Big Data Algorithms

Xiaotie Deng

AIMS Lab
Department of Computer Science
Shanghai Jiaotong University

September 18, 2017

Course Information

Data Ownership

- Data Ownership:http://www.niu.edu/rcrportal/datamanagement/dotopic.html
- Data Acquisition and Originality
- Grammar Check: https://www.grammarly.com/
- Similarity Check: https://guides.turnitin.com/01_Manuals_and_Guides/Instructor/Class
- Plagiarism Check: http://smallseotools.com/plagiarism-checker/
- Value or Cost (entropy?) of Data?

Orientation in bigdata algorithm design

- algorithms for problems vs for data
- public/referential data integrity and blockchain
- polynomial time vs one pass algorithms
- the ranking and quantile problem
- mapping and navigation in networks
- dna sequence and computational genomics
- classification (supervised and unsupervised)
- approximation and randomization and parallelization

Learning in a PhD Course

- Help from MOOC. Examples:
  - https://www.edx.org/course/machine-learning-data-science-analytics-columbiax-ds102x-0
  - Find your favorite ones, and share.
  - Mooc courses are complementary reading/watching study materials but NOT the teaching materials in this course.
- Research materials
  - Develop research topics in bigdata algorithms.
  - Everyone is expected to develop a focus topic (in bigdata algorithms) you become an expert on.
- Study=learn+research

Assessment Information

- 50% coursework plus 50% final examination
- Coursework:
  - 4 Assignments, 5% each, maximum 20%
  - 1 Middle term test, maximum 20%
  - Extra-ordinary work (For A^+ work) maximum 10%.
    - One project leading to a publishable research paper.
    - Student expert self study: Script one lecture note and give a suitable exam question (not searchable from Internet) with a correct solution.
- Final Examination: Problems given out 24 hours before examination and each redoes it in class. Two parts separately marked and both counted in final examination evaluation.

Teaching Style

- To teach key issues in intended outcomes.
  - Teach it simple and clear. Avoid complicated subjects/techniques.
  - Open ended questions are left to expert student requirement and project, out of students' own interests.
- Expert student requirement: A crystal clear understanding in one subject matter, close to an expert level, ability to venture into research.
- Project: Marked for its suitability of submission to a quality journal/conference

Contact Information

- Classroom: Chen Rui Qiu Building, 312.
- Time: Mondays and Wednesdays 10:00-11:40am
- Lecturer: Professor Xiaotie Deng
- Phone number: 34208027
- Office: 3-428
- Email: deng-xt@cs.sjtu.edu.cn
- TA: Someone

References

- Harvard Course http://people.seas.harvard.edu/ minilek/cs229r/fall15/lec.html
- Jimmy Lin and Chris Dyer, Data Intensive Text Processing with Map Reduce, 2010.
- Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. Mining of Massive Datasets, 2014.
- Charu C. Aggarwal and Philip S. Yu. A Survey of Synopsis Construction in Data Stream, 2006.
- Probabilistic Data Structures for Web Analytics and Data Mining. 2012
- Damian Cryski, Probability Data Structure for Go, 2014

Online Resources

- Big Data Structure http://www.structuredata.com/data-2016/about/
- http://www.dummies.com/how-to/computers-software/Big-Data/Engineering.html
- http://www.dummies.com/how-to/computers-software/Big-Data.html
- http://stackoverflow.com/questions/11094645/what-data-structure-to-use-for-big-data
- Big Data Analytics

  http://data-informed.com/rethink-your-org-chart-for-big-data-analytics-teams/

Course Intended Learning Outcomes

Big data has posed a great challenge to algorithmic studies for the delivery of efficient solutions to process massive data, especially with real time data analysis needs. This course aims to provide students with

1. the conceptual framework for good algorithms, and approximate answers in big data analysis.
2. the proper algorithmic tools for the design and analysis of algorithms for big data, including
   1. randomized algorithms,
   2. probabilistic data structures and techniques
3. Empirical framework to solve practical problems and the corresponding evaluation criteria.

Syllabus

- Bigdata algorithms: stream algorithms, randomized algorithms, approximation algorithms, secondary memory algorithms, bsp algorithms, map reduce algorithms;
- Big Data Structure: Probabilistic data structure such as Bloom filters, hyperLogLog, count-min sketch;
- Big Data Analytics; dimension reduction, compressed sensing, matrix sparsification;
- Internet applications such as median and quantile, exploring/mapping/querying, genomics, recommendation system, computational advertising, large-scale machine learning.

