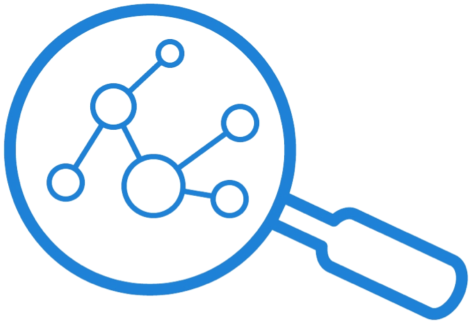
数据科学与大数据技术的数学基础

第四讲



计算机学院
余皓然

2023/5/5

课程内容

Part1 随机化方法
一致性哈希 布隆过滤器 CM Sketch方法 最小哈希 欧氏距离下的相似搜索 Jaccard相似度下的相似搜索
Part2 谱分析方法
主成分分析 奇异值分解 谱图论
Part3 最优化方法
压缩感知

最小哈希

不同元素统计问题

不同元素统计问题

例：

- 零售店收银员希望根据信用卡号统计在一段时间内有多少位（不同的）顾客购买了东西

不同元素统计问题

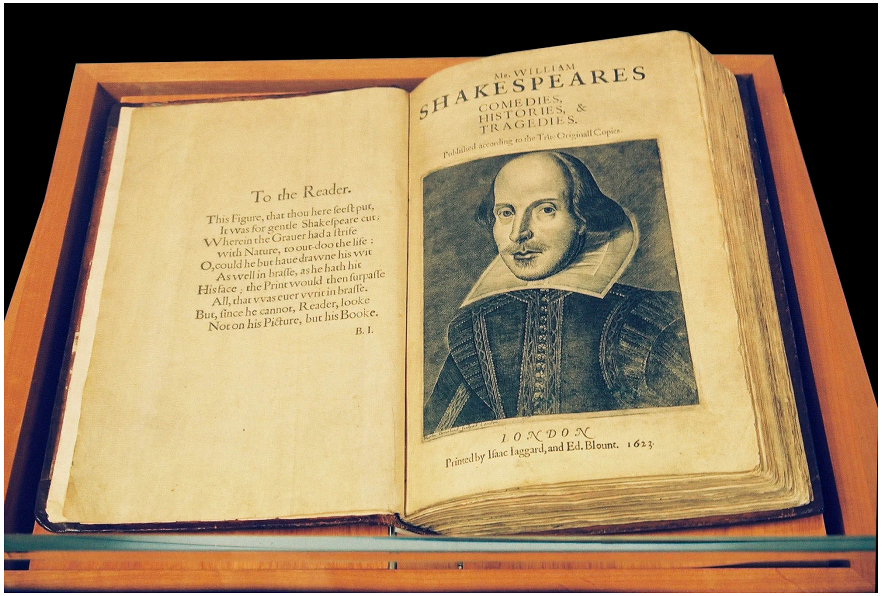
例：

- 零售店收银员希望根据信用卡号统计在一段时间内有多少位（不同的）顾客购买了东西
- 网站/广告商希望根据IP地址统计在一段时间内有多少个（不同的）人访问了网站/广告

不同元素统计问题

例：

- 零售店收银员希望根据信用卡号统计在一段时间内有多少位（不同的）顾客购买了东西
- 网站/广告商希望根据IP地址统计在一段时间内有多少个（不同的）人访问了网站/广告
- 统计莎士比亚在他的各类作品中共用了多少个（不同的）单词



共31534个词

不同元素统计问题

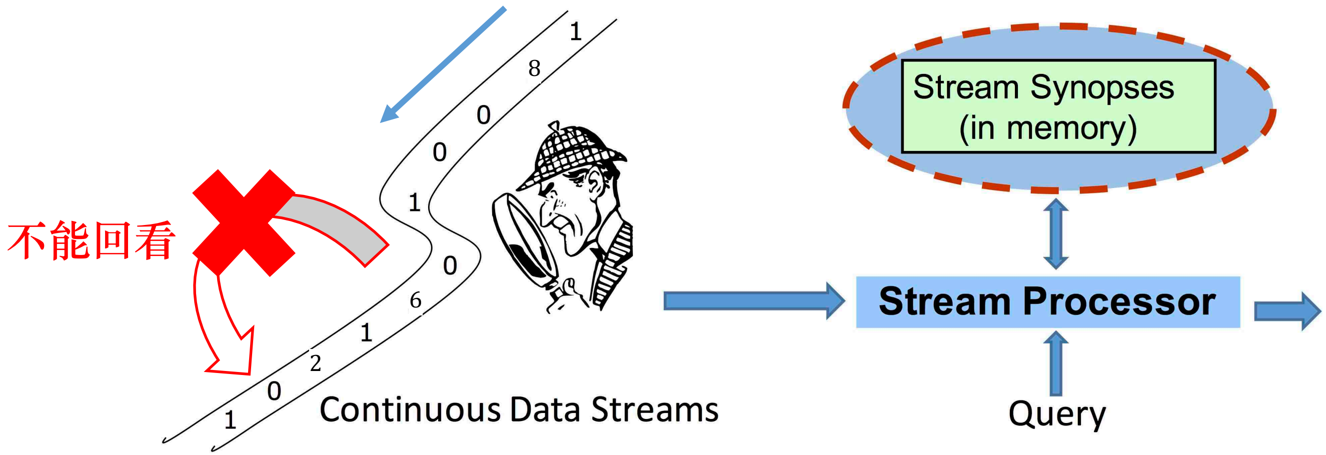
不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的数据流$x_1, \cdots, x_n$，如何计算其中不同元素的个数?



