第2届全国高校大数据教学研讨会
http://dblab.xmu.edu.cn/post/bdts2018

大数据算法

从研究到教学的实践

报告人：王宏志
wangzh@hit.edu.cn
http://homepage.hit.edu.cn/wang

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

目录

目录

大数据


BIG
DATA
STORAGE
ANALYTICS
TECHNOLOGIES
INTERNET
SOFTWARE
USE
ONE
SINCE
USING
TYPES
WORKING
GARTNER
MASSIVELY
BIOLOGICAL
HUNDREDS
CREATED
RELATED
ALSO
PERFORMANCE
DISK
ZETTABYTES
SHARED
TIME
COMPUTING
TOOLS
SET
GENOMICS
WITHIN
PROCESS
COMBAT
SIGNIFICANT
RELATIONAL
INCLUDE
SYSTEMS
DEFINITION
BIOGEOCHEMICAL
NETWORKS
SEARCH
RECONSIDER
OPPORTUNITIES
INFORMATION
CONNECTOMICS
RECORDS
COST
CONTINUES
CITATION
TERABYTES
MPP
RESEARCH
COMPLEX
DATABASES
INDEXING
DIFFICULTY
TARGET
ABILITY
SETS
SENSOR
ARCHIVES
EXAMPLES
APPLIED
AMOUNT
DESCRIBING
ELAPSED
CURRENT
THOUGHT
EVERY
LARGER
CASE
STORE
DEFINING
SOCIAL
LARGE
MANAGE
GROW
MAY
MANAGEMENT
USED
DISTRIBUTED
CAPTURE
PETABYTES
SIZE
NEEDED
QUALITIES
PROCESSING
BURIED
LOGS
PARALLEL
TOLERABLE
SAN
SOLID
WIRELESS
RADIO-FREQUENCY
COMPLEXITY
UBIQUITOUS
REQUIRING
BUSINESS
MOVING
ORGANIZATIONS
DESKTOP
CURRENTLY
FC
TENS
CAPACITY
FORMS
PRESENTATIONS
WORLD'S
PRACTITIONERS
NOW

大数据的定义和特点

大数据是通过传统数据库技术和数据处理工具不能处理的庞大而复杂的数据集合。

规模大 (Volume)

速度快 (Velocity)

类型多 (Variety)

价值密度低 (Value)

