大数据近似算法

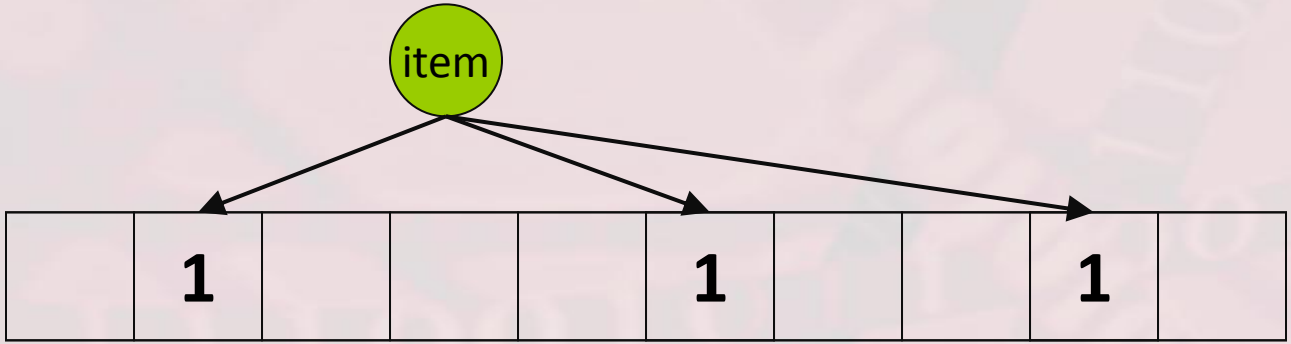
- 研究背景与计算模型
- 随机采样算法
- 基于计数的近似算法
- 基于哈希的近似算法
- 研究成果简介

布隆过滤器

- 1970年由布隆提出
- 针对字典问题(Dictionary problem)：
  - 输入：一个集合S，大小为n
  - 查询：给定一个元素x，问x是否属于S
- 字典问题是很多实际问题的抽象：路由器查找url，数据库查找记录是否存在。。。
- 优点是空间效率和查询时间都远远超过一般的算法，缺点是有一定的误识别率和删除困难

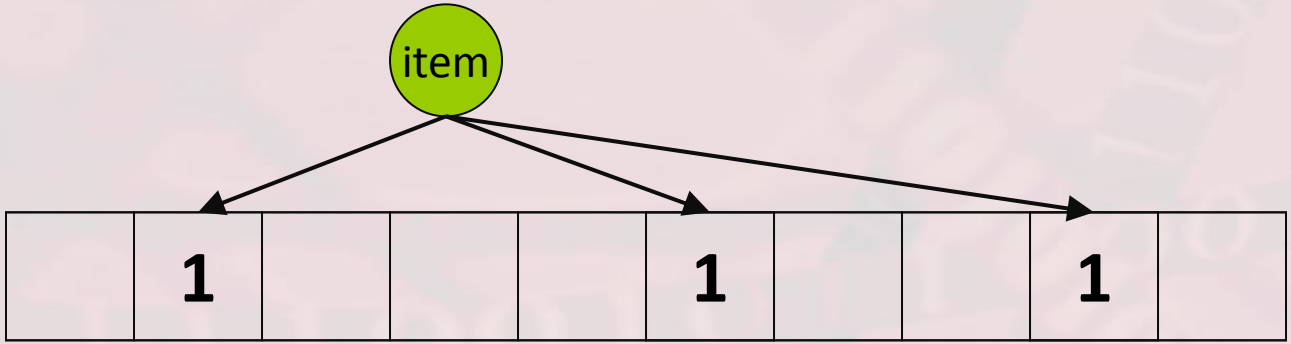
布隆过滤器

- 布隆过滤器(Bloom Filter)的构造
  - 使用一个长度为m的0-1比特数组，以及k个哈希函数$h_1,\ldots,h_k$,
  - 插入元素x：将$h_1(x), \ldots, h_k(x)$设为1
  - 查询元素x是否属于S: 检测$h_1(x), \ldots, h_k(x)$是否都为1
- 性质：若x属于S，则$h_1(x), \ldots, h_k(x)$都为1