图片取自S. Raskhodnikova and A. Smith <Scalable, Private Algorithms for Continual Data Analysis >

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）
给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假设$x_1, \cdots, x_n \in \{1,2,\dots,m\}$，最直接的方法是：

- 方法一：用一个长度为m的0-1向量记录（消耗m比特用于存储，即空间复杂度为$O(m)$）

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假设$x_1, \cdots, x_n \in \{1,2,\dots,m\}$，最直接的方法是:

- 方法一：用一个长度为m的0-1向量记录（消耗m比特用于存储，即空间复杂度为$O(m)$）
- 方法二：对输入的每个首次出现的x，用$\log_2 m$比特记录（最多消耗$n\log_2 m$比特用于存储，即空间复杂度为$O(n\log_2 m)$）

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假设$x_1, \cdots, x_n \in \{1,2,\ldots,m\}$，最直接的方法是：

- 方法一：用一个长度为m的0-1向量记录（消耗m比特用于存储，即空间复杂度为$O(m)$）
- 方法二：对输入的每个首次出现的x，用$\log_2 m$比特记录（最多消耗$n\log_2 m$比特用于存储，即空间复杂度为$O(n\log_2 m)$）

m和n值在实际应用中都非常大，是否有方法可以降低空间复杂度?

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

m和n值在实际应用中都非常大，是否有方法可以降低空间复杂度?

不可能定理（Impossibility Result）

当$n = O(m)$，在仅浏览数据一遍的约束下，**不存在**任何确定性算法可以用**少于m比特**的存储空间准确计算不同元素的个数。

定理引自A. Blum, J. Hopcroft, and R. Kannan <Foundations of Data Science> 第183面

不同元素统计问题

不可能定理（Impossibility Result）

当$n = O(m)$，在仅浏览数据一遍的约束下，**不存在**任何确定性算法可以用**少于m比特**的存储空间准确计算不同元素的个数。

Lower bound on memory for exact deterministic algorithm

We show that any exact deterministic algorithm must use at least m bits of memory on some sequence of length $m+1$. Suppose we have seen $a_1, \ldots, a_m$, and suppose for sake of contradiction that our algorithm uses less than m bits of memory on all such sequences. There are $2^m - 1$ possible subsets of $\{1, 2, \ldots, m\}$ that the sequence could contain and yet only 2^{m-1} possible states of our algorithm's memory. Therefore there must be two different subsets S_1 and S_2 that lead to the same memory state. If S_1 and S_2 are of different sizes, then clearly this implies an error for one of the input sequences. On the other hand, if they are the same size, then if the next element is in S_1 but not S_2, the algorithm will give the same answer in both cases and therefore must give an incorrect answer on at least one of them.

假设$m = 3$，则对于出现的不同元素共有7个可能情况：{1,2,3}, {1,2}, {1,3}, {2,3}, {1}, {2}, {3}

证明思路引自A. Blum, J. Hopcroft, and R. Kannan <Foundations of Data Science> 第183面

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）
给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

改用**随机化**方法**近似**估计不同元素的个数，以**精确度**换**存储空间**

不同元素统计问题

不同元素统计问题（Distinct Element Counting Problem）
给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

改用**随机化**方法**近似**估计不同元素的个数，以**精确度**换**存储空间**

（与解决从属判断问题、高频元素寻找问题的思路类似）

从属判断问题（membership query）	近似求解的随机化方法：布隆过滤器
如何存储关于集合S的信息从而可以准确判断关于“元素x是否属于集S”的问题?	

高频元素寻找问题（Heavy Hitters Problem）	近似求解的随机化方法：CM Sketch
给定长度为n的数组A和数值k，如何寻找出所有出现次数大于等于n/k的元素?	

最小哈希

最小哈希方法

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假如有一个随机哈希函数 $h\colon \mathcal{U} \to [0,1]$，即输出是连续而非离散值

初始化 $s \leftarrow 1$

对$i = 1, \ldots, n$:　$s \leftarrow \min\{s, h(x_i)\}$

最后如何估计不同元素的个数?

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的数据流$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假如有一个随机哈希函数 $h\colon \mathcal{U} \to [0,1]$，即输出是连续而非离散值

初始化 $s \leftarrow 1$

对$i = 1, \ldots, n$:　$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假如有一个随机哈希函数 $h\colon \mathcal{U} \to [0,1]$，即输出是连续而非离散值

