VE 281 Homework 2 (Due 10/31 11:59PM)

Name: 赵翰云 ID: 518370910091

- **Exercise 1. Open Addressing**

Consider inserting the keys $10, 22, 31, 4, 15, 28, 17, 88, 59$ into a hash table of length $m = 11$ using open addressing with the auxiliary hash function $h(k) = k$. Illustrate the result of inserting these keys using linear probing, using quadratic probing with $c_1 = 1$ and $c_2 = 3$, and using double hashing with $h_1(k) = k$ and $h_2(k) = 1 + (k \mod (m-1))$.

Solution:

We use T_t to represent each time stamp t starting with $i = 0$, and if encountering a collision, then we iterate i from $i = 1$ to $i = m - 1 = 10$ until there is no collision.

Linear probing:

$h(k,i) = (k+i) \mod 11$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
0 mod 11		22	22	22	22	22	22	22	22
1 mod 11								88	88
2 mod 11									
3 mod 11									
4 mod 11				4	4	4	4	4	4
5 mod 11					15	15	15	15	15
6 mod 11						28	28	28	28
7 mod 11							17	17	17
8 mod 11									59
9 mod 11			31	31	31	31	31	31	31
10 mod 11	10	10	10	10	10	10	10	10	10

Quadradic probing:

$h(k,i) = (k+i+3i^2) \mod 11$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
0 mod 11		22	22	22	22	22	22	22	22
1 mod 11								88	88
2 mod 11									
3 mod 11							17	17	17
4 mod 11				4	4	4	4	4	4
5 mod 11									
6 mod 11							28	28	28
7 mod 11									59
8 mod 11					15	15	15	15	15
9 mod 11			31	31	31	31	31	31	31
10 mod 11	10	10	10	10	10	10	10	10	10

Double hashing:

$h(k,i) = (k+i(1+k \mod 10)) \mod 11$	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
0 mod 11		22	22	22	22	22	22	22	22
1 mod 11									
2 mod 11									59
3 mod 11							17	17	17
4 mod 11					4	4	4	4	4
5 mod 11						15	15	15	15
6 mod 11							28	28	28
7 mod 11								88	88
8 mod 11									
9 mod 11				31	31	31	31	31	31
10 mod 11	10	10	10	10	10	10	10	10	10

- **Exercise 2. Slot-Size Bound for Chaining**

Suppose that we have a hash table with n slots, with collisions resolved by chaining, and suppose that n keys are inserted into the table. Each key is equally likely to be hashed to each slot. Let M be the maximum number of keys in any slot after all the keys have been inserted. Your mission is to prove an $O(\lg n / \lg\lg n)$ upper bound on $\mathrm{E}[M]$, the expected value of M

a. Argue that the probability Q_k that exactly k keys hash to a particular slot is given by

$$Q_k = \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \binom{n}{k}$$

b. Let P_k be the probability that $M = k$, that is, the probability that the slot containing the most keys contains k keys. Show that $P_k \le nQ_k$.

c. Use Stirling's approximation to show that $Q_k < e^k/k^k$.

d. Show that there exists a constant $c > 1$ such that $Q_{k_0} < 1/n^3$ for $k_0 = c\lg n/\lg\lg n$. Conclude that $P_k < 1/n^2$ for $k \ge k_0 = c\lg n/\lg\lg n$

e. Argue that

$$\mathrm{E}[M] \le \Pr\left\{M > \frac{c\lg n}{\lg\lg n}\right\} \cdot n + \Pr\left\{M \le \frac{c\lg n}{\lg\lg n}\right\} \cdot \frac{c\lg n}{\lg\lg n}$$

Conclude that $\mathrm{E}[M] = O(\lg n/\lg\lg n)$

Solution:

a. $(\frac{1}{n})^k$ possibility of k keys inserted into a particular slot

$(1-\frac{1}{n})^{n-k}$ all other keys not in this slot

$\binom{n}{k}$ choose k keys from n

so $Q_k = (\frac{1}{n})^k (1-\frac{1}{n})^{n-k} \binom{n}{k}$

b. nQ_k: there's one slot containing k keys (but may not be the only one)

nQ_k = P[one slot contain k keys and all others no longer ~~less than~~ k] + P[one slot contains k key and there some more than k]

$= P_k + P_{\text{other}}$

so $P_k \le nQ_k$

c. $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{(\frac{n}{e})^n}{\sqrt{2\pi n}(\frac{k}{e})^k (\frac{n-k}{e})^{n-k}} = \frac{n^n}{\sqrt{2\pi n}\, k^k (n-k)^{n-k}}$

