Big Data Algorithms

Xiaotie Deng

AIMS Lab
Department of Computer Science
Shanghai Jiaotong University

October 3, 2017

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
  - $E(S_n) = 0$
  - $E(S_n^2) = n$
  - $E(|S_n|) \rightarrow \sqrt{\frac{2}{\pi} \cdot n}$ (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 $O(\sqrt{n})$.