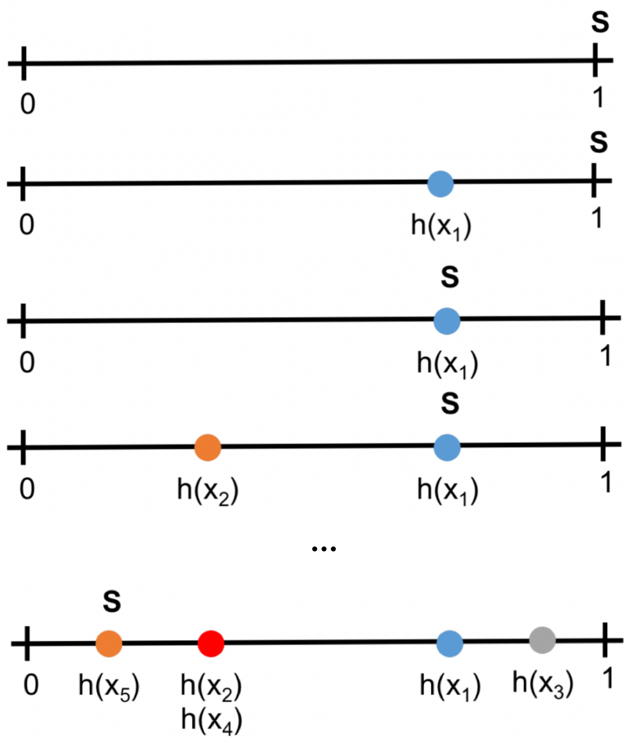
初始化 $s \leftarrow 1$

对$i = 1, \ldots, n$:　$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

若是相同元素，则哈希值相等

例如，若$x_2 = x_4$，有$h(x_2) = h(x_4)$



图例取自Cameron Musco <COMPSCI 514 - Algorithms for Data Science>

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

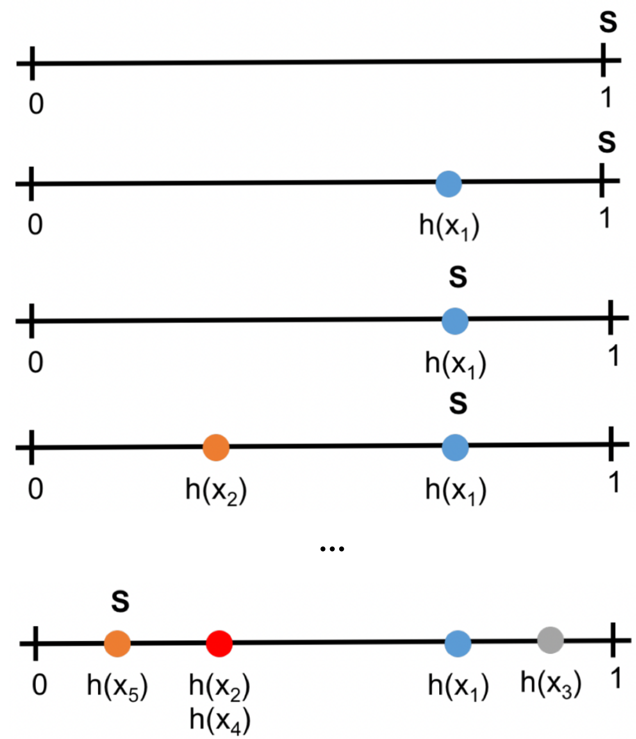
给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

假如有一个随机哈希函数 $h\colon \mathcal{U} \rightarrow [0,1]$，即输出是连续而非离散值

初始化 $s \leftarrow 1$

对$i = 1, \ldots, n$:　$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计



s等于d个（注意不是n个）取值服从在$[0,1]$均匀分布的随机变量的最小值，如何分析随机变量s与d的关系?　（* d为不同元素个数）

d越大，s如何变化?

图例取自Cameron Musco <COMPSCI 514 - Algorithms for Data Science>

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

s等于d个取值服从在$[0,1]$均匀分布的随机变量的最小值，如何分析随机变量s与d的关系?

最小哈希

不同元素统计问题（Distinct Element Counting Problem）
给定长度为n的数据流$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

s等于d个取值服从在$[0,1]$均匀分布的随机变量的最小值，如何分析随机变量s与d的关系?

提示：分析s的累积分布函数 $F(\tilde{s}) = \Pr[s \leq \tilde{s}] = 1 - (1 - \tilde{s})^d$

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的数据流$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

s等于d个取值服从在$[0,1]$均匀分布的随机变量的最小值，如何分析随机变量s与d的关系?

对于$s \in [0,1]$，有$F(s) = 1 - (1 - s)^d$

$$\mathbb{E}\{s\} = \int_{s=0}^{1} s\, dF(s) = 1 - \int_{s=0}^{1} F(s)\, ds = 1 + \frac{1}{d+1} - 1 = \frac{1}{d+1}$$

因此有$d = \frac{1}{\mathbb{E}\{s\}} - 1$

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1,\cdots,x_n$，如何计算其中不同元素的个数?

s等于d个取值服从在$[0,1]$均匀分布的随机变量的最小值，如何分析随机变量s与d的关系?

对于$s\in[0,1]$，有$F(s)=1-(1-s)^d$

$$\mathbb{E}\{s\}=\int_{s=0}^{1} s\, dF(s)=1-\int_{s=0}^{1}F(s)\, ds=1+\frac{1}{d+1}-1=\frac{1}{d+1}$$

因此有$d=\frac{1}{\mathbb{E}\{s\}}-1$

初始化 $s\leftarrow 1$

对$i=1,\dots,n$： $s\leftarrow\min\{s,h(x_i)\}$

取$\frac{1}{s}-1$作为对不同元素个数的估计$\hat{d}$

注意$d\neq\mathbb{E}[\hat{d}]$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

上一讲用于分析CM Sketch方法的马尔可夫不等式能否用于最小哈希的分析?

马尔可夫不等式（Markov's Inequality）

若X是一个非负随机变量且$\mathbb{E}[X] \neq 0$，对任何常数$c > 0$，有 $\Pr\left[X \geq c\mathbb{E}[X]\right] \leq \frac{1}{c}$.

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1 - s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

上一讲用于分析CM Sketch方法的**马尔可夫不等式**能否用于最小哈希的分析?

马尔可夫不等式（Markov's Inequality）

若X是一个非负随机变量且$\mathbb{E}[X] \neq 0$，对任何常数$c > 0$，有 $\Pr\left[X \geq c\mathbb{E}[X]\right] \leq \frac{1}{c}$.

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

可以同时刻画$X \geq \mathbb{E}[X]$时和$X \leq \mathbb{E}[X]$时X偏离$\mathbb{E}[X]$的概率

最小哈希

马尔可夫不等式（Markov's Inequality）

若X是一个非负随机变量且$\mathbb{E}[X] \neq 0$，对任何常数$c > 0$，有 $\Pr\left[X \geq c\mathbb{E}[X]\right] \leq \frac{1}{c}$.

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

可以利用马尔可夫不等式证明切比雪夫不等式:

最小哈希

马尔可夫不等式（Markov's Inequality）

若X是一个非负随机变量且$\mathbb{E}[X] \neq 0$，对任何常数$c > 0$，有 $\Pr\left[X \geq c\mathbb{E}[X]\right] \leq \frac{1}{c}$.

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

可以利用马尔可夫不等式证明切比雪夫不等式:

$$\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right]=\Pr\left[(X - \mathbb{E}[X])^2 \geq c^2\mathbf{Var}[X]\right]$$

将$(X - \mathbb{E}[X])^2$视作随机变量（满足非负条件），用马尔可夫不等式及$\mathbf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

需要先算$\mathbf{Var}[s]$

$$\mathbb{E}\{s^2\} = \int_{s=0}^{1} s^2\, dF(s) = 1 - 2\int_{s=0}^{1} sF(s)\, ds = 1 - 2\int_{s=0}^{1} s\, d\left(s + \frac{(1-s)^{d+1}}{d+1}\right)$$

分部积分

$$= 1 - 2\left(\frac{1}{2} - \frac{1}{d+1}\frac{1}{d+2}\right) = \frac{2}{(d+1)(d+2)}$$

分部积分

$$\mathbf{Var}[s] = \mathbb{E}\{s^2\} - (\mathbb{E}\{s\})^2 = \frac{d}{(d+1)^2(d+2)}$$

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

如何分析$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right]$?

