Jooseung Song

CS 435 Programming Assignment Report #1

7 October 2019

1. For the classes BloomFilterFNV explain the process via which you are generating k-hash values, and the rationale behind your process.

Generated hash value of String s by setting initialValue and prime and XOR with ^= and *= operators. To make hash functions not identical and generate k different hash values, generated random int value n and added it to back of the String s.

2. The random hash function that you used for the class BloomFilterRan, explain how you generated k hash values.

For generating k hash values with BloomFilterRan, each random hash function generates its own random values of a and b. These integers lets each hash function generates unique hash value.

3. The experiment designed to compute false positives and your rationale behind the design of the experiment.

First, generate each one instances of each BloomFilterFNV, BloomFilterRan, and DynamicFilter. For each filter, I add same sets of random strings to each using random string generator, thus each filter should now contain same set of random strings. After that, for total number of tests, counted all the random strings that each bloom filters return true but are not in the string sets that were previously added to the bloom filters. Probability of each bloom filter is percent = ((double) false positives count / number of experiments) * 100.

4. For all the Bloom filter classes report the false probabilities when bitsPerElement are 4, 8 and 10. How do false positives depend on bitsPerElement? Which filter has smaller false positives? If there is a considerable difference between the false positives, can you explain the difference? How far away are the false positives from the theoretical predictions?

bitsPer Element	BloomFilterFNV			BloomFilterRan			DynamicFilter		
4	12.74%	13.74%	13.99%	19.31%	19.73%	19.74%	3.71%	3.57%	3.48%
8	2.04%	2.32%	1.94%	2.55%	2.12%	2.21%	0.19%	0.00%	0.00%
10	0.93%	0.74%	0.76%	0.93%	0.76%	0.91%	0.00%	0.01%	0.00%

In most of the cases, dynamic filter had the lowest probability and next was FNV. The reason BloomFilterRan has so many more false positives is because the other hash functions are deterministic. My algorithm to generate random strings probably isn't random enough.

5. Write a program to evaluate the accuracies of the methods from the class Statistics. Report accuracy results.

I referenced some information from bloom filter wiki. The function estimateSetSize was very accurate. However for both methods, one from the assignment pdf and the other from wiki page(https://en.wikipedia.org/wiki/Bloom_filter#The_union_and_intersection_of_sets), gave negative value sometimes.

6. Evaluate the (approximate) efficiency of the Bloom Filter for differential files application as follows: Recall that in this application that Bloom Filter is stored in the main memory and all other files are stored in secondary memory. Suppose that to access contents of a file in the secondary memory takes 1 second whereas accessing main memory takes 1 milli second. Compare the time taken by programs that use Bloom Filter and that do not use Bloom Filter. Use your experiment that compared the performances of NaiveDifferential and BloomDifferential to arrive the times.

```
time in micro second
Naive, Bloom, InDiff, InFilter
3775,  1771,  false, false
4090,  2071,  false, false
5385,  3299,  false, false
6076,  4096,  false, false
7726,  5839,  false, false
9912,  8498,  false, false
10997, 8770,  false, false
11736, 11607, false, true
12370, 10590, false, false
1074,  1127,  true,  true
15905, 13480, false, false
18034, 16512, false, false
```

For the test of 12 cases, in most of the cases, method that use Bloom Filter took less time compared to naive way. Bloom filter could make a noticeable speedup of about 15% amount. It seems that the Bloom filter used in BloomDifferential was able to prevent many potential accesses to DiffFile.

7. Consider the computation of Distributed Join on the relations given. What would be the communication cost (in number of bytes) if a naive strategy was followed (Sending one of the tables to the other)? What would be the communication cost using Bloom Filter? You do not have to give exact communication cost, an estimate suffices.

In case of the methods built in this assignment, bloom filter will reduce communication cost from sending whole table to just sending one category relation and filter bitset. Also, it will reduce time taken just as we tested in EmpiricalComparison Class.