CS2040C Semester 2 2018/2019

Data Structures and Algorithms

Tutorial 06 - Table ADT 1, Hash Table

For Week 08

Document is last modified on: March 13, 2019

1 Introduction and Objective

In this tutorial, we will continue our discussion about Non-Linear Data Structures. Next up is Hash Table, **one possible efficient** implementation of Table ADT (unordered). We will heavily use `https://visualgo.net/en/hashtable` in this tutorial.

2 Tutorial 06 Questions

Hash Table Basics

Q1). (Choose 2 out of 3 to be discussed live): A good hash function is essential for good Hash Table performance. A good hash function is easy/efficient to compute and will evenly distribute the possible keys. Comment on the flaw (if any) of the following hash functions. Assume that for this question, the load factor α = number of keys N / Hash Table size $M = 0.3$ (i.e. low enough) for all cases below:

1. $M = 100$. The keys are positive even integers. The hash function is h(key) = key % 100.
2. $M = 1\,009$. The keys are valid email addresses. The hash function is h(key) = (sum of ASCII values of the **last 10 characters**) % 1 009. See `http://www.asciitable.com` for ASCII values.
3. $M = 101$. The keys are integers in the range of [0, 1 000]. The hash function is h(key) = floor(key * random) % 101, where $0.0 \leq$ random ≤ 1.0.

Q2). Hashing or No Hashing: Hash Table is a Table ADT that allows for `search(v)`, `insert(new-v)`, and `delete(old-v)` operations in O(1) average-case time, **if properly designed**. However, it is not without its limitations. For each of the cases described below, state if Hash Table can be used. If not possible to use Hash Table, explain why is Hash Table not suitable for that particular case. If it is possible to use Hash Table, describe its design, including:

1. The <Key, Value> pair
2. Hashing function/algorithm
3. Collision resolution (OA: LP/QP/DH or SC; give some details)

The cases are:

1. A mini-population census is to be conducted on every person in your (not so large) neighbourhood. No two person have the same name but there can be two or more person with the same age. You can assume that age is an integer and within reasonable human age range [0..150] years old. We are only interested in storing every person's name and age. The operations to perform are: retrieve age by name and retrieve list of names (in any order) by age.
2. A much larger population census is also conducted across the country, similarly containing only every person's name and age. Again, no two person have the same name but there can be two or more person with the same age. You can assume that age is an integer and within reasonable human age range [0..150] years old. The operation to perform is: retrieve list of names (in any order) of people eligible for voting. Only people above legal age 17 years old (or older) are eligible for voting. However, we still need to store the rest of the data.
3. A different population census similarly contains only the name (in full name, distinct) and the age of every person. The operation to perform is: Retrieve person's full name and his/her age given a last (sur)-name. Note that although the full names are distinct, their last (sur-)names may not.
4. A grades management program stores a student's index number and his/her final marks in one GCE 'O' Level subject. There are 100,000 students, each scoring final marks in [0.0, 100.0]. The operation to perform is: Retrieve a list of students who passed in ranking order (highest final marks to passing marks). A student passes if the final marks are more than 65.5. Whether a student passes or not, we still need to store all students' performance.

Basic Hash Table Stuffs

Q3). Quick check: Let's review all 4 modes of Hash Table (use the Exploration mode of `https://visualgo.net/en/hashtable`). During the tutorial session, the tutor will randomize the Hash Table size M, the selected mode (LP, QP, DH, or SC), the initial keys inside, and then ask student to `Insert(random-integer)`, `Remove(existing-integer)`, or `Search(integer)` operations. This part can be skipped if most students are already comfortable with the basics.

Hash Table Discussions

Q4). (Choose 2 out of 3 to be discussed live): The following topics require deeper understanding of Hash Table concept. Please review `https://visualgo.net/en/hashtable?slide=1`, use the Exploration Mode, or Google around to help you find the initial answers and we will discuss the details in class. For some questions, there can be more than one valid answer.

1. What is/are the main difference(s) between List ADT basic operations (see `https://visualgo.net/en/list?slide=2-1`) versus Table ADT basic operations (see `https://visualgo.net/en/hashtable?slide=2-1`)?

2. At `https://visualgo.net/en/hashtable?slide=4-4`, Steven mentions about Perfect Hash Function. Now let's try a mini exercise. Given the following strings, which are the names of Steven's current family members: {"Steven Halim", "Grace Suryani Halim", "Jane Angelina Halim", "Joshua Ben Halim", "Jemimah Charissa Halim"}, design any valid **minimal perfect hash function** to map these 5 names into index [0..4] without any collision. Steven and Grace are not planning to increase their family size so you can assume that $N = 5$ will not change.

3. Thus far, which collision resolution technique is better (in your opinion or Google around): One of the Open Addressing technique (LP, QP, DH) or the Separate Chaining technique?