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

分析$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right]$

$$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right] = \Pr\left[|s - \mathbb{E}[s]| \geq \frac{\varepsilon\mathbb{E}[s]}{\sqrt{\mathbf{Var}[s]}}\sqrt{\mathbf{Var}[s]}\right] \leq \frac{\mathbf{Var}[s]}{\varepsilon^2(\mathbb{E}[s])^2} = \frac{d}{\varepsilon^2(d+2)} \leq \frac{1}{\varepsilon^2}$$

最小哈希

对于$s \in [0,1]$，有$F(s) = 1-(1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev’s Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

结论为：$\Pr[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]] \leq \frac{1}{\varepsilon^2}$

初始化 $s \leftarrow 1$

对$i = 1, \dots, n$：　$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

当$\varepsilon < 1$时，不等式右边大于1，不等式没有价值

如何改进方法，使不等式右边取值更小?

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\textbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\textbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\textbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

结论为：$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon \mathbb{E}[s]\right] \leq \frac{1}{\varepsilon^2}$

初始化 $s \leftarrow 1$

对$i = 1, \ldots, n$： $s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

当$\varepsilon < 1$时，不等式右边大于1，不等式没有价值

如何改进方法，使不等式右边取值更小?

使用多个哈希函数，令随机变量取值更接近期望

最小哈希

对于$s \in [0,1]$，有$F(s) = 1 - (1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

结论为：$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right] \leq \frac{1}{\varepsilon^2}$

使用多个哈希函数，令随机变量取值更接近期望

初始化 $s \leftarrow 1$

对$i = 1, \dots, n$：　$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$：$s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

取　　作为对不同元素个数的估计

最小哈希

对于$s \in [0,1]$，有$F(s) = 1-(1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

结论为：$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right] \leq \frac{1}{\varepsilon^2}$

使用多个哈希函数，令随机变量取值更接近期望

初始化 $s \leftarrow 1$

对$i = 1, \dots, n$： $s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

取____作为对不同元素个数的估计

用$\frac{1}{\frac{s_1+s_2+\cdots+s_k}{k}} - 1$还是$\frac{1}{k}\left(\frac{1}{s_1} - 1 + \frac{1}{s_2} - 1 + \cdots + \frac{1}{s_k} - 1\right)$?

最小哈希

对于$s \in [0,1]$，有$F(s) = 1-(1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

结论为：$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right] \leq \frac{1}{\varepsilon^2}$

使用多个哈希函数，令随机变量取值更接近期望

初始化 $s \leftarrow 1$

对$i = 1, \dots, n$：$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

注意$d = \frac{1}{\mathbb{E}\{s\}} - 1 \neq \mathbb{E}\left\{\frac{1}{s} - 1\right\}$

例有某随机变量$s \sim \mathcal{U}[0,1]$

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$：$s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

取　　作为对不同元素个数的估计

用$\frac{1}{\frac{s_1+s_2+\cdots+s_k}{k}} - 1$还是$\frac{1}{k}\left(\frac{1}{s_1} - 1 + \frac{1}{s_2} - 1 + \cdots + \frac{1}{s_k} - 1\right)$?

最小哈希

对于$s\in[0,1]$，有$F(s)=1-(1-s)^d$，$\mathbb{E}\{s\}=\frac{1}{d+1}$，$\mathbf{Var}[s]=\frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X]\neq 0$，对任何$c>0$，有 $\Pr\left[|X-\mathbb{E}[X]|\geq c\sqrt{\mathbf{Var}[X]}\right]\leq\frac{1}{c^2}$.

结论为：$\Pr\left[|s-\mathbb{E}[s]|\geq\varepsilon\mathbb{E}[s]\right]\leq\frac{1}{\varepsilon^2}$

使用多个哈希函数，令随机变量取值更接近期望

初始化 $s\leftarrow 1$

对$i=1,\ldots,n$： $s\leftarrow\min\{s,h(x_i)\}$

取$\frac{1}{s}-1$作为对不同元素个数的估计

注意$d=\frac{1}{\mathbb{E}\{s\}}-1\neq\mathbb{E}\left\{\frac{1}{s}-1\right\}$

$\mathbb{E}\left\{\frac{1}{s}-1\right\}$趋近于$\infty$

初始化 $s_1,s_2,\ldots,s_k\leftarrow 1$

对$i=1,\ldots,n$： $s_1\leftarrow\min\{s_1,h_1(x_i)\},\ldots,s_k\leftarrow\min\{s_k,h_k(x_i)\}$

取　　作为对不同元素个数的估计

用$\frac{1}{\frac{s_1+s_2+\cdots+s_k}{k}}-1$还是$\frac{1}{k}\left(\frac{1}{s_1}-1+\frac{1}{s_2}-1+\cdots+\frac{1}{s_k}-1\right)$?

最小哈希

对于$s \in [0,1]$，有$F(s) = 1-(1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

结论为：$\Pr\left[|s - \mathbb{E}[s]| \geq \varepsilon\mathbb{E}[s]\right] \leq \frac{1}{\varepsilon^2}$

使用多个哈希函数，令随机变量取值更接近期望

初始化 $s \leftarrow 1$

对$i = 1, \dots, n$：　$s \leftarrow \min\{s, h(x_i)\}$

取$\frac{1}{s} - 1$作为对不同元素个数的估计

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$：$s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1 + s_2 + \cdots + s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

最小哈希

对于$s \in [0,1]$，有$F(s) = 1-(1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c>0$，有 $\Pr\left[|X-\mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}}-1$作为对不同元素个数的估计

对随机变量$\bar{s}$如何求期望和方差?

最小哈希

对于$s \in [0,1]$，有$F(s) = 1-(1-s)^d$，$\mathbb{E}\{s\} = \frac{1}{d+1}$，$\mathbf{Var}[s] = \frac{d}{(d+1)^2(d+2)}$

随机变量s在其期望$\mathbb{E}\{s\}$附近取值的概率?

