大数据算法

第二讲 亚线性算法概述

哈尔滨工业大学
王宏志
wangzh@hit.edu.cn

本讲内容

2.1 亚线性算法的定义

2.2 水库抽样—空间亚线性算法

2.3 平面图直径—时间亚线性计算算法

2.4 全0数组判定—时间亚线性判定算法

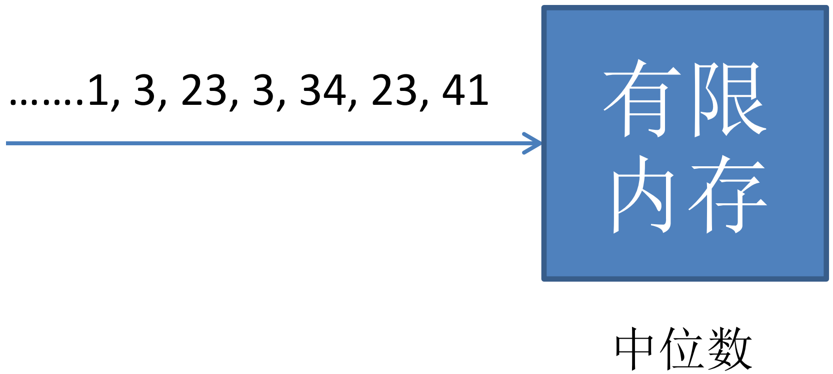
亚线性的含义

- **时间/空间/IO/通讯/能量等消耗是o(输入规模)**
- **亚线性时间算法**
    - 亚线性时间近似算法
    - 性质检测算法
- **亚线性空间算法**
    - 数据流算法

亚线性时间问题

- **给定一个社交网络，如何平均每个人的朋友个数，即在大图中计算其结点的平均度**
- **能否在不访问所有顶点的情况下完成此任务？**
  - 精确计算需要访问最少n-1个顶点
  - 是否可以简单的抽样？

亚线性空间问题



- **一个（源源不断到来的）数据集合(流)，只能扫描一次，如何求其中位数？**
  - ➢ 不能存储所有数据->不能对其进行排序
  - ➢ 应当存储哪些数据？

本讲内容

2.1 亚线性算法的定义

2.2 水库抽样—空间亚线性算法

2.3 平面图直径—时间亚线性计算算法

2.4 全0数组判定—时间亚线性判定算法

水库抽样——一个亚线性空间算法

- **输入：一组数据，其大小未知**
- **输出：这组数据的k个均匀抽样**
- **要求：**
  - 仅扫描数据一次
  - 空间复杂性为$O(k)$
  - 扫描到数据的前n个数字时（$n>k$），保存当前已扫描数据的k个均匀抽样

水库抽样算法

1. 申请一个长度为k的数组A保存抽样
2. 保存首先接收到的k个元素
3. 当接收到第i个新元素t时，以k/i的概率随机替换A中的元素(即生成[1,i]间随机数j, 若$j \le k$, 则以t替换$A[j]$)

性质1：该采样是均匀的

$$\frac{k}{i} \times \left(1 - \frac{1}{i+1}\right) \times \left(1 - \frac{1}{i+2}\right) \times \cdots \times \left(1 - \frac{1}{n}\right) = \frac{k}{n}$$

性质2: 空间复杂性是O(k)

本讲内容

2.1 亚线性算法的定义

2.2 水库抽样—空间亚线性算法

2.3 平面图直径—时间亚线性计算算法

2.4 全0数组判定—时间亚线性判定算法

平面图的直径——一个亚线性时间计算算法

- **输入：*m*个顶点的平面图，任意两点之间的距离存储在矩阵*D*中，即点*i*到点*j*的距离为D_{ij}**
  - 输入大小是$n=m^2$
  - 最大的D_{ij}是图的直径
  - 点之间的距离对称且满足三角不等式
- **输出：该图的直径和距离最大的D_{ij}**
- **要求：**
  - 运行时间为$o(n)$

平面图的直径近似算法

- **无法在要求的时间内得到精确解，寻找近似算法**
- **近似算法**

1. 任意选择$k \leq m$
2. 选择使得D_{kl}最大的l
3. 输出D_{kl}和(k, l)