5亿用户
8亿商品
20亿PV/天

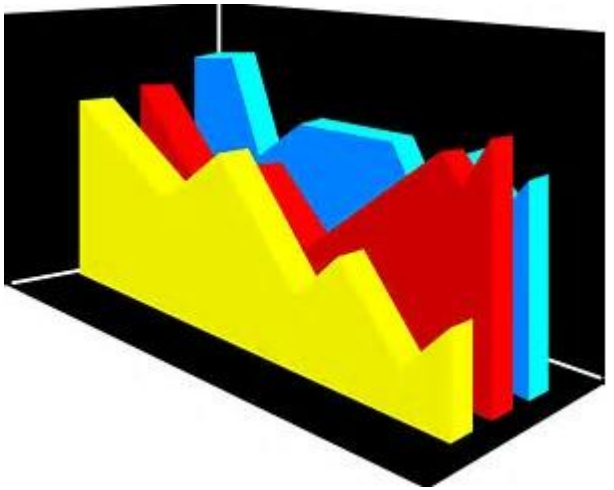
3万条/秒

5万订单/分钟

用户评论

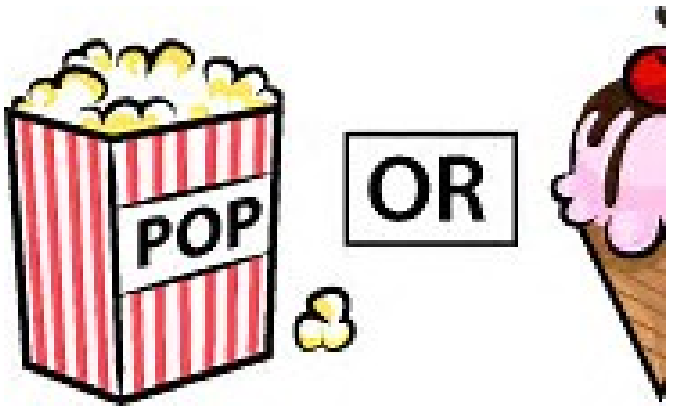
大数据需要的新思路

系统



建模

实现



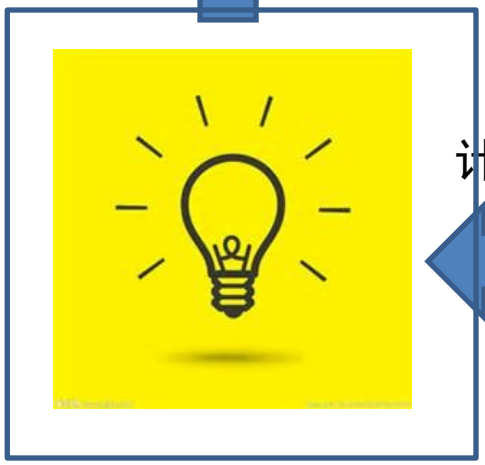
折衷

大数据思维

感知、采集

建模、抽象

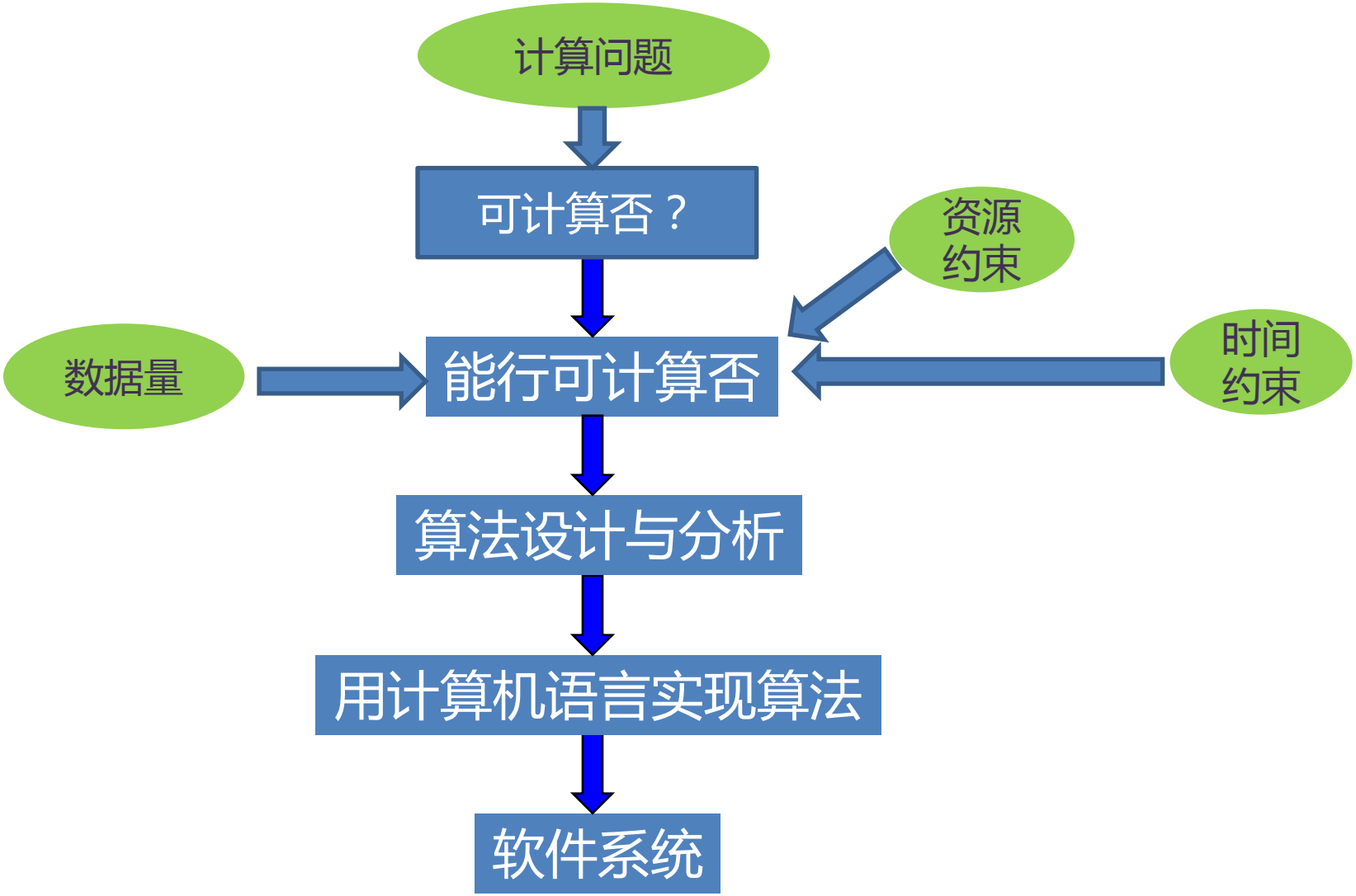
计算问题求解



分析结果回馈

大数据问题求解

大数据上问题求解计算问题的过程

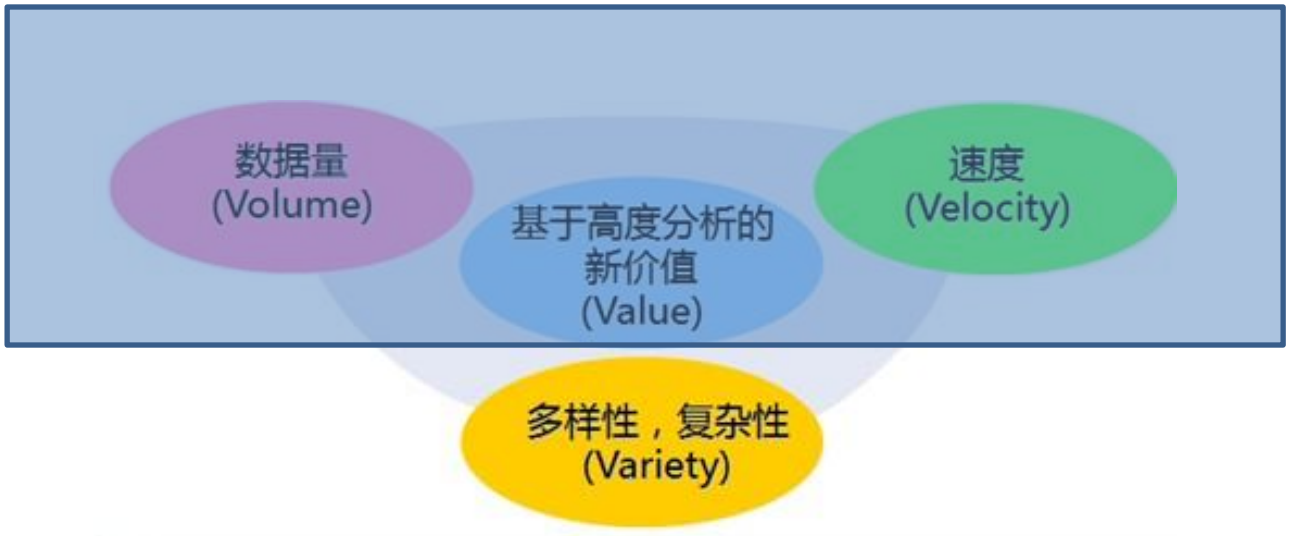


大数据算法

- 大数据算法的定义
  - 在给定的资源约束下，以大数据为输入，在给定时间约束内可以生成满足给定约束结果的算法。
- 大数据算法可以不是：
  - 精确算法
  - 串行算法
  - 内存算法
  - 仅在电子计算机上运行的算法
- 大数据算法不仅是：
  - 云计算
  - 大数据分析和挖掘的算法
  - MapReduce
  - 数据库中的算法

和“算法设计与分析”课程中算法大不相同

大数据的特点与大数据算法



大数据算法的难度

- 访问全部数据时间过长
  - 读取部分数据

时间亚线性算法

- 数据难于放入内存计算
  - 将数据存储到磁盘上
  - 仅基于少量数据进行计算

外存算法

空间亚线性算法

- 单个计算机难以保存全部数据，计算需要整体数据
  - 并行处理

并行算法

- 计算机计算能力不足或知识不足
  - 人来帮忙

众包算法

目录

大数据的算法设计技术

- 精确算法设计方法
- 并行算法
- 近似算法
- 随机算法
- 在线算法/数据流算法
- 外存算法
- 面向新型体系结构的算法
- 现代优化算法

- 时间空间复杂性
- IO复杂性
- 结果质量 (近似比、competitive ratio)
- 通讯复杂性

大数据算法

- 亚线性算法
- 外存算法
- 并行算法
- 众包算法

大数据算法课程内容

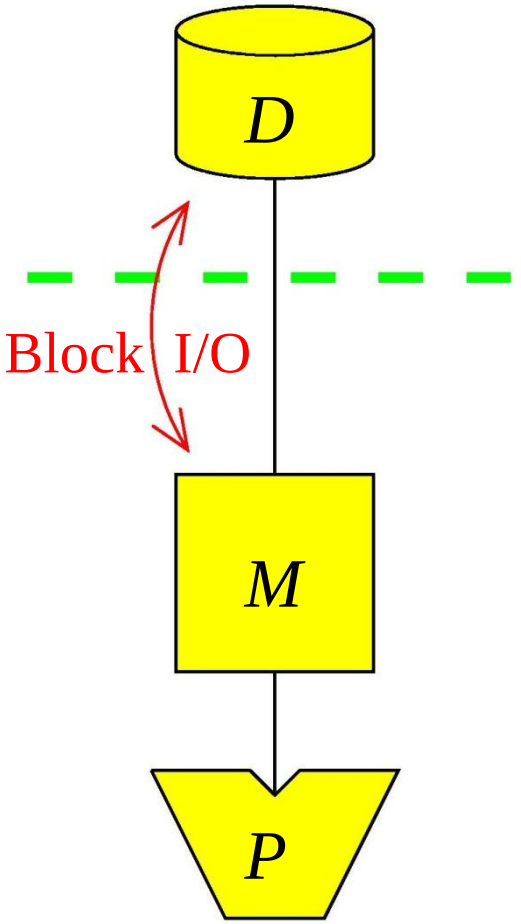
- 绪论
- 时间亚线性算法
- 空间亚线性算法
- 外存算法
- 外存图算法
- 外存数据结构
- MapReduce并行算法
- 非MapReduce并行算法
- 众包算法