布隆过滤器

- 布隆过滤器(Bloom Filter)的性质：
  - 若x属于S，则$h_1(x), \ldots, h_k(x)$已经被设为1，回答正确。布隆过滤器没有false negative；
  - 若x不属于S， $h_1(x), \ldots, h_k(x)$仍然有可能被其他元素设为1，可能出现错误。布隆过滤器有可能出现false positive。



布隆过滤器分析

- 布隆过滤器(Bloom Filter)出现false positive的概率：
  - 某个元素y和某个哈希函数h_j将$h_1(x)$设为1的概率：$\frac{1}{m}$
  - 一共有n个元素，k个哈希函数
  - $h_1(x)$没有被任何元素的哈希值设成1的概率：$\left(1-\frac{1}{m}\right)^{kn}$
  - $h_1(x), \ldots, h_k(x)$中全部被设成1的概率为

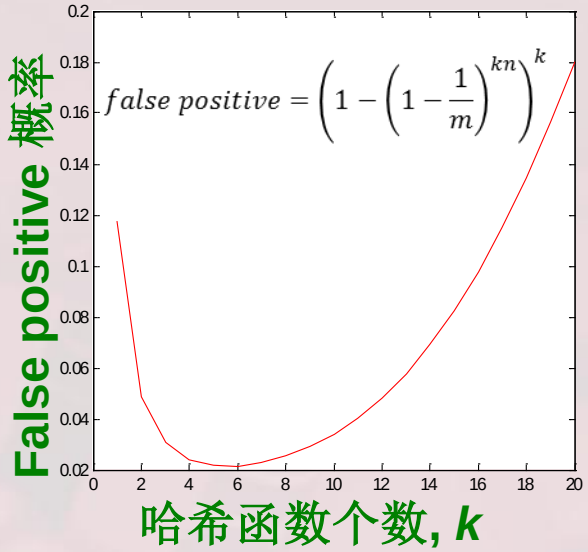
$$\left(1-\left(1-\frac{1}{m}\right)^{kn}\right)^{k}$$

布隆过滤器分析

- ***n* = 1 billion, *m* = 8 billion**
  - **k = 1: False positive = 0.1175**
  - **k = 2: False positive = 0.0493**

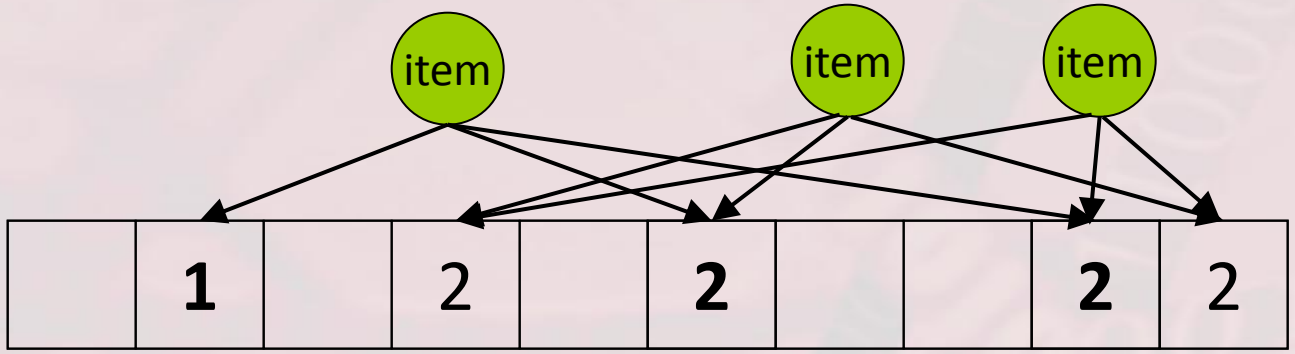
- **当我们增加k时，false positive的概率会先降后升**

- ***K*的最优值: *m*/*n* ln(2)**
  - **上例中最优的k = 8 ln(2) = 5.54 ≈ 6**
    - **false positive概率= $(1 - e^{-1/6})^2 = 0.0235$**



