streaming lib.

金奕成 黄道吉 杨天正

算分29班

March 23, 2018

content

streaming algorithm

- 流算法(streaming algorithm)在网络流, 数据库和文本处理当中有比较重要的应用
- 两种重要的方面是hashing和sketch, 我们将完成这两个部分的一些基本算法, 并封装成库.

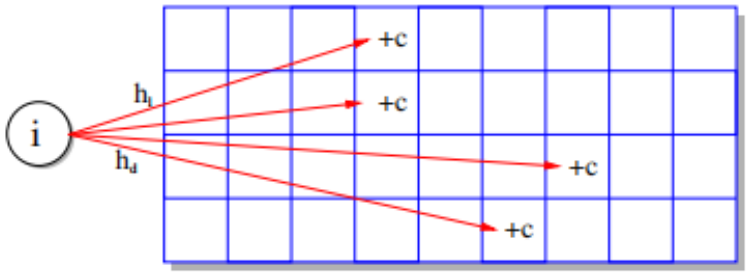
references

目前主要参考以下几篇论文

- Cormode, Graham. “Sketch Techniques for Approximate Query Processing.” (2010).
- Charikara, Moses and Martin Farach-Coltonc. “Finding frequent items in data streams.” (2003).
- Muthukrishnan, S.. “Data streams: algorithms and applications.” SODA (2003).

Frequency based sketch

- Count Sketch, Count-Min Sketch是两种常用的基于频率的sketch



- Fig. 1.3 Count-Min sketch data structure with $w = 9$ and $d = 4$
- 此外也有AMS sketch之类的sketch.

Distinct Count Estimation

- 主要有Flajolet-Martin Sketch, k Minimum Value estimator之类的算法
- 基于hash还可以将前面的算法做到更好
- 支持一些基础的skipping, sampling技术.

goals

我们的目标是实现

- 主流, 常见的一些sketch, 及其变体
- 实现必要的hash函数支持
- 提供测试程序
- 提供对多种数据的支持(五元组)