hashing Strings

- Assume A $\mapsto 0$, B $\mapsto 1$, etc.
- Assume 26 characters and $m = 101$.
- Each character can be mapped to a *binary* string of length 5 ($2^5 = 32$).

map each character to
0-25

A — 00000
B — 00001
C — 00010

✓ $2^4 = 16 < 26$
$2^5 = 32$

Hashing Strings

- Assume A $\mapsto 0$, B $\mapsto 1$, etc.
- Assume 26 characters and $m = 101$.
- Each character can be mapped to a *binary* string of length 5 ($2^5 = 32$).

We can think of a string as a long binary number:

$$\text{M Y K E Y} \mapsto 0110011000010100010011000 \ (= 13379736)$$

$$13379736 \bmod 101 = 64$$

So 64 is the position of string M Y K E Y in the hash table.

Hashing Strings

We deliberately chose m to be prime.

$$13379736 = \overbrace{\boxed{12 \times 32^4 + 24 \times 32^3 + 10 \times 32^2 + 4 \times 32}}^{\text{will be cancelled by 32}} + \overset{\text{only this matter}}{24}$$

With $m = 32$, the hash value of any key is the last character's value! if last character is the same, many collisions

another problem for hashing strings:

① very long string, the number calculated is very large → overflow

Hashing Long Strings

Assume *chr* be the function that gives a character's number, so for example, $chr(\mathtt{c}) = 2$.

Hashing Long Strings

Assume *chr* be the function that gives a character's number, so for example, $chr(\mathrm{c}) = 2$.

Then we have

$$h(s) = (\sum_{i=0}^{|s|-1} chr(s_i) \times 32^{|s|-i-1}) \bmod m,$$

where m is a prime number. For example,

$$h(\mathrm{V\ E\ R\ Y\ L\ O\ N\ G\ K\ E\ Y}) = (21 \times 32^{10} + 4 \times 32^{9} + \cdots)\ mod\ 101$$

Hashing Long Strings

Assume *chr* be the function that gives a character's number, so for example, $chr(\mathrm{c}) = 2$.

Then we have

$$h(s) = (\sum_{i=0}^{|s|-1} chr(s_i) \times 32^{|s|-i-1}) \bmod m,$$

where m is a prime number. For example,

$$h(\mathrm{V\ E\ R\ Y\ L\ O\ N\ G\ K\ E\ Y}) = (21\times 32^{10} + 4\times 32^{9} + \cdots)\ mod\ 101$$

The term between parenthesis can become quite large and result in overflow.

Horner's Rule

Instead of

$$21 \times 32^{10} + 4 \times 32^9 + 17 \times 32^8 + 24 \times 32^7 \cdots$$

factor out repeatedly:

$$(\cdots ((21 \times 32 + 4) \times 32 + 17) \times 32 + \cdots) + 24$$

Horner's Rule

Instead of

$$21 \times 32^{10} + 4 \times 32^{9} + 17 \times 32^{8} + 24 \times 32^{7} \cdots$$

factor out repeatedly:

instead of calculating big number

$$(\cdots ((21 \times 32 + 4) \times 32 + 17) \times 32 + \cdots) + 24$$

Now utilize these properties of modular arithmetic:

$$\begin{aligned} (x + y) \bmod m &= ((x \bmod m) + (y \bmod m)) \bmod m \\ (x \times y) \bmod m &= ((x \bmod m) \times (y \bmod m)) \bmod m \end{aligned}$$

So for each sub-expression it suffices to take values modulo m.

Collisions

Happens when the hash function give identical results to two different keys.

Collisions

Happens when the hash function give identical results to two different keys.

We saw two solutions:

- Separate Chaining
- Linear Probing

Collisions

Happens when the hash function give identical results to two different keys.

We saw two solutions:

- Separate Chaining
- Linear Probing

Practical efficiency will depend on the table **load factor**: $\alpha = n/m$ n = # total of records

Separate Chaining

Assign multiple records per cell (usually through a linked list)

Separate Chaining

Assign multiple records per cell (usually through a linked list)

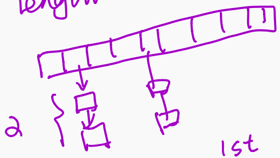
- Assuming even distribution of the n keys.

