Programming Assignment 2: Hash Table Probe Analysis

Sean Barag
<sjb89@drexel.edu>

August 28, 2011

Programming assignment two required students to compare the open and closed hash tables as a means of storing a dictionary. This was measured in the average number of probes required to insert and delete fixed number of elements to a hash table with varying number of buckets.

1 Open Hashing

The complete Project Gutenberg Etext of *Alice's Adventures In Wonderland* was used as an input to the hash table. With the number of buckets ranging from just one up to 15,000 (an arbitrary point at which the average number of probes appeared to stabilize) at increments of 250, each word in the input was inserted into the table. Once all words had been successfully added, the input was again parsed so that each word could be deleted. The results of this experiment can be seen in Table 1.

B	Insert Total Probes	Delete Total Probes	Insert Average Probes	Delete Average Probes
1	85182903	89749809	2822.86926697	2486.97098759
250	370874	354549	12.2903632025	10.3333916237
500	197411	178054	6.54198700954	5.57394189832
750	132108	116625	4.37791622481	3.70050133266
1000	108861	84298	3.60753579003	3.00516915618
1250	100875	74059	3.34288838812	2.53314406896
1500	78721	58611	2.60872879109	2.25496306556
1750	74531	51440	2.46987672322	2.00272532607
2000	72315	46447	2.39644088017	1.9908701243
2250	59252	36705	1.96354718982	1.80910838385
2500	64694	35551	2.14388918346	1.70173759035
2750	59473	32954	1.97087089077	1.62575234336
3000	54585	31027	1.8088878579	1.52152805022
3250	57631	30673	1.90982900318	1.56999539336
3500	51873	27345	1.71901511135	1.43731931669
3750	49325	27350	1.6345771474	1.44525470302
4000	52278	25636	1.73243637328	1.4238267148
4250	46724	22667	1.54838282078	1.35560074158
4500	45270	22177	1.50019883351	1.34863780102
4750	47160	23324	1.56283138918	1.33968983343
5000	46325	18900	1.53516039236	1.27944760357
5250	44251	20304	1.46643027572	1.21203438395
5500	43998	20132	1.45804612937	1.27224469161
5750	41806	17413	1.38540562036	1.21735178971
6000	43213	18502	1.43203207847	1.24399919317
6250	43030	19479	1.42596765642	1.25719633406
6500	41869	17020	1.38749337222	1.17850713198
6750	39501	15217	1.30902041357	1.19988960732
7000	39729	15391	1.31657608696	1.16793140082
7250	47020	15940	1.55819194062	1.14453938393
7500	40456	14795	1.34066808059	1.17401999683
7750	40029	16200	1.32651776246	1.18093016475
8000	42967	17215	1.42387990456	1.21764040175
8250	39874	14917	1.32138123012	1.12589629406
8500	39589	14626	1.31193663839	1.11717079132
8750	38097	12398	1.26249337222	1.11623300621
9000	37258	13182	1.23468981972	1.12158597805
9250	41438	12762	1.37321049841	1.12065331928
9500	39011	14576	1.29278234358	1.21103356597
9750	38682	14760	1.28187963945	1.14135477884
10000	39054	12412	1.29420731707	1.11119068935
10250	36220	11764	1.20029162248	1.12628051699
10500	36358	11551	1.20486479321	1.09789943922
10750	39010	14729	1.29274920467	1.12935132648
11000	35758	11914	1.18498144221	1.08555808656
11250	36326	10789	1.20380434783	1.0965545279
11500	36250	12026	1.20128579003	1.07260078487
11750	37949	12937	1.2575888123	1.08677755376
12000	36712	12605	1.21659597031	1.07901044342
12250	35662	11924	1.18180010604	1.05859375
12500	36486	10834	1.20910657476	1.07246089883
12750	35065	11154	1.16201617179	1.0765370138
13000	35487	10318	1.17600079533	1.07289175419
13250	37766	9736	1.25152439024	1.05219928672
13500	35063	10956	1.16194989396	1.07972799842
13750	37340	11525	1.23740721103	1.06604384423
14000	34626	10414	1.14746818664	1.06657107743
14250	35173	12312	1.16559517497	1.04961636829
14500	38943	11923	1.29052889714	1.06894387664
14750	34147	9975	1.13159464475	1.07431340872
15000	35340	10186	1.17112937434	1.05162089614

Table 1: Probe data resulting from the open hashing of *Alice's Adventures in Wonderland*.