大数据算法的例子

- 亚线性算法
  - 时间亚线性计算算法
    - 平面图直径
    - 最小生成树
  - 时间亚线性判定算法
    - 全0数组判定
    - 序列有序的判定
  - 空间亚线性算法
    - 水库抽样
    - 数据流中频繁元素

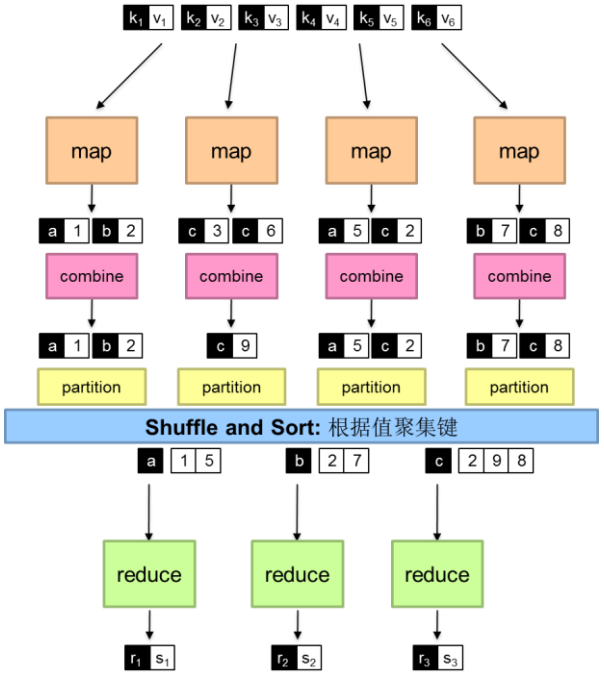
大数据算法的例子

- 外存算法
  - 外存排序(归并排序、分布排序)
  - 表排序
  - 欧拉回路
  - 最大独立集
  - 最小生成树
- 外存数据结构
  - 外存查找树
  - B树
  - KD树



大数据算法的例子

- MapReduce算法
  - Wordcount
  - 平均数计算
  - 词频计算
  - 连接算法
  - 单源最短路径
- Pregel算法
  - 单源最短路径
  - 子图匹配
- 改造平台上的MapReduce算法
  - PageRank
  - 传递闭包
  - K-Means



大数据算法的MOOC实践

中国大学MOOC 课程 名校 学·问 学校云 考研 客户端 搜索感兴趣的课程 登录 | 注册

首页 > 全部课程 > 计算机

大数据算法 国家精品

分享

第2次开课

开课时间： 2015年03月02日 ~ 2015年07月20日 当前开课已结束

学时安排： 3-5小时/周

已有57917人参加

已结束，查看内容

学堂在线 xuetangx.com 首页 课程 院校 微学位 学堂云 雨课堂 请输入课程、老师、学校 APP下载 注册 | 登录

大数据算法 随堂模式

来自于：哈尔滨工业大学 | 分类：计算机(383)

课程描述

大数据不论在研究还是工程领域都是热点之一，算法是大数据管理与计算的核心主题。本课程试图简要介绍大数据计算中涉及到的基本算法设计方法。适用于大数据研究与开发人员，也适用于数据科学爱好者。

开课时间：课程已完结 学习时长：5小时/周

课程进度：-- 报名人数：6650人

先修知识：算法设计与分析、概率论

关注课程

大数据算法的教学方法

- 经典方法和新方法相结合
- 着重介绍思想
- 典型问题作为案例
- 凝练知识点形成短视频
- 3成设计7成分析
- 按需补充知识
- 开放式题目+学生互评

大数据算法MOOC得失谈

- 优点
  - 填补空白
  - 面向前沿
  - 热门领域
- 不足
  - 理论太多
  - 难度太大
  - 实践不足

目录

水库抽样——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
- 算法描述
  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

性质2: 空间复杂性是O(k)

数组有序性判定——时间亚线性算法

- 输入：n个数的数组, $x_1, x_2, \ldots, x_n$
- 输出：这个数组是否有序？
  - 需要访问这n个数，时间是Ω(n)
- 近似版本
  - 这个数组是有序的还是 ε 远离有序的？
- ε 远离
  - 我们必须删除大于 εn 个元素才能保证剩下的元素有序

算法

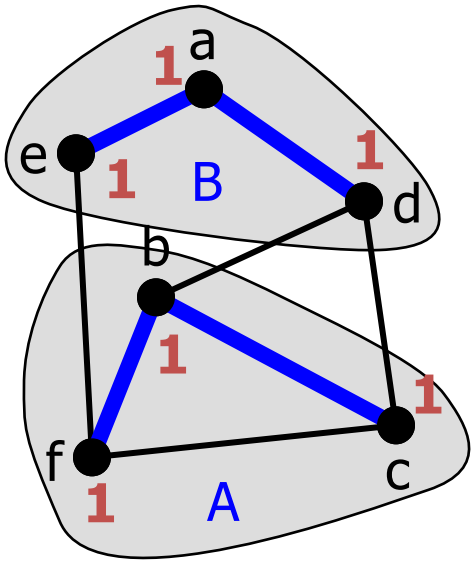
```
for k=1 to 2/ε do
    选择数组中第i个元素xi
    用xi在数组中做二分查找
     if 发现i<j  但是xi>xj then//碰到了“坏”索引
          return false
return true
```

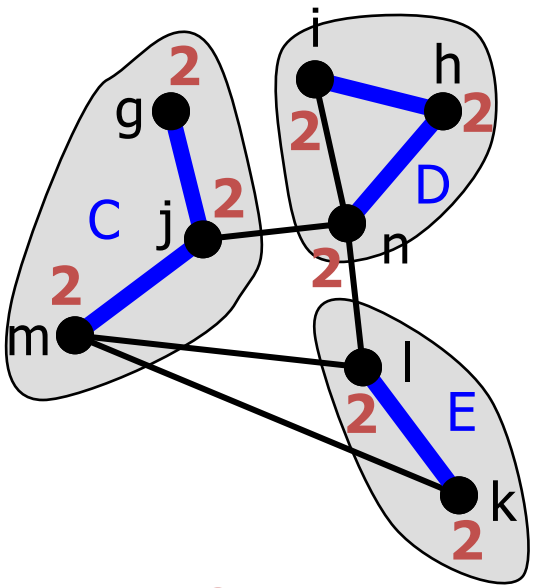
算法的时间复杂性：$O(\frac{1}{\varepsilon}\log n)$

