Algorithms and Data Structures for Big Data

Hu Fu @ SUFE. Sept 16, 2021

Teaching Staff

- Instructor: Hu Fu 伏虎
- Office: 504 School of Information Management & Engineering
- Office hour: Tuesday 3-5pm or appointment by email
- Email: fuhu@mail.shufe.edu.cn
- Website: https://bb.shufe.edu.cn/

  http://www.fuhuthu.com/BigData2021/
- Teaching Assistant: Qun Hu 胡群
- Email: 2019212804@163.sufe.edu.cn

What is Big Data?

- Broadly speaking, big data is simply exceptionally large datasets
  - In popular media, it is sometimes synonym of machine learning with large training datasets
- Applications include:
  - Retail and wholesale trade
  - Banking and securities
  - Communications, media and entertainment
  - Healthcare

What is Big Data?

- Broadly speaking, big data is simply exceptionally large datasets
  - In popular media, it is sometimes synonym of machine learning with large training datasets
- Applications include:
  - Insurance
  - Government
  - Scientific research
  - Transportation…

Focus of this course

- This course focuses on basic operations on such datasets, such as
  - Accessing and storing such datasets
  - Estimating simple statistics
  - Extracting meaningful sketches to be used by upper level applications
- We do not look at upper level applications such as learning
  - For that you should take machine learning or statistical learning theory (the latter not offered this year)

(Tentative) Syllabus

- Review of basic probability theory
- Hashing
- Search trees
- Concentration Inequalities
- Skip lists and SkipNet
- Dimensionality Reductions
- Streaming Algorithms

Coursework

- Homework:
  - Students are encouraged to work in groups of up to 3 people
  - Everyone should be able to explain the solutions turned in
  - Typesetting your solutions is highly encouraged
- Project: literature survey on a chosen topic (I can provide candidate topics)
  - Done in groups of up to 4 people
  - Presentation at the end of the semester
- Take-home final: 1-3 days' work, done independently. Time TBD
- Grade makeup: 45% homework + 20% project + 35% final

Prerequisites

- We will assume basic familiarity of data structures and algorithms
  - At the very least, you should have some rough idea on how computer programs work
  - Comfortable with basic running time analysis (e.g. familiarity with the big $O(\cdot)$ notation)
  - Knowledge of basic data structures. We will use arrays, linked lists, trees.
  - Comfort with basic probability theory will go a long way, but is not strictly required. We start with a quick review.

This is a *Theory* course

- All materials are proof-based, and so is the homework
- Implementation of algorithms is not required; coding things up may help with understanding
- Mathematical maturity helps
  - Grasping the mathematical essence is often more important than the “knowledge”
  - Ideas, intuitions, tricks, facts

A Brain Teaser

- The following problem gives a taste of streaming algorithm
  - As datasets grow large, traditionally acceptable running time or memory usage may no longer be practical
  - It's pretty challenging if you are seeing it for the first time
- Say you have a very large array of n entries, but strictly more than half of them have the same content. Design an algorithm to find out this content.
  - Your algorithm must run in linear time ($O(n)$ time)
  - You have only $O(1)$ memory

One Solution

- Keep the content of an entry (initiated to empty) and a counter (initiated to 0), and go over the array
- At each new entry, do the following:
  - If the counter is 0, copy the current entry's content to the stored content, and set the counter to 1
  - Otherwise, compare the current entry's content and the stored content
    - If they are the same, counter++; otherwise counter--
- At the end, output the stored entry.