- **近似比**

$$D_{ij} \leq D_{ik} + D_{kj} \leq D_{kl} + D_{kl} \leq 2\, D_{kl}$$

因而近似比为2

- **运行时间**

$$\mathbf{O}(m)=\mathbf{O}(\sqrt{n}) = o(n)$$

近似算法

- 什么是近似算法
  - 近似算法主要用来解决优化问题
  - 能够给出一个优化问题的近似优化解的算法
- 近似算法解的近似度
  - 问题的每一个可能的解都具有一个代价
  - 问题的优化解可能具有最大或最小代价
  - 我们希望寻找问题的一个误差最小的近似优化解
- 我们需要分析近似解代价与优化解代价的差距
  - Ratio Bound
  - 相对误差
  - $(1+\varepsilon)$-近似

近似比

- **Ratio Bound**

设A是一个优化问题的近似算法, A具有ratio bound $p(n)$, 如果

$$\max\left\{\frac{C}{C^*},\frac{C^*}{C}\right\}\le p(n)$$

其中n是输入大小, C是A产生的解的代价, C^*是优化解的代价.

- 如果问题是最大化问题, max$\{C/C^*, C^*/C\}=C^*/C$
- 如果问题是最小化问题, max$\{C/C^*, C^*/C\}=C/C^*$
- 由于$C/C^*<1$当且仅当$C^*/C>1$, Ratio Bound不会小于1
- Ratio Bound越大, 近似解越坏

- **相对误差**

相对误差: 对于任意输入, 近似算法的相对误差定义为$|C-C^*|/C^*$, 其中C是近似解的代价, C^*是优化解的代价.

相对误差界: 一个近似算法的相对误差界为$\varepsilon(n)$, 如果$|C-C^*|/C^*\le\varepsilon(n)$.

本讲内容

2.1 亚线性算法的定义

2.2 水库抽样—空间亚线性算法

2.3 平面图直径—时间亚线性计算算法

2.4 全0数组判定—时间亚线性判定算法

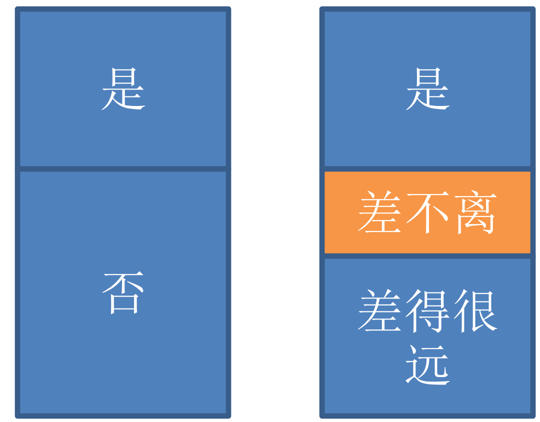
全0数组的判定——一个亚线性时间判定算法

- **输入：包含n个元素的0,1数组A**
- **输出：A中的元素是否全是0**
- **要求：**
  - 运行时间为$o(n)$

判定问题的近似

- 无法在要求的时间内得到精确解，寻找近似解
  - ➢ **判定问题如何近似？**
- 输入满足某种性质或者**远**非满足此性质

问题：图片中是否包含“猫”



- ε-远离
  - ➢ 对于输入x，如果着从x到L中任意字符串的汉明距离至少为 $\varepsilon|x|$，则x是ε-远离 L的.
- 全0数组判定问题的近似
  - ➢ **是否A=00…0或者其包含1的个数大于εn?**

全0数组的判定近似算法

- **算法描述**

1. 在A中随机独立抽取s=2/ ε个位置上的元素
2. 检查抽样，若不包含1，则输出“是”，若包含1，则输出“否”

- **判定精确性分析**
  - **如果A是全0数组，始终输出”是”**
  - **如果A是ε-远离的，Pr[error]=Pr[抽样中没有1]$\leq(1-\varepsilon)^s\approx e^{-\varepsilon s}=e^{-2}<\frac{1}{3}$**

- **运行时间：**$O(s)$

- **证据引理**

如果一次测试以大于等于p的概率获得一个证据，那么s=2/p轮测试得到证据的概率大于等于2/3

判定算法的定义

- **对于判定问题L，其查询复杂性为q(n)和近似参数ε的性质测试算法是一个随机算法，其满足对于给定L的是一个实例x, 最多进行q(|x|)次查询，并且满足下述性质：**
  - **如果x在L之中，该算法以最少2/3的概率返回“是”**
  - **如果x 是ε远离L的, 该算法以最小2/3的概率返回“否”**

致谢

- 本讲义部分内容来自于Qi Zhang和Sofya Raskhodnikova的讲义