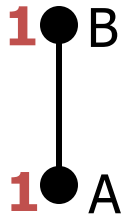
输入数列有序，则总返回True

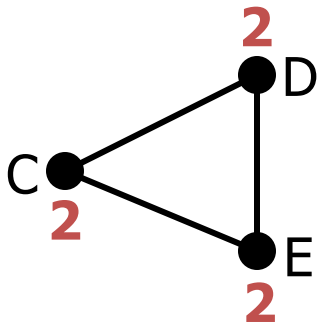
当输入数列ε远离有序时，算法返回false的概率大于2/3

连通性Connectivity——外存算法







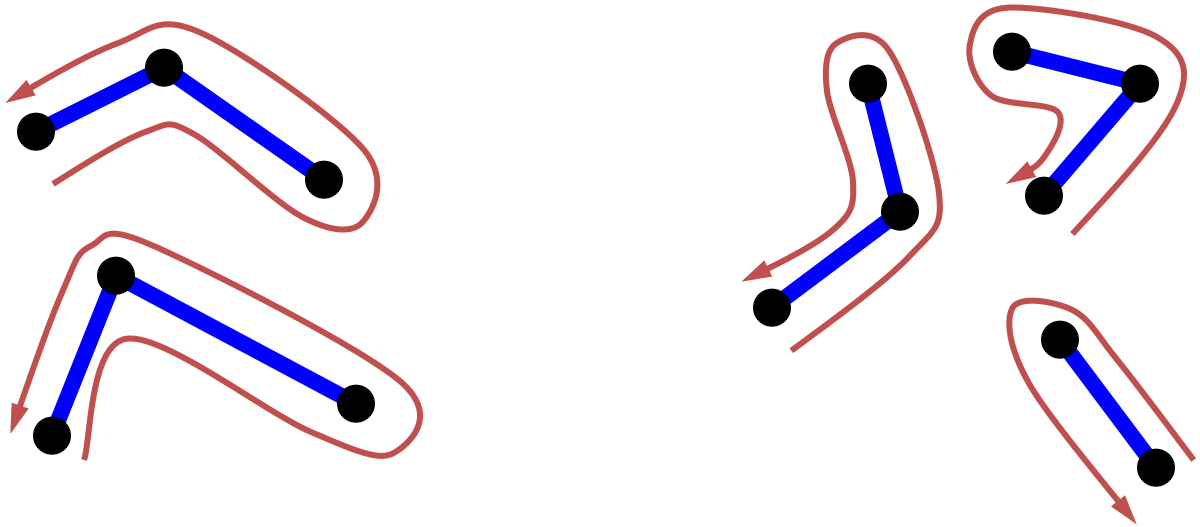


连通性Connectivity

主要步骤:

- 对于每个点找到最小的邻居(容易)
- 计算图H由选定的边导出的连通分量
- 将每个连通分量缩为一个结点(容易)
- 递归调用上述过程
- 对每个$v \in G'$，将其连通度复制到其代表的G中的每个结点上(容易)