$Q_k = \frac{n^n}{e\sqrt{2\pi n}\, k^k (n-k)^{n-k}} \cdot \frac{1}{n^k}(1-\frac{1}{n})^{n-k} = \frac{1}{\sqrt{2\pi n}\, k^k} \cdot (1+\frac{k}{n-k})^{n-k} \cdot (1-\frac{1}{n})^{n-k} < \underbrace{\frac{1}{\sqrt{2\pi n}}(1-\frac{1}{n})^{n-k}}_{<1} \cdot \frac{e^k}{k^k} < \frac{e^k}{k^k}$ ($1+\frac{k}{n-k} < e^k$)

d. ~~$\lg Q_k < -k \lg k$~~ $\ln Q_k < -k \ln k$ $\lg Q_k = -\frac{k \lg e \lg k}{k \lg e \lg k}$

when $k = k_0 = c \ln n / \ln\ln n$ $\to \frac{1}{2} \ln\ln n$

$\ln Q_{k_0} < -c(\ln n / \ln\ln n) \cdot (\ln\ln n - \ln\ln\ln n)$

$\ln\ln n > 2 \ln\ln\ln n$ for large n

so $\ln Q_{k_0} < \frac{c}{2} \ln n$

$c = 6 \Rightarrow Q_{k_0} < -3 \ln n \Rightarrow Q_{k_0} < \frac{1}{n^3}$

$P_{k_0} < n Q_{k_0} < \frac{1}{n^2}$

e. $M_{max} = n$ let $O = \frac{c \lg n}{\lg\lg n}$

$E[M] = \sum P_r\{M = i > O\} \cdot i + \sum P_r\{M = i \le O\} \cdot i$

$\le P_r\{M > O\} \cdot n + P_r\{M \le O\} \cdot O$

~~For $k \le \frac{c \lg n}{\lg\lg n}$~~

$\le \left(\sum_{O}^{n} 1\right) \cdot \frac{1}{n^2} \cdot n + \left(1 - \sum_{O}^{n} \frac{1}{n^2}\right) \cdot O$

$\downarrow$ $O(1)$ $\qquad$ $\downarrow$ $O(O)$

so $E[M] = O\left(\frac{c \lg n}{\lg\lg n}\right)$

- **Exercise 3. D-select**

  In the deterministic linear-time selection algorithm, instead of partitioning all the inputs into a number of groups of size 5, we partition the inputs into a number of groups of size 7. The rest steps of the algorithm are similar to the original version. Will the runtime of this new algorithm still be O(n), where n is the input size? Prove your claim.

  Solution:

Suppose there's a C that

$T_n \le an + T(\frac{n}{7}) + T(\frac{5n}{7})$ ← $1 - \frac{4}{7} \times \frac{1}{2}$

and $T(1) \le C$

we hope there's a constant a such that

$T(n) \le an$ for all $n > 1$

we choose $a \ge 10C$ and then prove it

first, Because $T(1) \le C$, so $T(1) \le 7aC$

if $T(m) \le am$ is true for all $m \le k$

then $T(k+1) \le c(k+1) + (k+1)c + 5(k+1)c$

$= 7c(k+1)$

By induction, $T(n) \le an$ for all pos int n ($a \ge 7c$)

therefore $T(n) = O(n)$

- **Exercise 4. Silly Blue Tiger**

  1. Imagine that an algorithm requires us to hash strings containing English phrases. Knowing that strings are stored as sequences of characters, Blue Tiger decides to simply use the sum of those character values (modulo the size of its hash table) as the string's hash. Will the performance of the implementation match the expected value shown in lecture? (Yes/No and the brief reasoning are what we expect.)
  2. Blue Tiger decides to implement both collision resolution and rehashing for its hash table. However, it does not want to do more work than necessary, so it wonders if it needs both to maintain the correctness and performance it expects. After all, if it has rehashing, it can resize to avoid collisions; and if it has collision resolution, collisions don't cause correctness issues. Which statement about these two properties true? (Choose one or two of the statements and briefly explain your choice.)
  3. Blue Tiger wants to implement rehashing. Assuming it is enlarging a table of size m into a table of size m', and the table contains n elements, what is the best time complexity to achieve this rehashing step (expanding a m-slot table with n elements into a m'-slot table)? Answer in big theta notation in terms of the variables in the context. Intermediate results of your complexity analysis should be displayed.
  4. In lecture, we discussed approximately doubling the size of our hash table. Blue Tiger begins to implement this approach (that is, it lets m' be a prime close to but greater than $2m$) but stops when it thinks that it might be able to avoid wasting half of the memory the table occupies on empty space by letting m' be a prime number close to $(m + k)$ instead, where k is some constant. Does this work equally well comparing to picking a prime number close to $2m$? Explain your answer.