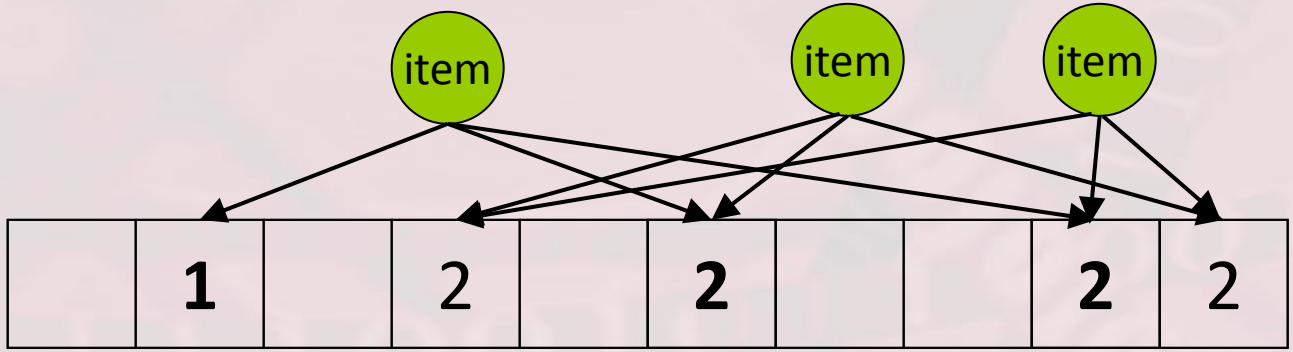
计数布隆过滤器

- 计数布隆过滤器(Counting Bloom filter)，支持删除
  - 将m个比特改为m个计数器
  - 插入元素x：将$h_1(x)$, …, $h_k(x)$对应的计数器加1
  - 删除元素x：将$h_1(x)$, …, $h_k(x)$对应的计数器减1



计数布隆过滤器

- 计数布隆过滤器(Counting Bloom filter)，支持删除
  - 由于哈希函数平均分配元素，每个计数器无需记录太多元素
  - 可证明：在没有重复元素的情况下，每个计数器只需要4比特
- 布隆过滤器仍然是一个活跃的研究领域
  - 在数据库中，通常被称为“Signature file”



大数据近似算法

- 研究背景与计算模型
- 随机采样算法
- 基于计数的摘要算法
- 略图算法
- 研究成果简介

个人简介

魏哲巍

信息学院 副教授 zhewei@ruc.edu.cn

- 个人背景：北大数学院本科毕业，香港科技大学计算机博士，丹麦奥胡斯大学数据科学博士后
- 研究方向：近似数据算法
- 研究成果：于CCF A类会议/期刊发表论文十余篇
- 讲授课程：海量数据算法；算法设计与分析；运筹学
- 个人主页：www.weizhewei.com

研究成果简介

- **Mergeable Summaries**
  - 提出了可合并摘要的概念
  - 设计了随机采样、MG摘要的合并算法
  - 发表于Transaction on Database Systems(TODS), 2014

- **Summary Queries: Theory and Practice**
  - 提出摘要查询，一种基于摘要的新型数据库查询
  - 设计了支持摘要查询的快速查询索引
  - 发表于Transaction on Database Systems(TODS), 2104

研究成果简介

- **Persistent Data Sketching：**
  - 历史查询：从2014年5月到2015年3月，最常被检索的10个关键词是什么？
  - 设计了可支持历史查询的略图算法, 只需使用亚线性空间
  - 发表于ACM SIGMOD International Conference on Management of Data (SIGMOD2015)
- **Matrix Sketching Over Sliding Windows:**
  - 设计了支持窗口查询的矩阵略图算法
  - 发表于ACM SIGMOD International Conference on Management of Data (SIGMOD2016)

谢谢！