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

对随机变量$\bar{s}$有：

期望不变

方差缩小至$\frac{1}{k}$

$\mathbb{E}\{\bar{s}\} = \frac{\mathbb{E}\{s_1\}+\cdots+\mathbb{E}\{s_k\}}{k} = \frac{1}{d+1}$（期望的线性性质），$\mathbf{Var}[\bar{s}] = \frac{1}{k}\frac{d}{(d+1)^2(d+2)}$（$s_1, s_2, \dots, s_k$为独立变量）

If X and Y are independent, then: $E(XY) = E(X)E(Y)$.

利用期望定义及独立事件概率关系

最小哈希

切比雪夫不等式（Chebyshev's Inequality）

若X是随机变量且$\mathbf{Var}[X] \neq 0$，对任何$c > 0$，有 $\Pr\left[|X - \mathbb{E}[X]| \geq c\sqrt{\mathbf{Var}[X]}\right] \leq \frac{1}{c^2}$.

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

$$\mathbb{E}\{\bar{s}\} = \frac{\mathbb{E}\{s_1\}+\cdots+\mathbb{E}\{s_k\}}{k} = \frac{1}{d+1}, \quad \mathbf{Var}[\bar{s}] = \frac{1}{k}\frac{d}{(d+1)^2(d+2)}$$

$$\Pr[|\bar{s} - \mathbb{E}[\bar{s}]| \geq \varepsilon\mathbb{E}[\bar{s}]] = \Pr\left[|\bar{s} - \mathbb{E}[\bar{s}]| \geq \frac{\varepsilon\mathbb{E}[\bar{s}]}{\sqrt{\mathbf{Var}[\bar{s}]}}\sqrt{\mathbf{Var}[\bar{s}]}\right] \leq \frac{\mathbf{Var}[\bar{s}]}{\varepsilon^2(\mathbb{E}[\bar{s}])^2} \leq \frac{1}{k\varepsilon^2}$$

因为考虑了k个哈希函数，即便$\varepsilon < 1$，也可以通过增大k使不等式右边小于1

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

$$\Pr\left[|\bar{s} - \mathbb{E}[\bar{s}]| \geq \varepsilon\mathbb{E}[\bar{s}]\right] \leq \frac{1}{k\varepsilon^2}$$

$\mathbb{E}\{\bar{s}\}$与实际d的关系 $\mathbb{E}\{\bar{s}\} = \dfrac{1}{d+1} \Rightarrow d = \dfrac{1}{\mathbb{E}\{\bar{s}\}} - 1$

$\bar{s}$与估计值$\hat{d}$的关系 $\hat{d} = \dfrac{1}{\bar{s}} - 1$

回顾：$\mathbb{E}\left[\hat{d}\right] \neq d$，那么如何刻画$d$相对$\mathbb{E}\left[\hat{d}\right]$的偏差

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

$$\Pr\left[|\bar{s} - \mathbb{E}[\bar{s}]| \geq \varepsilon \mathbb{E}[\bar{s}]\right] \leq \frac{1}{k\varepsilon^2}$$

$\mathbb{E}\{\bar{s}\}$与实际d的关系 $\mathbb{E}\{\bar{s}\} = \frac{1}{d+1} \Rightarrow d = \frac{1}{\mathbb{E}\{\bar{s}\}} - 1$ $\bar{s}$与估计值$\hat{d}$的关系 $\hat{d} = \frac{1}{\bar{s}} - 1$

回顾：$\mathbb{E}[\hat{d}] \neq d$，那么如何刻画$d$相对$\mathbb{E}[\hat{d}]$的偏差

先求$\bar{s}$与d的关系：

$$\Pr\left[\left|\bar{s} - \frac{1}{d+1}\right| \geq \frac{\varepsilon}{d+1}\right] \leq \frac{1}{k\varepsilon^2} \Rightarrow \Pr\left[\bar{s} \geq \frac{\varepsilon+1}{d+1} \; or \; \bar{s} \leq \frac{1-\varepsilon}{d+1}\right] \leq \frac{1}{k\varepsilon^2}$$

$$\Rightarrow \Pr\left[\hat{d} \leq \frac{d-\varepsilon}{1+\varepsilon} \; or \; \hat{d} \geq \frac{d+\varepsilon}{1-\varepsilon}\right] \leq \frac{1}{k\varepsilon^2}$$

为得到简洁表达式，利用$\varepsilon \leq 0.5$（一般关心较小的ε）和d取值一般较大进行缩放

$$\Pr\left[\hat{d} \leq (1-3\varepsilon)d \; or \; \hat{d} \geq (1+3\varepsilon)d\right] \leq \frac{1}{k\varepsilon^2}$$

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

初始化 $s_1, s_2, \ldots, s_k \leftarrow 1$

对$i = 1, \ldots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \ldots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

$$\Pr\left[\left|\hat{d} - d\right| \geq 3\varepsilon d\right] \leq \frac{1}{k\varepsilon^2}$$

（当$\varepsilon \leq 0.5$以及d取值较大时成立）

例，若要保证估计值$\hat{d}$距离真实值d的偏差大于等于$3\varepsilon d$的概率小于等于δ，该如何设置k?

最小哈希

不同元素统计问题（Distinct Element Counting Problem）

给定长度为n的**数据流**$x_1, \cdots, x_n$，如何计算其中不同元素的个数?

初始化 $s_1, s_2, \ldots, s_k \leftarrow 1$

对$i = 1, \ldots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \ldots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

$$\Pr\left[\left|\hat{d} - d\right| \geq 3\varepsilon d\right] \leq \frac{1}{k\varepsilon^2}$$

（当$\varepsilon \leq 0.5$以及d取值较大时成立）

例，若要保证估计值$\hat{d}$距离真实值d的偏差大于等于$3\varepsilon d$的概率小于等于δ，该如何设置k?

令$\frac{1}{k\varepsilon^2} = \delta$，有$k = \frac{1}{\delta\varepsilon^2}$ 不受n和d的影响

最小哈希方法的空间复杂度?

最小哈希

最小哈希的应用

多个哈希函数

初始化 $s_1, s_2, \ldots, s_k \leftarrow 1$

对$i = 1, \ldots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \ldots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

在实际应用中，为每个元素分别计算k个哈希值非常耗时，也不容易构建k个独立的哈希函数

有无其它替代方法?