Solution:

1. no. words with same characters but different sequence will be hashed to same slot "top" "pot"

2. collision resolution or let collision don't cause correctness issue is true. It will just take longer true time (longer and longer) to carry out. rehashing reduce the possibility of collision but cannot avoid it

}

- **Exercise 5. Rehashing**

Given a sequence of inputs 4371, 1323, 6173, 4199, 4344, 9679, 1989, insert them into a hash table of size 10. Suppose that the hash function is h(x) = x mod 10. Hash table uses the double hashing with the second hash function as g(x) = (7 - x) mod 7.(Be careful with this mod calculation)
If any given input cannot be inserted correctly, do rehashing to guarantee a successful insertion, using a new table size as introduced in class (the first prime number after doubling the original hash table size) and a new hash function h(x) = x mod new_size. Note that the order in rehashing depends on the order stored in the old hash table, not their previous inserting order.
Solve this question step by step and show intermediate results.

Solution:

The initial hash table:

rehash

	T_0	T_1	T_2	T_3	T_4	T_5	T_6
entry0							
entry1	4371	4371	4371	4371	4371	4371	
entry2							
entry3		1323	1323	1323	1323	1323	
entry4			6173	6173	6173	6173	
entry5						9679	
entry6							
entry7					4344	4344	
entry8							
				4199	4199	4199	

Does the hash table need rehashing? If it does, when and why (answer in 1-2 sentences)? Also finish the rehashing process by drawing your own progress table that is similar to the one above:

T6 hi = (9+6i)%10 , 13579 are occupied

	T_0	T_1	T_2	T_3	T_4	T_5	T_6
0							
1	4371						
2							
3							
4							
5							
6							
7							
8							
9		6173					
10							
11						1989	
12		1323					
13					4199		
14							
15							
16							
17							
18							
19						9679	
20					4344		
21							
22							
23							

- **Exercise 6**

  Suppose we want to design a hash table containing at most 1162 elements using quadratic probing. We require it to guarantee successful insertions, S(L) < 2 and U(L) < 4. Please determine a proper hash table size and show all intermediate steps.

$\frac{1}{2}\left[1+\frac{1}{1-L}\right]<2 \Rightarrow L<\frac{3}{4}$

$\frac{1}{2}\left[1+\left(\frac{1}{1-L}\right)^2\right]<4 \Rightarrow L<0.763$

$\frac{1}{1-L}<4 \Rightarrow L<0.75$

$\begin{cases} \frac{1}{L}\ln\frac{1}{1-L}<2 \Rightarrow L<0.797 \end{cases}$

so $L<0.75$

$\left\lceil \frac{1162}{0.75} \right\rceil = 1550$

take the closest prime 1553

- **Exercise 7**

  In a few sentences (in an itemized fashion), explain why a hash table with open addressing and rehashing (doubling the size every time when the load factor exceeds a threshold) achieves insertion in $O(1)$ time.

  Answer the following questions: 1) When rehashing is not triggered, is the insertion time $O(1)$ (does it have an upper bound)? Why? 2) Show that at any given point, the amortized (distributed to all items in the hash table) cost of rehashing is $O(1)$ for each item in the hash table.

Because open addressing cost O(1) time, and rehashing cost O(1) time amotize over each insert.

1. hashing: O(1)
2. open addressing: O(1) before rehashing
3. rehash: O(1) if amortized.

(1) No. the worst case is O(n) e.g. linear probing, and all elements have the same original bucket (O(M) cost)

(2) suppose size M→2M, and we have ML items already. (L load factor). Adding another triggers rehash, we need to insert (M+1) items, cost O(M) time, created a new table O(1)

Rate $\frac{2q(M)}{LM+1} = O(1)$

What if during rehashing, the hash table is expanded by a factor x with $x > 2$? Is the amortized cost of rehashing still $O(1)$? If no, explain. If yes, show your work. If it depends, show both.

Yes. suppose each time the table expanded x times, & suppose the last rehashed table is of size M. then the total time on rehashing is $O(M)+\frac{1}{x}O(M)+\frac{1}{x^2}O(M)\cdots$

$= O(M)$, and the hashed items number is $O(M)$

so still O(1), if amortized.