Tentative Schedule

- Input Size: hard disk memory and stream algorithms
- Output Size: reduced database
- Data Type
  - Single data: Kolmogorov complexity, data sketching, instance-specific algorithms
  - Dynamic data: instance-optimality, phase transition, online algorithms
- Data Structures and External Memory
  - B-trees, Bloom Filters, Merkle Trees
  - Secondary Memory Algorithms: Funnel Sorting
- Data Issues
  - Data model: classification and prediction
  - Statistical Model: confusion matrix
- Data Processing Efficiency
  - Approximate linear algebra: Principle Component Analysis

Selected Topics in Technology and Applications

- Sorting and Webpage Ranking
- GPS shortest path (fastest path/optimal path)
- Match.com scheduling/pricing (ad pricing, ad proposals, user happiness prediction)
- Originality, Similarity and Plagiarism Test
    - https://btckan.com/news/topic/22654
- Bio-3D Reconstruction
- Blockchain Coding
- Machine learning algorithms
- Unmanned-flight algorithms

Projects

- Mapping: Unmanned-flight/walk
- Blockchain: Assignment of original ideas in a discussion group.
- Algorithmic efficiency in the identification of Caocao's descendants. https://en.wikipedia.org/wiki/Cao_Cao

Algorithmic Issues in Data Science

What Data?

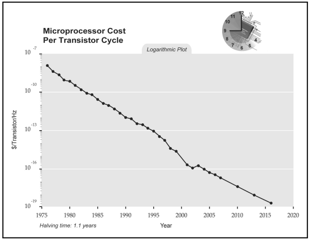
1. Permanent data: The example of Voyager Golden Record.
   - The first images show mathematical and physical quantities, the Solar System and its planets, DNA, and human anatomy and reproduction.
2. Data in Action: http://www.huffingtonpost.com/steve-rosenbaum/data-through-the-ages_b_1025913.html
   - Observation: data gathered five humans sensors and computer connected sensors
   - Mapping: ordered data through organizing tools
   - Prediction: models that turn data into reproducible algorithms
   - Control: designs of a path pointing to the desirable future
   - Reaction: best response of environment and agents with respect to control parameters
   - Equilibrium: Harmonic data sets

Folk's Laws of computing

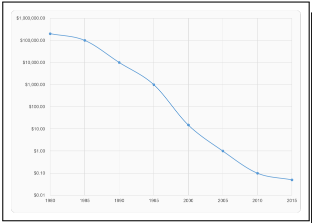
1. (Gordon) Moore's law: chip performance doubles every 18 months (David House)
2. (Mark) Kryder's law: hard disk space had a 50 million-fold increasing from 2000 to 100 billion bits in one square inch during a period of 50 years (1956-2005).
   - Walter, Chip (August 2005). "Kryder's Law". Scientific American.
3. Power Law: $f(x) = ax^{-k}$ for a wide variety of physical, biological, and man-made phenomena (and virtual?).
4. References: WikiBooks of Data Science–An Introduction/A History of Data Science, Wiki: Moore's law, Power Law.

Progress of Computer, Storage, Algorithms(Genomics)

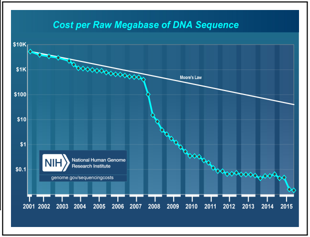
Per CycleCost



Harddisk Cost



DNA Sequencing Cost



The Economics of Big Data

1. Parkinson's Law: "Data expands to fill the space available for storage"(?)
   - Parkinson, Cyril Northcote (November 19, 1955). "Parkinson's Law". The Economist.
   - Mikhail Gorbachev: Parkinson's law works everywhere
     - Reference: O'Sullivan, John (June 2008). "Margaret Thatcher: A Legacy of Freedom". Imprimis. Hillsdale College. 37 (6): 6.
2. New Challenge: analyzing and utilizing big data collected by government and corporations, pending efficient algorithms.
3. Are those laws demanded by the economics of obsoleting equipments?
   - "WEEE ? Combating the obsolescence of computers and other devices". SAP Community Network. 2012-12-14. http://scn.sap.com/docs/DOC-34208

The Evolving Role of Data

1. Data: Numbers and Counting:
   - Alex the African grey parrot could count up to six, in "More Animals Seem to Have Some Ability to Count". By Michael Tennesen, Scientific American, September 1, 2009.
   - According to a chinese idiom, a monkey know 4 is larger than 3 but does not know $3 + 4 = 4 + 3$.
   - When can we start teaching counting?
   - http://vedicsciences.net/articles/history-of-numbers.html
2. Database: Conceptual Abstraction.
3. Base of Scientific Discovery:
   - The End of Theory: The Data Deluge Makes the Scientific Methods Obsolete
   - Chris Anderson, Science 06.23.08, 12.00pm

