Data Structures

Topic: Hashing

By

Ravi Kant Sahu

Asst. Professor

Lovely Professional University, Punjab

Outlines

- Introduction
- Hashing
- Hash Functions
  - Division Method
  - Midsquare Method
  - Folding Method
- Collision Resolution
- Open Addressing : Linear Probing & Modifications
  - Quadratic Probing
  - Double Hashing

Introduction

- The search time of all the algorithms depends on the number n of the elements in the collection of Data.

- A searching technique which is essentially independent of n is called Hash Addressing or Hashing.

- F is a file with n records and a set K of Keys which uniquely determine the records in F.

- F is maintained in memory by a table T of m memory locations and L is a set of memory addresses of the locations in T.

Example

- Suppose a company with 168 employees assigns a 5 digit Emp_No. to each employee which is used as primary key in Employee file.

- Emp_No can be used as address of record in memory but we will require 100000 memory locations.

Hashing

- Hashing is a searching technique which is independent of the number of elements in file.

- The general idea of using the Key to determine the address of records is an excellent idea. But it must be modified to prevent the wastage of space.

- The Modification takes the form of a function H from the set K of keys into the set L of memory addresses.

$$\mathbf{H}: K \rightarrow L$$

Hash Functions

- Hash function H is a mapping between set of Keys K and set of memory locations L.

  H: K → L

- Such a function H may not yield distinct values.

- It is possible that two different keys K1 and K2 will yield the same hash address.

- This situation is called Collision.

Hash Functions…

- Two principle criteria used in selecting a hash function H are:

  1. H should be very easy and quick to compute.

  2. H should be uniformly distribute the hash address throughout the set L. So that the number of collisions are minimized.

- Some popular hash Functions are:
  - Division Method
  - Midsquare Method
  - Folding Method

Division Method

- Choose a number *m* larger than the number *n* of Keys(usually a prime number) . Hash function is defined as:

$$H(K) = k \pmod{m}$$

or $$H(K) = k \pmod{m} + 1$$

(when we want hash address to range from 1 to m rather than 0 to m-1)

Midsquare Method

- The key is squared and some digits are deleted from both sides to obtained l digits.

$$H(k) = l$$

where l is obtained by deleting digits from both ends of k^2 .

Folding Method

- The key k is partitioned into a number of parts k1, k2, k3… k_r, where each part (except possibly the last) has the same number of digits as the required address.

- Then the parts are added together, ignoring the last carry.

$$H(k) = k1 + k2 + \ldots + k_r$$

Hash Table

- A hash table (also hash map) is a data structure used to implement an associative array, a structure that can map keys to values.

- A hash table uses a hash function to compute an index into an array of buckets or slots (or in memory), from which the correct value can be found.

- Ideally, the hash function should assign each possible key to a unique bucket, but this ideal situation is rarely achievable in practice.

Collision Resolution

Collision Resolution

- Collision is a situation when two or more keys map to the same memory location.

- Load Factor: The ratio of number n of keys in K to the number m of hash addresses in L.

$$\lambda = n/m$$

- Efficiency of a hash function is measured by the average number of probes needed to find the location of record with a given key k.

Open Addressing

➢ Linear Probing

- Resolves collisions by placing the data into the next open slot in the table.

- We assume that the table T with m locations is circular, so that T[1] comes after T[m].

Linear Probing

- Insert pairs whose keys are 6, 12, 34, 29, 28, 11, 23, 7, 0, 33, 30, 45

- divisor = b (number of buckets) = 17.
- H = key % 17.

0				4				8				12				16
34	0	45				6	23	7			28	12	29	11	30	33

Problems with Linear Probing

- Identifiers(keys) tend to cluster together
- Adjacent cluster tend to coalesce
- Increase the search time

- Worst Case Complexity is $\Theta(n)$. This happens when all keys are in the same cluster.

- Two techniques are used to minimize clustering:
  1. Quadratic Probing
  2. Double Hashing

Quadratic Probing

- Quadratic probing uses a quadratic function of *i* as the increment

- Examine buckets H(x), $(H(x)+i^2)\%b$

Double Hashing

- A second hash function H` is used for resolving the collision.