Separate Chaining

Assign multiple records per cell (usually through a linked list)

- Assuming even distribution of the n keys.
- A sucessful search requires $1 + \alpha/2$ operations on average.

ave length $= \alpha$



1st 1 op

2st 2 op

$\vdots$

α ops

$$\frac{1}{\alpha}(1+\cdots+\alpha) = \frac{1}{\cancel{\alpha}} \cdot \frac{\cancel{\alpha}(\alpha+1)}{2} = \frac{\alpha+1}{2}$$

if $0<\alpha<1$, $1+\alpha/2 \sim 1$ operation

if $\alpha \geq 1$

eg. $\alpha=2$, each cell has average 2 elements

$\rightarrow$ 2 operations

$\Leftrightarrow$ each cell has linked list of 2 elements, $\rightarrow$ search for two operation

Separate Chaining

Assign multiple records per cell (usually through a linked list)

- Assuming even distribution of the *n* keys.
- A sucessful search requires $1 + \alpha/2$ operations on average.
- An unsucessful search requires α operations on average.

try to look up but fails

$\boxed{\text{why ?}}$

look through whole linked list

Separate Chaining

Assign multiple records per cell (usually through a linked list)

- Assuming even distribution of the *n* keys.
- A sucessful search requires $1 + \alpha/2$ operations on average.
- An unsucessful search requires α operations on average.
- Almost same numbers for Insert and Delete.

Separate Chaining

Assign multiple records per cell (usually through a linked list)

- Assuming even distribution of the n keys.
- A sucessful search requires $1 + \alpha/2$ operations on average.
- An unsucessful search requires α operations on average.
- Almost same numbers for Insert and Delete.
- Worst case $\Theta(n)$ only with a bad hash function (load factor is more of an issue).

mapping all elements into same cell

Separate Chaining

Assign multiple records per cell (usually through a linked list)

- Assuming even distribution of the *n* keys.
- A sucessful search requires $1 + \alpha/2$ operations on average.
- An unsucessful search requires α operations on average.
- Almost same numbers for Insert and Delete.
- Worst case $\Theta(n)$ only with a bad hash function (load factor is more of an issue).
- Requires extra memory.

Linear Probing

Populate successive empty cells.

move until find an empty cell

Linear Probing

we will only have $0<\alpha<1$ in this case

Populate successive empty cells.

- Much harder analysis, simplified results show:
- A sucessful search requires $(1/2) \times (1 + 1/(1-\alpha))$ operations on average.
- An unsucessful search requires $(1/2) \times (1 + 1/(1-\alpha)^2)$ operations on average.

only make sense for $0<\alpha<1$

* we don't allocate extra space, if $\alpha>1$, means all spaces are fully allocated, when $\alpha=1$, worst case $O(n)$

Linear Probing

Populate successive empty cells.

- Much harder analysis, simplified results show:
- A sucessful search requires $(1/2) \times (1 + 1/(1 - \alpha))$ operations on average.
- An unsucessful search requires $(1/2) \times (1 + 1/(1 - \alpha)^2)$ operations on average.
- Similar numbers for Insert. Delete virtually impossible.

we don't actually delete element, just assign a flag show it's deleted

(如果真delete, 我们在搜索下一个会loss track.

不删除的缺点, 占用内存.

Linear Probing

Populate successive empty cells.

- Much harder analysis, simplified results show:
- A sucessful search requires $(1/2) \times (1 + 1/(1-\alpha))$ operations on average.
- An unsucessful search requires $(1/2) \times (1 + 1/(1-\alpha)^2)$ operations on average.
- Similar numbers for Insert. Delete virtually impossible.
- Does not require extra memory.

Linear Probing

Populate successive empty cells.

- Much harder analysis, simplified results show:
- A sucessful search requires $(1/2) \times (1 + 1/(1-\alpha))$ operations on average.
- An unsucessful search requires $(1/2) \times (1 + 1/(1-\alpha)^2)$ operations on average.
- Similar numbers for Insert. Delete virtually impossible.
- Does not require extra memory.
- Worst case $\Theta(n)$ with a bad hash function *and/or* clusters.

eg ?

Double Hashing

A generalisation of Linear Probing.

Apply a second hash function in case of collision.

Double Hashing

A generalisation of Linear Probing.

Apply a second hash function in case of collision.