Data Dynamics

1. Persistent data is or is likely to be in the context of the execution of a program.
2. Static data is in the context of the business historical data, regardless of any one application or program.
3. The "dynamic" data is the new/updated/revised/deleted data in both cases, but again over different time horizons.
4. An example: Your paycheck stub is dynamic data for 1 week, or 1 day, then it becomes read-only and read-rarely, which would be either and both static and persistent.
5. Reference: https://en.wikipedia.org/wiki/Dynamic_data

Dealing with Big Data

1. Observation: Acquisition and Accessing
2. Mapping: Data Structure
3. Prediction: Machine Learning
4. Actions
    - Control: Dynamic systems
    - Reaction: Optimization
    - Equilibrium: Best plays among participants

Algorithmic Design and Analysis

1. The big-O notation and polynomial time algorithms
2. Stream algorithms
3. Parallel algorithms
4. Bayesian algorithms
5. Approximation algorithm
6. Neural network and deep learning algorithms

Evolution of Algorithms

1. Constant memory algorithm: Monkeys can count: http://www.sciencemag.org/news/2014/04/monkeys-can-do-math
2. Ordinality: 4 wallnuts in the morning and 3 in the afternoon.
3. External memory: Counts of 10s, 100s, 1000s, 10000s.
4. Size of CPU: constant, polylog, polynomial in terms of memory space
5. Memory structure: Cache, memory, hard-disk.

Algorithmic Thinking

1. Algorithmic Ideas Independent of Specific Platforms/Models
2. Simultaneously Consider Implementation for Different Platforms.
3. Randomized Algorithm Methods All from the Beginning (statistical parameters)
4. Bayesian Setting from the Beginning (and Heuristic Work for Them, Powerlaw)
5. Generic Ideas of Algorithmic Design
6. Basic Algorithms Aligned with The Above classification of data.

Median and Ranking

Median in Linear Time

1. Recursive equation.
   - Blum, Floyd, Pratt, Rivest and Tarjan (1973)
   - Munro and Paterson $\theta(N^{1/p})$ memory for finding a median in p passes of data(1980).
2. Randomization.
3. Query in $\log(N)$ Time

Median in SubLinear Time?

One pass algorithm for the median

Munro and Paterson Algorithm

- J. I. Munro and M. S. Paterson, Selection and Sorting with Limited Storage, Theoretical Computer Science, vol. 12, pp. 315-323, 1980.
- Keep a memory of size s.
- Read the n numbers one by one
- maintain s of them in memory and discard one each time
- Find the median of the s number in the end and report it.

Selection Policy

- Set $H = L = 0$ initially, representing the sets of numbers already removed as higher and lower than the median.
- Insert the first s numbers in the set S.
- Sort S.
- If the new number is larger than $\max(S)$ or smaller than $\min(S)$ remove it to place in H or L accordingly
- If the new number is in $(\min(S), \max(S))$, then keep it and remove $\max(S)$ or $\min(S)$ to make L or H more balanced.

Correctness and Complexity

Analysis

- Each datum is read into memory once.
- At all the time, $\forall i \in L, \forall j \in S, \forall k \in H, i < j < k$.
- Algorithm terminates with the median found if
  - $|H| \leq n/2$ and $|L| \leq n/2$.
- How big should $|S|$ be to satisfy this condition with high probability?

Random Permutation Model

- Random Permutation Model
  - Data enter the memory as a random permutation.
- Balanced Condition:
  - $d = |H| - |L|$
  - Starting at zero until there are S items in the memory
  - $|H|$ or $|L|$ increases by one at each of the next steps, which happens at probability $1/2$ each.
- D follows the standard random walk
- $E(|S_n|) \to \sqrt{\frac{2}{\pi} \cdot n}$ (http://mathworld.wolfram.com/RandomWalk1-Dimensional.html)

Complexity of The Algorithm

- Hard disk size n
- One read of each datum
- Memory size $O(\sqrt{n})$
- J.I. Munro, and M.S. Paterson. SELECTION AND SORTING WITH LPMITED STORAGE. TCS 12 (1980), 315-323.

Random Walk on the Line

Simple Random Walk

- $S_n = \sum_{j=1}^{n} Z_j$, $S_0 = 0$.
  - Z_js are iid (*identical independent distribution*) random variables,
  - all uniform in $\{0, 1\}$: $Pr(Z_j = 1) = Pr(Z_j = -1) = 1/2$.
- Properties
  - $E(S_n) = 0$
  - $E(S_n^2) = n$
  - $E(|S_n|) \to \sqrt{\frac{2}{\pi} \cdot n}$ (http://mathworld.wolfram.com/RandomWalk1-Dimensional.html)

Length of Random Walk Model

- http://mathworld.wolfram.com/RandomWalk1-Dimensional.html
  - With $1/2$ probability a new item is in H/L.
  - With High probability, length of S is $O(\sqrt{n})$.