连通性

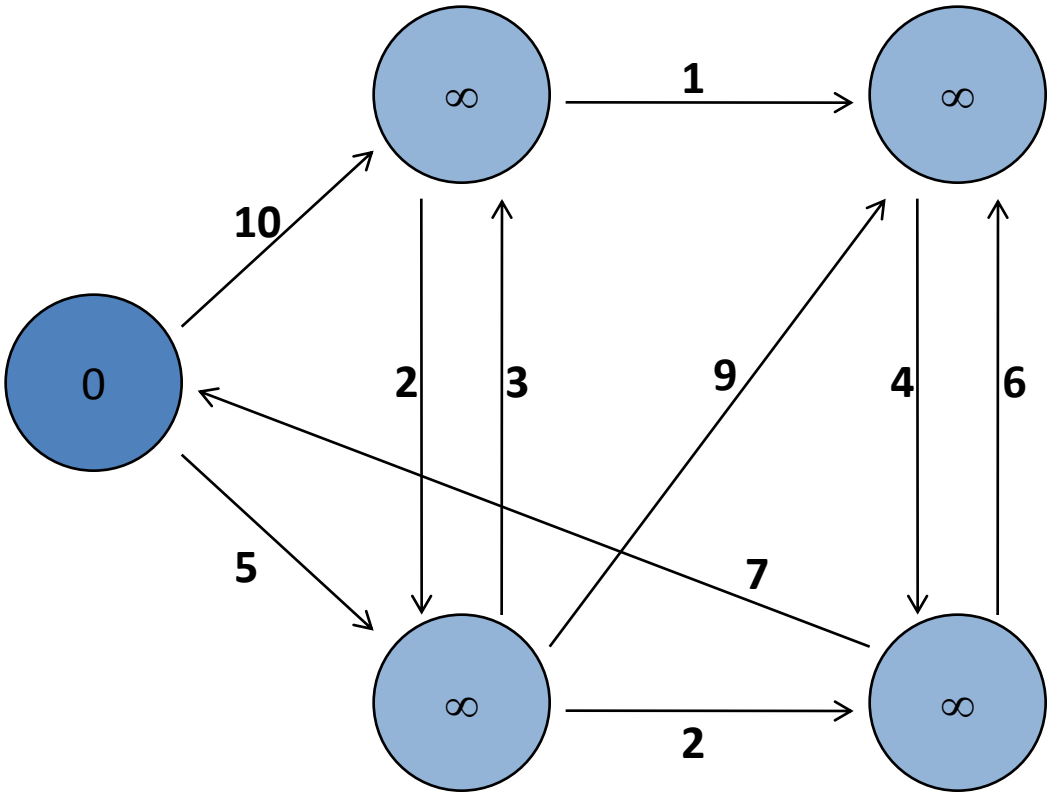


- 每个强连通分量H的大小至少为2
  - $|V'| \le |V|/2$
  - $\log(|V|/M)$次递归调用

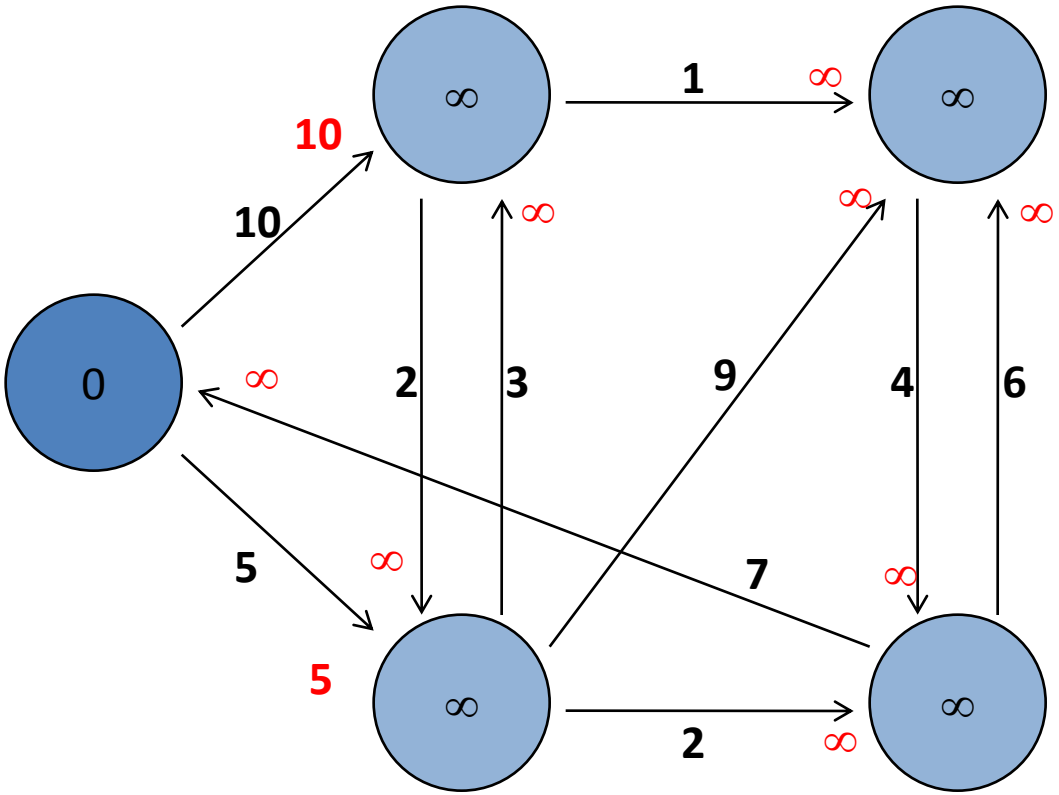
计算图G = (V,E) 的连通分量的I/O复杂度为

$$O(sort(|E|)\log(|V|/M))$$

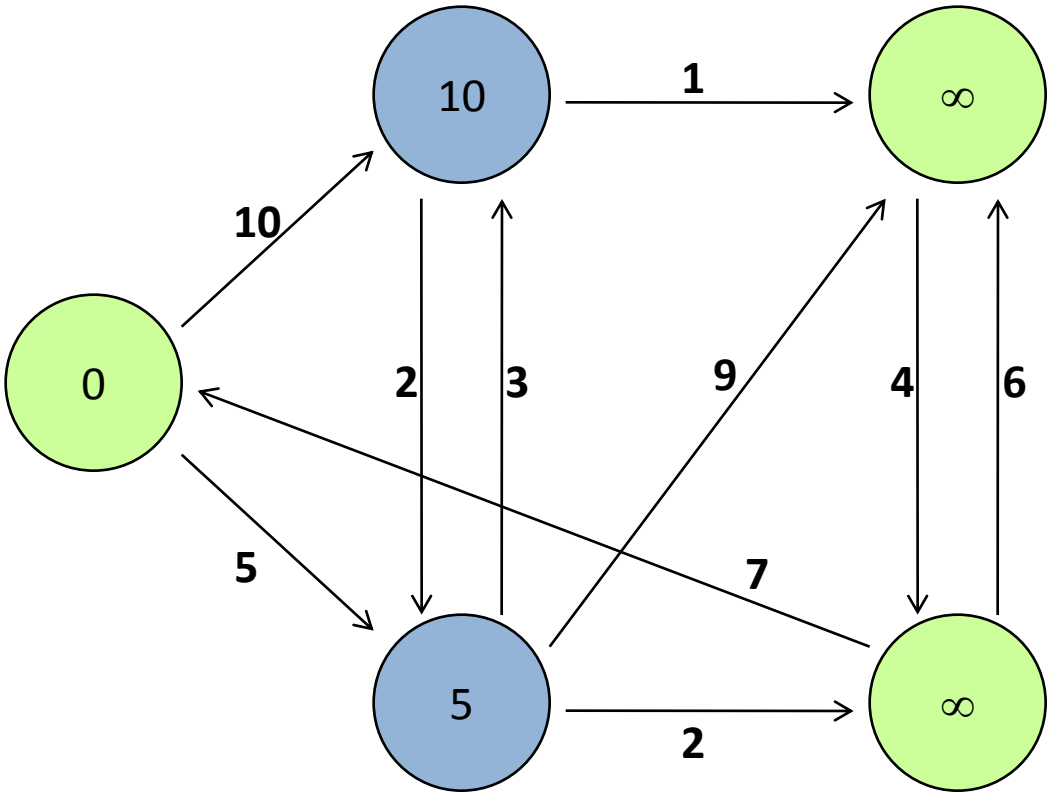
SSSP – Pregel并行广度优先搜索



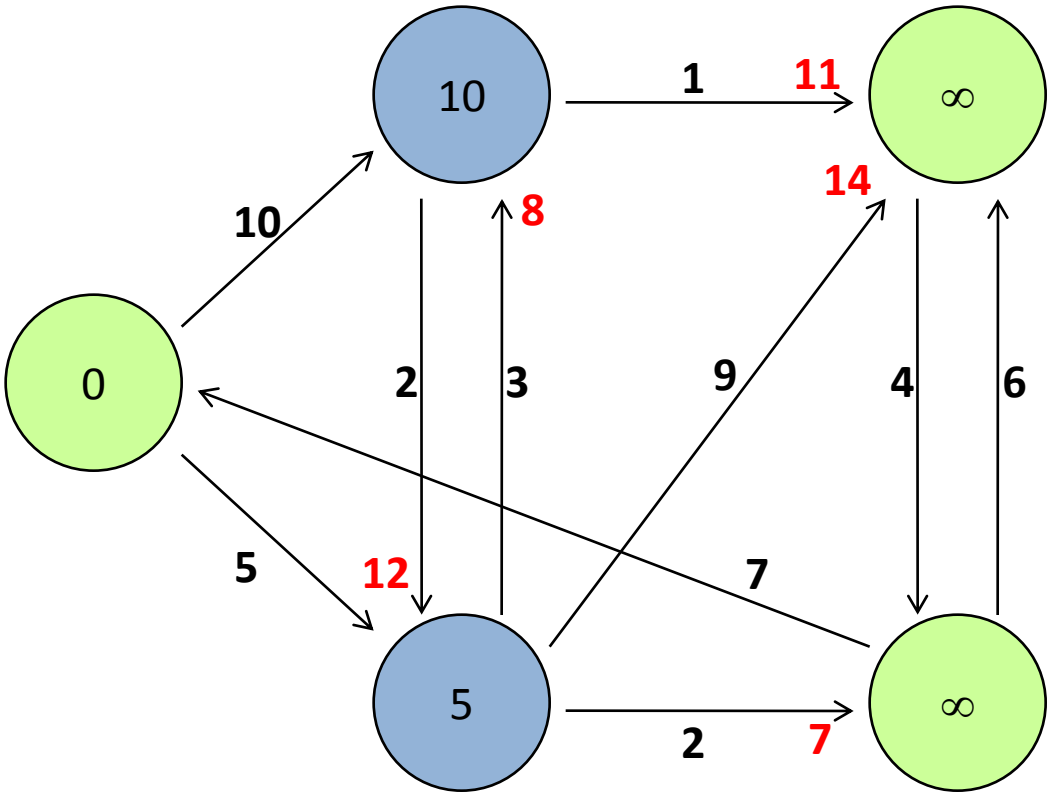
SSSP – Pregel并行广度优先搜索



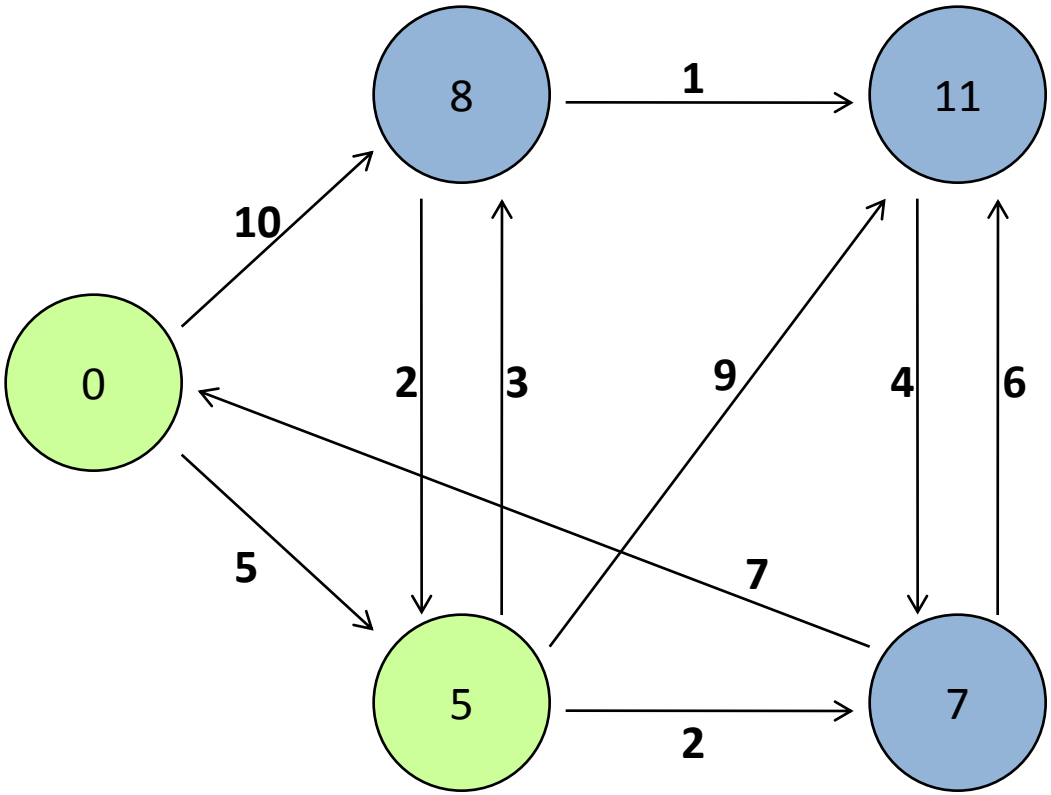
SSSP – Pregel并行广度优先搜索



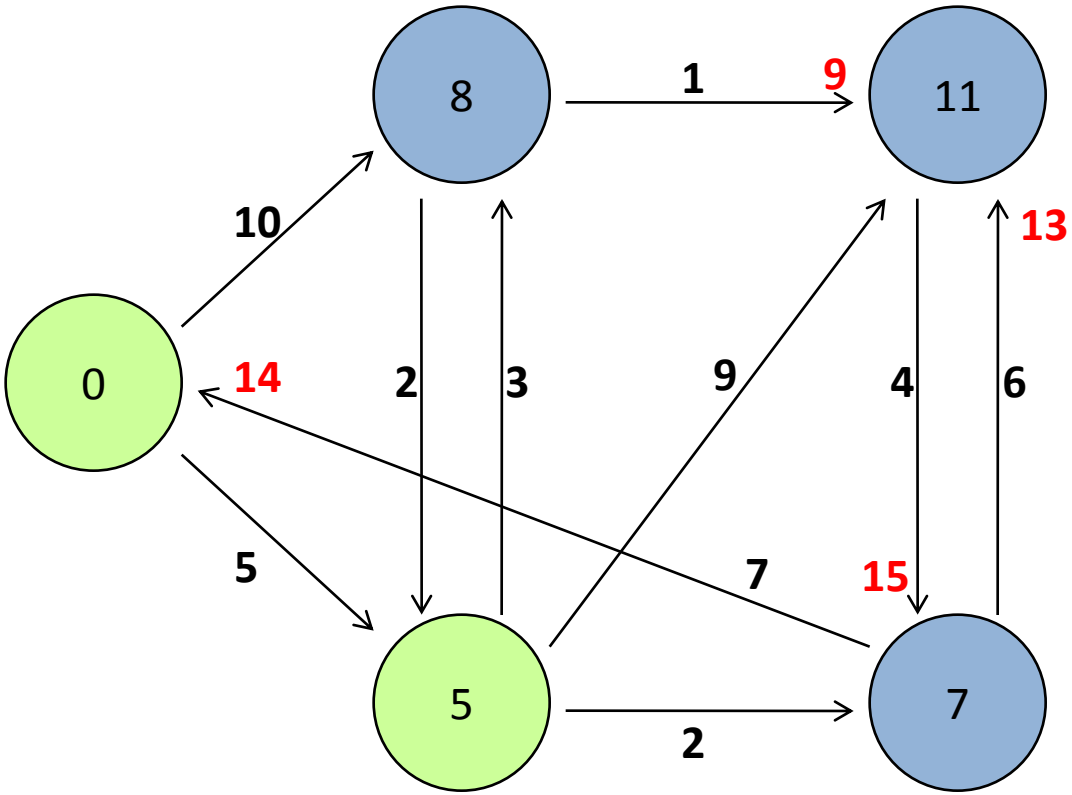