- Let $H(k) = h$ and $H`(k) = h` \neq m$

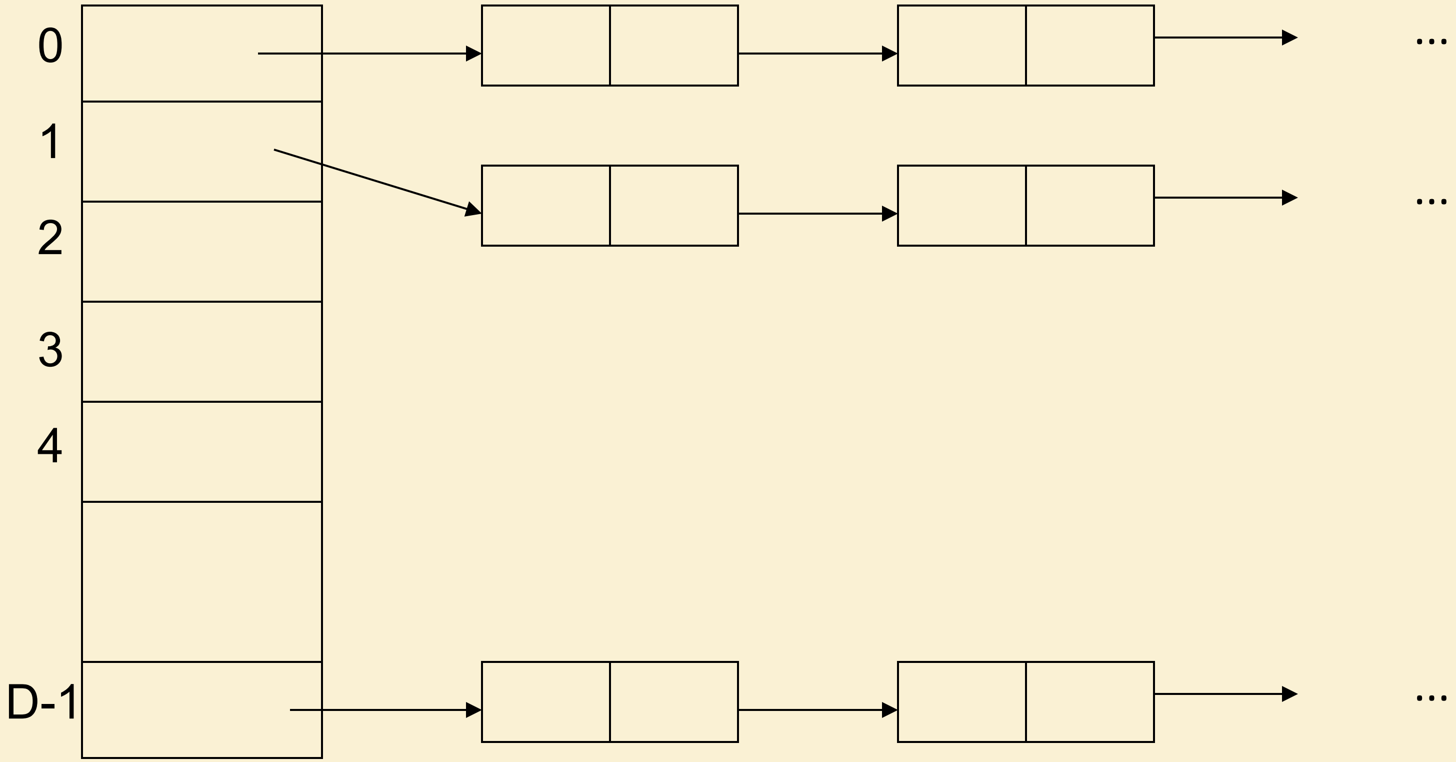
  Then we linearly search the locations with addresses:

  $$h, h+h`, h+2h`, h+3h`, \ldots$$

Open Hashing or Separate Chaining

- Each bucket in the hash table is the head of a linked list.
- All elements that hash to a particular bucket are placed on that bucket's linked list.
- Records within a bucket can be ordered in several ways
  - by order of insertion,
  - by key value order, or
  - by frequency of access order

Open Hashing Data Organization



Analysis

- Open hashing is most appropriate when the hash table is kept in main memory, implemented with a standard linked list.

- We hope that number of elements per bucket roughly equal in size, so that the lists will be short.

- If there are *n* elements in set, then each bucket will have roughly *n/D* elements.

- If we can estimate *n* and choose *D* to be roughly as large, then the average bucket will have only one or two members

Brainstorming-1

In a hash table of size 13 which index positions would the following two keys map to?

27, 130

(A) 1, 10
(B) 13, 0
(C) 1, 0
(D) 2, 3

Brainstorming-2

Suppose you are given the following set of keys to insert into a hash table that holds exactly 11 values:

113 , 117 , 97 , 100 , 114 , 108 , 116 , 105 , 99

Which of the following best demonstrates the contents of the has table after all the keys have been inserted using linear probing?

(A) 100, __, __, 113, 114, 105, 116, 117, 97, 108, 99

(B) 99, 100, __, 113, 114, __, 116, 117, 105, 97, 108

(C) 100, 113, 117, 97, 14, 108, 116, 105, 99, __, __

(D) 117, 114, 108, 116, 105, 99, __, __, 97, 100, 113

Brainstorming-3

Suppose you are given the following set of keys to insert into a hash table that is capable of holding exactly 12 values:

93 , 47 , 97 , 106 , 15 , 121 , 108 , 31 , 9

Find out the average number of probes for Successful Search and Unsuccessful search if Linear Probing is used for Collision Resolution.