The expected trend for both insertion and deletion for this test is $O(1 + \frac{N}{B})$, where N is the number of elements being inserted (in this case, 28,198 words according to Bash `wc -l`). While the data certainly decays, it does so at a much faster rate than is expected, resembling a logarithmic function more closely than the provided formula. This is most likely due to the underlying data structure of a linked list, which requires $O(\log_2 n)$ time to traverse. Since each bucket potentially contains a fully qualified linked list and a probe was considered one comparison of a linked list node to the word in question, the measured logarithmic behavior is not very surprising.

2 Closed Hashing

Closed hashing was tested in a nearly identical way to open hashing: *Alice in Wonderland* was parsed, with each word getting added to the closed hash table where space allowed. At the end of the insertions, the story was again parsed so that each word could be deleted. B, the number of buckets, was varied from one to 34,000 in increments of 1,000. The resulting total and average number of probes is shown in Table 2.

B	Insert Total Probes	Delete Total Probes	Insert Average Probes	Delete Average Probes
1	30174	0	0.872534844717	0.0
1000	12767845	12626	369.204933202	0.365103232896
2000	14261813	7899	412.405673472	0.228413625586
3000	14034777	5249	405.840523972	0.151784165173
4000	11295085	4778	326.61745995	0.138164362963
5000	6209139	3768	179.548291018	0.108958417674
6000	241348	2298	6.97900641952	0.0664507547279
7000	32741	1384	0.946764212596	0.0400208200798
8000	19917	1186	0.575935457753	0.034295298132
9000	10662	780	0.308310681858	0.0225550864612
10000	11142	791	0.322190735064	0.0228731710138
11000	6770	556	0.195766583772	0.016077728298
12000	5295	457	0.153114336938	0.013214967324
13000	5377	435	0.155485512694	0.0125787982187
14000	4342	268	0.125556647967	0.00774969637384
15000	4102	406	0.118616621364	0.0117402116708
16000	3662	281	0.105893239257	0.00812561448152
17000	3120	320	0.0902203458447	0.00925336880458
18000	2862	196	0.082759817246	0.00566768839281
19000	2565	178	0.0741715343242	0.00514718639755
20000	2828	272	0.0817766468105	0.00786536348389
21000	2586	177	0.074778786652	0.00511826962003
22000	2168	137	0.062691573651	0.00396159851946
23000	2518	170	0.072812445781	0.00491585217743
24000	1972	111	0.0570238852582	0.00320976230409
25000	2294	252	0.0663350876178	0.00728702793361
26000	1768	101	0.0511248626453	0.00292059452895
27000	1576	88	0.0455728413626	0.00254467642126
28000	1822	62	0.0526863686311	0.00179284020589
29000	1548	62	0.0447631715922	0.00179284020589
30000	1577	81	0.0456017581401	0.00234225897866
31000	1308	37	0.0378231449887	0.00106992076803
32000	1619	65	0.0468162627957	0.00187959053843
33000	1316	71	0.0380544792088	0.00205309120352
34000	1272	50	0.0367821409982	0.00144583887572

Table 2: Probe data resulting from the closed hashing of *Alice's Adventures in Wonderland*.

Insertion for closed hash tables is expected to require approximately $\frac{1}{2} \cdot \left(1 + \frac{1}{1-\frac{N}{B}}\right)$ probes, whereas deletion should need roughly $\frac{1}{2} \cdot \left(1 + \frac{1}{\left(1-\frac{N}{B}\right)^2}\right)$ probes of the table. It is clear in cases where $B < N$ that

this should result in a negative number of probes, something that is quite obviously not possible and which did not occur. This does not occur for deletions as a result of the polynomial in the denominator. Of note in particular is the fact that zero probes were required for a bucket of size one, as only one word can be stored in such a dictionary.

3 Conclusions

In conclusion, open hashing has a much more predictable behavior than closed hashing, at least on the scope of this trial. The drastic error in results shown here is most likely due to the low (relative to the total word count) number of buckets used throughout the experiment.