SSSP – Pregel并行广度优先搜索



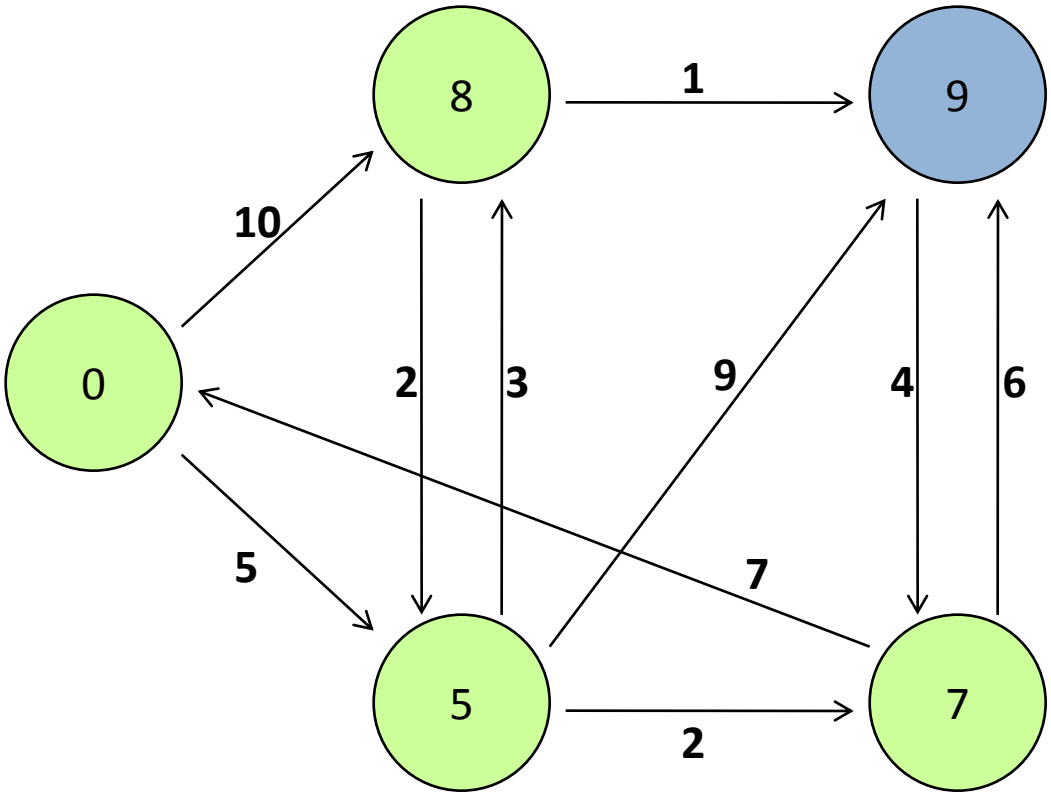
SSSP – Pregel并行广度优先搜索



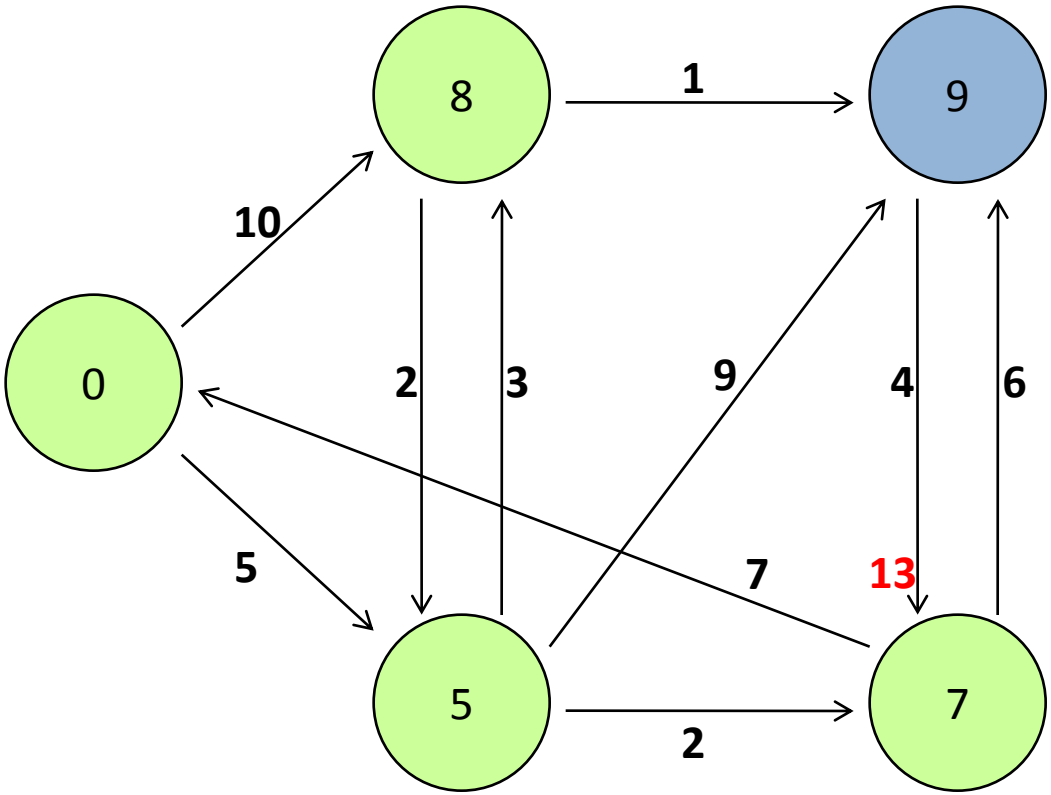
SSSP – Pregel并行广度优先搜索



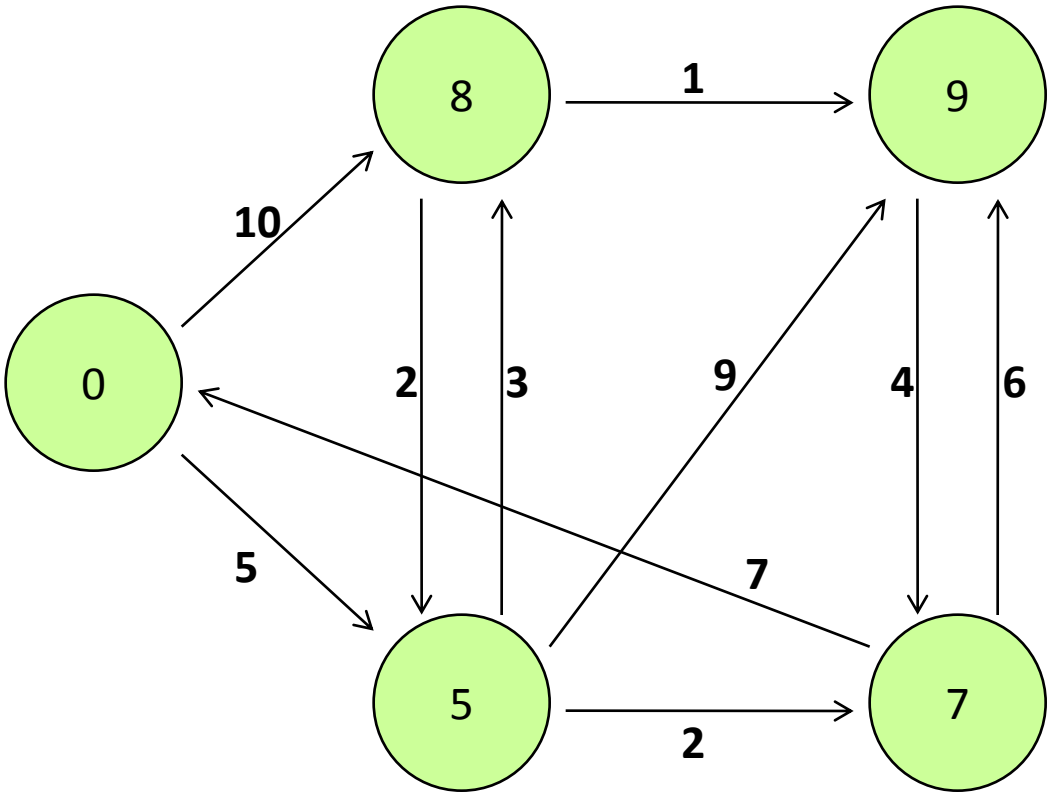
SSSP – Pregel并行广度优先搜索



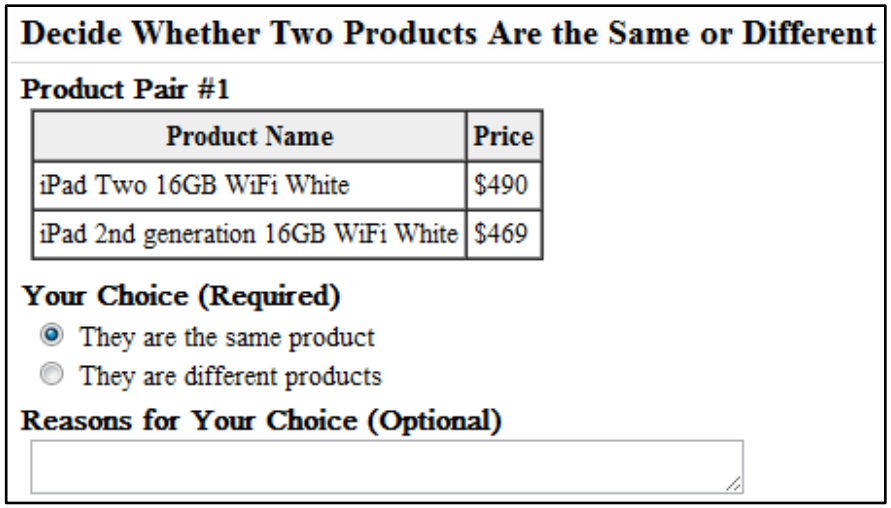
SSSP – Pregel并行广度优先搜索

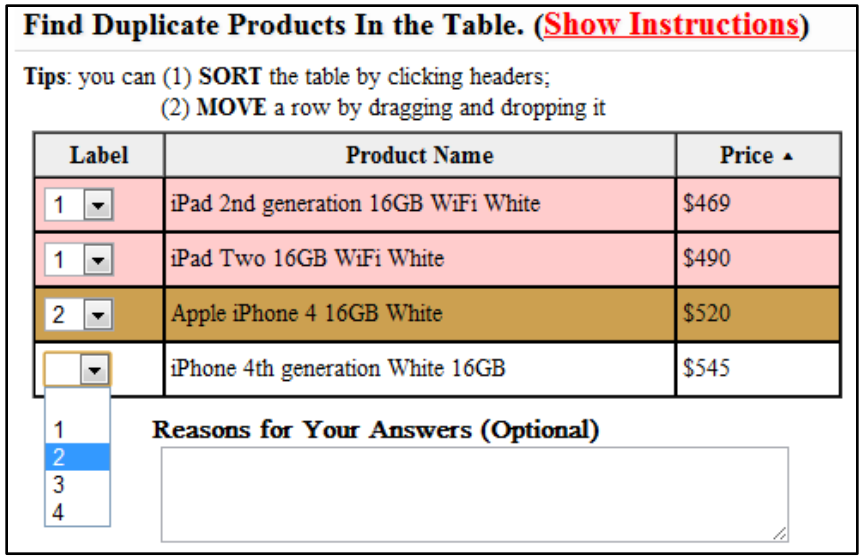


SSSP – Pregel并行广度优先搜索



众包算法：实体识别

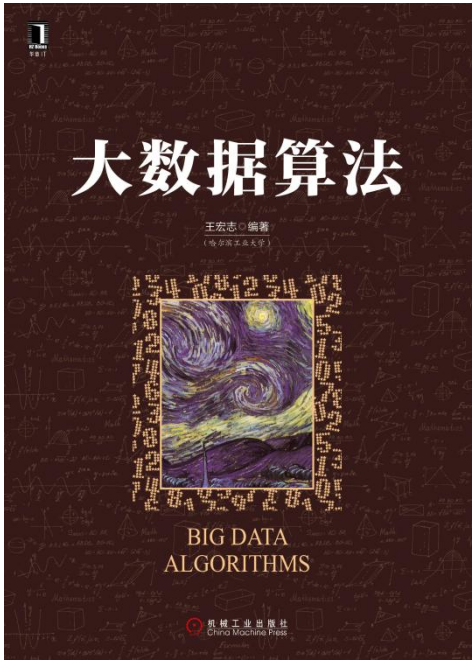

Decide Whether Two Products Are the Same or Different