相关评论参见P. Flajolet et al., <HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm>

多个哈希函数

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

在实际应用中，为每个元素分别计算k个哈希值非常耗时，也不容易构建k个独立的哈希函数

有无其它替代方法?

依然只用一个哈希函数，但把区间[0,1]均匀划成k段，为每个区间分别计算（标准化后的）最小哈希值，再对k个最小哈希值求 ?

方法提出于P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

多个哈希函数

初始化 $s_1, s_2, \ldots, s_k \leftarrow 1$

对$i = 1, \ldots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \ldots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

在实际应用中，为每个元素分别计算k个哈希值非常耗时，也不容易构建k个独立的哈希函数

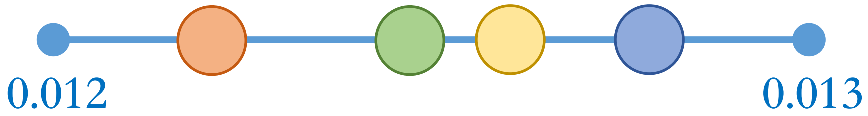
有无其它替代方法?

依然只用一个哈希函数，但把区间[0,1]均匀划成k段，为每个区间分别计算（标准化后的）最小哈希值，再对k个最小哈希值求均值

方法提出于P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

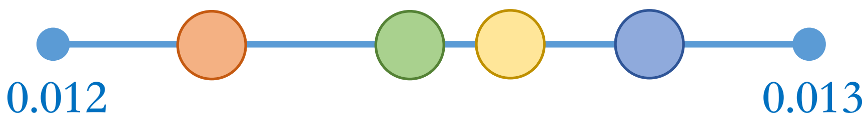
多个哈希函数

例如，把区间均分成1000段，先关注其中任一段：



多个哈希函数

例如，把区间均分成1000段，先关注其中任一段：

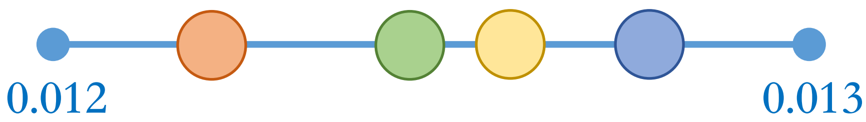


假设共q个不同的哈希值落入此段，这些哈希值都在[0.012,0.013]取值

将它们的哈希值记为$h(x_1), \ldots, h(x_q)$

多个哈希函数

例如，把区间均分成1000段，先关注其中任一段：



假设共q个不同的哈希值落入此段，这些哈希值都在[0.012,0.013]取值

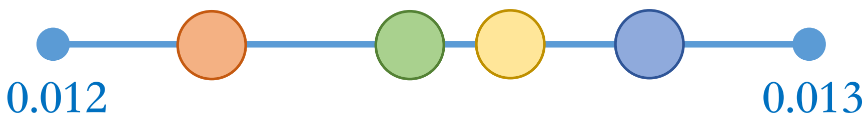
将它们的哈希值记为$h(x_1), \dots, h(x_q)$

分别将它们标准化为：$\frac{h(x_1)-0.012}{0.001}, \dots, \frac{h(x_q)-0.012}{0.001}$

标准化后为[0,1]区间**均匀分布**的随机变量

多个哈希函数

例如，把区间均分成1000段，先关注其中任一段：



假设共q个不同的哈希值落入此段，这些哈希值都在[0.012,0.013]取值

将它们的哈希值记为$h(x_1), \ldots, h(x_q)$

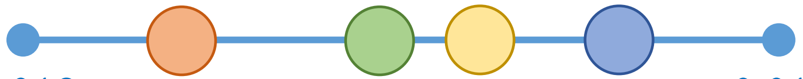
分别将它们标准化为：$\frac{h(x_1)-0.012}{0.001}, \ldots, \frac{h(x_q)-0.012}{0.001}$

标准化后为[0,1]区间**均匀分布**的随机变量

（标准化后的）最小哈希值即可用于估测q值：$q = \frac{1}{\mathbb{E}\{\text{minhash}\}} - 1$

多个哈希函数

例如，把区间均分成1000段，先关注其中任一段：



假设共q个不同的哈希值落入此段，这些哈希值都在[0.012,0.013]取值

将它们的哈希值记为$h(x_1), \ldots, h(x_q)$

分别将它们标准化为：$\frac{h(x_1)-0.012}{0.001}, \ldots, \frac{h(x_q)-0.012}{0.001}$

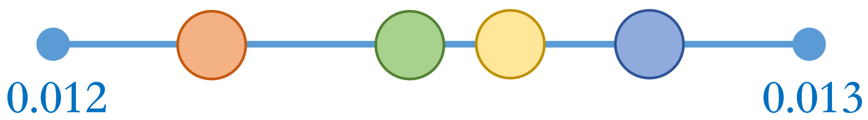
标准化后为[0,1]区间均匀分布的随机变量

（标准化后的）最小哈希值即可用于估测q值：$q = \frac{1}{\mathbb{E}\{\text{minhash}\}} - 1$

用1000段中得到的1000个（标准化后的）最小哈希值的均值$\overline{\text{minhash}}$近似$\mathbb{E}\{\text{minhash}\}$

多个哈希函数

例如，把区间均分成1000段，先关注其中任一段：



假设共q个不同的哈希值落入此段，这些哈希值都在[0.012,0.013]取值

将它们的哈希值记为$h(x_1), \dots, h(x_q)$

分别将它们标准化为：$\frac{h(x_1)-0.012}{0.001}, \dots, \frac{h(x_q)-0.012}{0.001}$

标准化后为[0,1]区间均匀分布的随机变量

（标准化后的）最小哈希值即可用于估测q值：$q = \frac{1}{\mathbb{E}\{\mathrm{minhash}\}} - 1$

用1000段中得到的1000个（标准化后的）最小哈希值的均值$\overline{\mathrm{minhash}}$近似$\mathbb{E}\{\mathrm{minhash}\}$

