CSCI181AF: Advanced Data Structures
09/21/2022

Comparison of Multiple Hash Table Strategies by Max Chain Length

Introduction

In this experiment, we are comparing the abilities of three different hash table structures, a traditional hash table, 2-choice hash table, and a 2-left hash table, that utilize separate chaining to minimize the length of the maximum length chain of the hash table. All hash tables have been implemented as lists in Python, with the hash function being a random number generator modded by n (number of buckets) to simulate a hash function. The control hash table (x_1) is an array of length n with our hash function. The 2-Choice hash table (x_2) hashes the same value with two different hash functions (two different randomly generated integers), then compares the length of the chains at each index, and adds the value to the shorter of the chains. 2-Left hash table (x_2) splits the array into a left and right half of the array (split at $n//2$ index), then hashes calculates a hash value for the left and right halves separately, places the value in the bucket with a shorter chain, or in the case of a tie, places in left.

Data

We tested each of these hash tables with n buckets by adding n values to our hash tables, then finding the max length chain in each table for values of $n = 500, 500000, 2000000$. These experiemnts were repeated 100 times, from which we calculated meaningful statistics. The last table represents the theoretical values for the length of the max bucket using literature provided in class where $\hat{x_1} = \frac{\ln(n)}{\ln(\ln(n))}$, $\hat{x_2} = \log_2(\ln(n))$, and $\hat{x_3} = 0.69 \log_2(\ln(n))$.

Strategy	Avg.	STDev.	95% CI
Control	5.11	0.7	±0.13
2-Choice	2.98	0.2	±0.04
2-Left	2.89	0.3	±0.06

Table 1: $n = 500$

Strategy	Avg.	STDev.	95% CI
Control	8.57	0.7	±0.13
2-Choice	3.97	0.17	±0.03
2-Left	3.02	0.14	±0.03

Table 2: $n = 500000$

Strategy	Avg.	STDev.	95% CI
Control	9.2	0.6	±0.11
2-Choice	4	0	±0
2-Left	3.13	3	±0.07

Table 3: $n = 2000000$

Strategy	500	500000	2000000
Control	3.40	5.10	5.42
2-Choice	2.64	3.71	3.86
2-Left	1.82	2.56	2.66

Table 4: Theoretical Quantities for n values

Analysis

We can note that in our theoretical quantities, there is a trend that as n increases, the difference between each of our hash tables increases, with the 2-Left perfoming best at each of our different n values. In our investigation, this was largely the case with our 2-Left hash table having a shorter maximum chain than the other hash tables on average in each of our scenarios. However, there is an overlap in the confidence intervals when $n = 500$ between the 2-Choice and the 2-Left tables, which in turn means that when $n = 500$, we cannot determine that there is a statisitically significant difference in the max chain length of the two hash tables in this case. For $n = 500, 20000000$, there is a significant difference in the max chain length of each of our hash tables with the 2-Left table performing the best.