Product Pair #1
Product Name
Price
iPad Two 16GB WiFi White
$490
iPad 2nd generation 16GB WiFi White
$469
Your Choice (Required)
They are the same product
They are different products
Reasons for Your Choice (Optional)

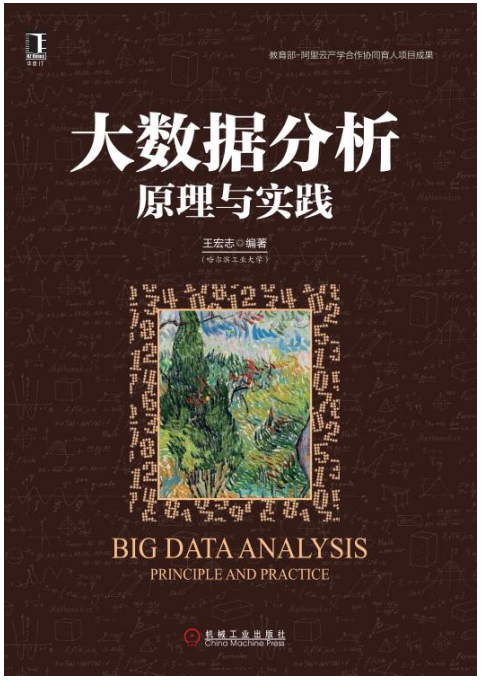

Find Duplicate Products In the Table. (Show Instructions)
Tips: you can (1) SORT the table by clicking headers;
(2) MOVE a row by dragging and dropping it
Label
Product Name
Price
1
iPad 2nd generation 16GB WiFi White
$469
1
iPad Two 16GB WiFi White
$490
2
Apple iPhone 4 16GB White
$520
iPhone 4th generation White 16GB
$545
1
2
3
4
Reasons for Your Answers (Optional)

目录

结论

- 大数据算法在大数据学科起着重要的作用
- 大数据算法和传统的算法设计与分析有显著的不同之处
- 大数据算法的一些设计思路
  - 亚线性算法
  - 外存算法
  - 并行算法
  - 众包算法
- 我们开设大数据算法课程，并且形成一些经验





谢谢！

Thanks for your attention!

报告人：王宏志
wangzh@hit.edu.cn
http://homepage.hit.edu.cn/wang