数据流中不同元素个数d可被估测为$1000\left(\frac{1}{\overline{\mathrm{minhash}}} - 1\right)$

多个哈希函数

算法一：为每个元素计算1000个哈希值

算法二：把区间均分成1000段，用$1000\left(\frac{1}{\overline{\mathrm{minhash}}}-1\right)$估测

算法三：把区间均分成1000段，用$1000\overline{\left(\frac{1}{\mathrm{minhash}}-1\right)}$估测

实验设置：跑20次实验计算估测的均值和标准差，不同元素个数d为766666

多个哈希函数

算法一：为每个元素计算1000个哈希值

算法二：把区间均分成1000段，用$1000\left(\frac{1}{\overline{\mathrm{minhash}}}-1\right)$估测

算法三：把区间均分成1000段，用$1000\overline{\left(\frac{1}{\mathrm{minhash}}-1\right)}$估测

实验设置：跑20次实验计算估测的均值和标准差，不同元素个数d为766666

算法一	$7.6836\mathrm{x}10^5 \pm 2.7702\mathrm{x}10^4$
算法二	$7.6475\mathrm{x}10^5 \pm 1.6956\mathrm{x}10^4$
算法三	$6.7950\mathrm{x}10^6 \pm 3.6642\mathrm{x}10^6$

20次估计值的均值及标准差

离散哈希值

初始化 $s_1, s_2, \dots, s_k \leftarrow 1$

对$i = 1, \dots, n$： $s_1 \leftarrow \min\{s_1, h_1(x_i)\}, \dots, s_k \leftarrow \min\{s_k, h_k(x_i)\}$

计算$\bar{s} = \frac{s_1+s_2+\cdots+s_k}{k}$，取$\frac{1}{\bar{s}} - 1$作为对不同元素个数的估计

将元素映射到连续空间（$\mathcal{U} \rightarrow [0,1]$）等价于将元素映射到无限长的二进制串（$\mathcal{U} \rightarrow \{0,1\}^{\infty}$）

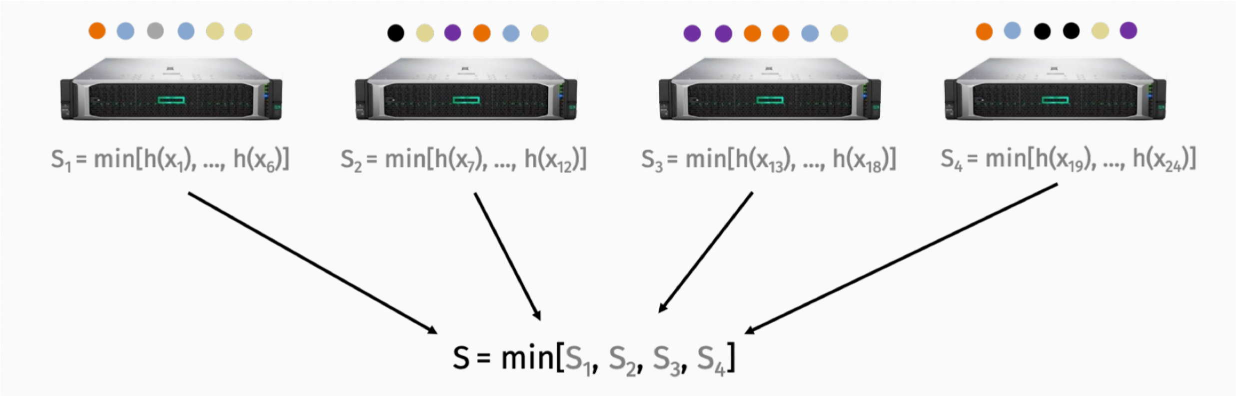
在实际中，一般将元素映射到离散空间$\{0,1,\dots,M-1\}$，若最小哈希值为S，可以用$\frac{M}{S}$估计d

$\frac{M}{S}$是近似后的值

类似结论可见A. Blum, J. Hopcroft, and R. Kannan <Foundations of Data Science>

最小哈希

可以分布式进行最小哈希的计算



图片取自Christopher Musco <NYU CS-GY 6763 (3943) Algorithmic Machine Learning and Data Science>

最小哈希

相关思想被Google PowerDrill, Facebook Presto, Twitter Algebird, Amazon Redshift等用于计算各类数据流中的不同元素个数，例如：

Use Case: Exploratory SQL-like queries on tables with 100's of billions of rows.

- Count number of distinct users in Germany that made at least one search containing the word 'auto' in the last month.
- Count number of distinct subject lines in emails sent by users that have registered in the last week, in comparison to number of emails sent overall (to estimate rates of spam accounts).

Answering a query requires a (distributed) linear scan over the database: 2 seconds in Google's distributed implementation.

例子取自Christopher Musco <NYU CS-GY 6763 (3943) Algorithmic Machine Learning and Data Science>

最小哈希

其它随机化方法

Flajolet — Martin算法

在前述方法（及相似方法）之外，有另一类随机化方法

前述方法

— *Order statistics observables*: these are based on order statistics, like the smallest (real) values, that appear in S. For instance, if $X = \min(S)$, we may legitimately hope that n is roughly of the order of $1/X$, since, as regards expectations, one has $\mathbb{E}(X) = 1/(n+1)$. The algorithms of Bar-Yossef *et al.* [2] and Giroire's MINCOUNT [16, 18] are of this type.

— *Bit-pattern observables*: these are based on certain patterns of bits occurring at the beginning of the (binary) S-values. For instance, observing in the stream S at the beginning of a string a bit-pattern $0^{\rho-1}1$ is more or less a likely indication that the cardinality n of S is at least 2^ρ. The algorithms known as *Probabilistic Counting*, due to Flajolet-Martin [15], together with the more recent LOGLOG of Durand-Flajolet [10] belong to this category.