- First try: $h(K)$ $(h(k)+0\cdot s(k)) \bmod m = h(k)$
- Second try: $(h(K)+s(K)) \mod m$
- Third try: $(h(K)+2s(K)) \mod m$
- ...

if $s(k)=1$

$(h(k)+0\cdot s(k)) \bmod m = h(k)$ since xx mod m $= h(k)$

$(h(k)+1\cdot 1) \bmod m$ $h(k) \bmod m = h(k)$

Double Hashing

A generalisation of Linear Probing.

Apply a second hash function in case of collision.

- First try: $h(K)$
- Second try: $(h(K) + s(K)) \mod m$
- Third try: $(h(K) + 2s(K)) \mod m$
- ...

? Another reason to use prime m in $h(K)$: will guarantee to find a free cell if there is one.

Double Hashing

A generalisation of Linear Probing.

Apply a second hash function in case of collision.

- First try: $h(K)$
- Second try: $(h(K) + s(K)) \mod m$
- Third try: $(h(K) + 2s(K)) \mod m$
- ...

Another reason to use prime m in $h(K)$: will guarantee to find a free cell if there is one.

Both Linear Probing and Double Hashing are sometimes referred as *Open Addressing* methods.

Rehashing

- High load factors deteriorate the performance of a hash table (for linear probing, ideally we should have $\alpha < 0.9$).

threshold

Rehashing

- High load factors deteriorate the performance of a hash table (for linear probing, ideally we should have $\alpha < 0.9$).
- *Rehashing* allocates a new table (usually around double the size) and move every item from the previous table to the new one.

Rehashing

- High load factors deteriorate the performance of a hash table (for linear probing, ideally we should have $\alpha < 0.9$).
- *Rehashing* allocates a new table (usually around double the size) and move every item from the previous table to the new one.
- *Very expensive operation*, but happens infrequently.

Summary

Hash Tables:

Summary

Hash Tables:

- Implement dictionaries.

Summary

Hash Tables:

- Implement dictionaries.
- Allow $\Theta(1)$ Search, Insert and Delete in the average case.

Summary

Hash Tables:

- Implement dictionaries.
- Allow $\Theta(1)$ Search, Insert and Delete in the average case.
- Preallocates memory (size m).

Summary

Hash Tables:

- Implement dictionaries.
- Allow $\Theta(1)$ Search, Insert and Delete in the average case.
- Preallocates memory (size m).
- Requires good *hash functions*.

Summary

Hash Tables:

- Implement dictionaries.
- Allow $\Theta(1)$ Search, Insert and Delete in the average case.
- Preallocates memory (size m).
- Requires good *hash functions*.
- Requires good collision handling.

Pros and Cons

If Hash Tables are so good, why bother with BSTs?

Pros and Cons

If Hash Tables are so good, why bother with BSTs?

- Hash Tables ignore *key ordering*, unlike BSTs.
- Queries like "give me all records with keys between 100 and 200" are easy within a BST but much less efficient in a hash table.

↑

if require ordering of keys

BST better

Pros and Cons

If Hash Tables are so good, why bother with BSTs?

- Hash Tables ignore *key ordering*, unlike BSTs.
- Queries like "give me all records with keys between 100 and 200" are easy within a BST but much less efficient in a hash table.
- Also: memory requirements of a hash table are much higher.

Pros and Cons

If Hash Tables are so good, why bother with BSTs?

- Hash Tables ignore *key ordering*, unlike BSTs.
- Queries like "give me all records with keys between 100 and 200" are easy within a BST but much less efficient in a hash table.
- Also: memory requirements of a hash table are much higher.

That being said, if hashing is applicable, a well-tuned hash table will typically outperform BSTs.

In Practice

Python dictionaries (*dict* type)

In Practice

Python dictionaries (*dict* type)

- Open addressing using *pseudo-random probing*
- Rehashing happens when $\alpha = 2/3$

In Practice

Python dictionaries (*dict* type)

- Open addressing using *pseudo-random probing*
- Rehashing happens when $\alpha = 2/3$

C++ *unordered_maps*

In Practice

Python dictionaries (*dict* type)

- Open addressing using *pseudo-random probing*
- Rehashing happens when $\alpha = 2/3$

C++ *unordered_maps*

- Uses chaining.
- Rehashing happens when $\alpha = 1$

In Practice

Python dictionaries (*dict* type)

- Open addressing using *pseudo-random probing*
- Rehashing happens when $\alpha = 2/3$

C++ *unordered_maps*

- Uses chaining.
- Rehashing happens when $\alpha = 1$

Next lecture: what happens if records/data is too large?