CS5330 Implementation Project 1

Reinaldo Maslim -
Vicknesh S/O Jegathesan - A0113973A
Xu Yunze -

April 26, 2020

Contents

Chapter 1

Introduction

The purpose of this implementation project is to test variations of the count sketch implementation for summarizing data streams. Using the Count-Min Sketch data structure, we analyse the query, insert and delete operations performed on the data streams. The goal would be to optimize the time and space needed for the operations on the data structure.

Count Min Sketch was first implemented to summarise large amounts of data streams via a high dimensional vector and perform operations on this data structure. These operations can be used to answer a variety of queries such as heavy hitters, quantile estimation and joint size estimation. The structure has an efficient update and delete process which makes it highly suitable for data streams.

Chapter 2

Preliminaries

The data structure itself is a 2 dimensional array with width w and depth d. Given the parameters $\epsilon and \delta$ we choose the width and depth. The d hash functions chosen are 2 universal.

Definition 1 *H is a 2-universal if for all $x, y \in U$ such that $x \neq y$ we have*

$Pr_{h-H}(h(x) = h(y)) \leq \frac{1}{n}$

where $h - H$ means that h is selected uniformly at random from H.

When an update arrives (i_t, c_t), then c_t is added to one entry in each row of the array count. In other words, for $1 \leq j \leq d \quad count[j, h_j(i_t)] < -count[s, j, h_j(i_t)] + c_t$.

In our report $\tilde{f}$ is the frequency vector and it is given by $f_j = \sum_{i:x_i=j} c_i$

The types of queries that can be answered by the count min sketch data structure is given below. Consider a vector $[x_1, \ldots, x_n]$

Point Queries: Given a i, return an approximation for x_i.

Range Query: Given l,r, return an approximation of $\sum_{i=l}^{r} x_i$

Heavy Hitters Query: Given $\phi \in (0, 1)$. a heavy hitter is given by $x_i \geq \phi||x||$. We return all the approximate heavy hitters i such that $x_i \geq (\phi - \epsilon)||x||$.

Quantile Query: Given $\phi \in (0, 1)$, return j such that $(\phi - \epsilon)||x|| \leq \sum_{i=1}^{j} x_j \leq (\phi + \epsilon)||x||$

Chapter 3

Algorithm Description

Here we will introduce the pseudo code for the algorithms used in the analysis.

3.1 Count Min Sketch

```
Initialize
k $< -$ $\lceil \frac{2}{\epsilon} \rceil$
t $< -$ $\lfloor \log_2(\frac{1}{\delta}) \rfloor$
c $< -$ t x k integer array, with all entries initiated to 0
Choose $h_1, \ldots, h_t : [M]- > [k]$ independently from a 2 universal family of hash functions.

for each i do
    for s = 1 to t do
        $C[s, h(x_i)] < -C[s, h(x_i)] + c_i$
    end for
end for
```

Output: $\tilde{f} < -(C, h_1 \ldots, h_t)$

$\tilde{f}_j = minC[s, h_s(j)] : s = 1 \ldots t$

3.2 Count Median Sketch

Chapter 4

Implementation

Chapter 5

Conclusions

Bibliography