评论引自P. Flajolet et al., <HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm>

Flajolet — Martin算法

通过分析二进制哈希值的末尾0个数进行估计

Probabilistic counting algorithms for data base applications
P Flajolet, GN Martin - Journal of computer and system sciences, 1985 - Elsevier
This paper introduces a class of probabilistic counting algorithms with which one can estimate the number of distinct elements in a large collection of data (typically a large file stored on …
☆ Save 99 Cite Cited by 1500 Related articles All 30 versions

Loglog counting of large cardinalities
M Durand, P Flajolet - European Symposium on Algorithms, 2003 - Springer
Using an auxiliary memory smaller than the size of this abstract, the LogLog algorithm makes it possible to estimate in a single pass and within a few percents the number of different …
☆ Save 99 Cite Cited by 407 Related articles All 23 versions

Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm
P Flajolet, É Fusy, O Gandouet… - Discrete Mathematics …, 2007 - hal.archives-ouvertes.fr
This extended abstract describes and analyses a near-optimal probabilistic algorithm, HYPERLOGLOG, dedicated to estimating the number of\emphdistinct elements (the cardinality) of very large data ensembles. Using an auxiliary memory of m units (typically," short bytes"), HYPERLOGLOG performs a single pass over the data and produces an estimate of the cardinality such that the relative accuracy (the standard error) is typically about $1.04/\sqrt {m} $. This improves on the best previously known cardinality estimator …
☆ Save 99 Cite Cited by 645 Related articles All 39 versions »

Flajolet — Martin算法

$h(x_1)$	1010010
$h(x_2)$	1001100
$h(x_3)$	1001110
⋮	
$h(x_n)$	1011000

核心思想：通过末尾0个数的最大值R估计不同元素的个数

不同元素越多，末尾0个数的最大值越大

图片取自Cameron Musco <COMPSCI 514 - Algorithms for Data Science>

Flajolet — Martin算法

方法：用一个0-1数组记录输入元素的（二进制）哈希值中右起一串零的最靠左的零出现的位置

$h(x_1)$	1010010
$h(x_2)$	1001100
$h(x_3)$	1001110
⋮	
$h(x_n)$	1011000

000001

详细分析见P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

Flajolet — Martin算法

方法：用一个0-1数组记录输入元素的（二进制）哈希值中**右起一串零的最靠左的零**出现的位置

$h(x_1)$	1010010
$h(x_2)$	1001100
$h(x_3)$	1001110
⋮	
$h(x_n)$	1011000

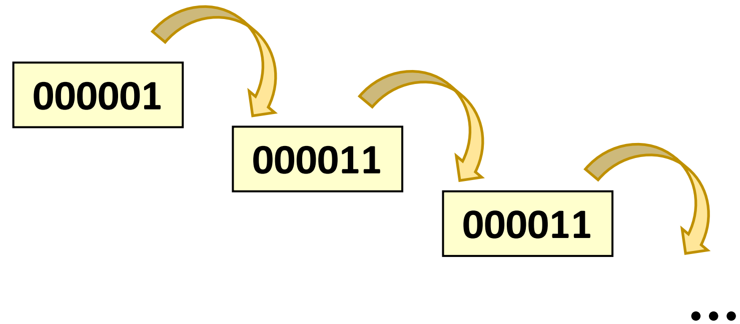
000001 → 000011

详细分析见P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

Flajolet — Martin算法

方法：用一个0-1数组记录输入元素的（二进制）哈希值中右起一串零的最靠左的零出现的位置

$h(x_1)$	1010010
$h(x_2)$	1001100
$h(x_3)$	1001110
⋮	
$h(x_n)$	1011000



实际的二进制位数远大于7，可取比如24、32

详细分析见P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

Flajolet — Martin算法

方法：用一个0-1数组记录输入元素的（二进制）哈希值中**右起一串零的最靠左的零**出现的位置

0 0 0 1 0 1 0 1 1 1 1 1

此部分分析为论文为介绍该方法做的基于直觉的分析，非严谨数学分析

大约有$\frac{d}{4}$个数可使此位置1（即mod4=2的数）

大约有$\frac{d}{8}$个数可使此位置1（即mod8=4的数）

右起第i位（$i = 1, \ldots$）：大约有$\frac{d}{2^{i+1}}$个数可使此位置1

详细分析见P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

Flajolet — Martin算法

方法：用一个0-1数组记录输入元素的（二进制）哈希值中右起一串零的最靠左的零出现的位置

0 0 0 1 0 1 0 1 1 1 1 1

此部分分析为论文为介绍该方法做的基于直觉的分析，非严谨数学分析

大约有$\frac{d}{4}$个数可使此位置1（即mod4=2的数）

大约有$\frac{d}{8}$个数可使此位置1（即mod8=4的数）

右起第i位（$i = 1, \ldots$）：大约有$\frac{d}{2^{i+1}}$个数可使此位置1

当$i \gg \log_2 d$时，第i位大概率为0
当$i \ll \log_2 d$时，第i位大概率为1
当$i \approx \log_2 d$时，第i位以一定概率分别取0和1

详细分析见P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications>

Flajolet — Martin算法

方法：用一个0-1数组记录输入元素的（二进制）哈希值中
右起一串零的最靠左的零出现的位置

0 0 0 1 0 1 0 1 1 1 1 1

此部分分析为论文为介绍该方法做的
基于直觉的分析，非严谨数学分析

大约有$\frac{d}{4}$个数可使此位置1（即mod4=2的数）

大约有$\frac{d}{8}$个数可使此位置1（即mod8=4的数）

右起第i位（$i = 1, \ldots$）：大约有$\frac{d}{2^{i+1}}$个数可使此位置1

当$i \gg \log_2 d$时，第i位大概率为0
当$i \ll \log_2 d$时，第i位大概率为1
当$i \approx \log_2 d$时，第i位以一定概率分别取0和1

若末尾0个数的最大值为R，有$R = O(\log_2 d)$

Flajolet — Martin算法

设末尾0个数的最大值为R，论文<Probabilistic counting algorithms for data base applications>严格分析了$\mathbb{E}\{R\}$与不同元素个数d的关系

论文中用n表示不同元素个数

THEOREM 3.A. *The average value of parameter R_n satisfies:*

$$\bar{R}_n = \log_2(\varphi n) + P(\log_2 n) + o(1),$$

where constant $\varphi = 0.77351 \ldots$ *is given by*

$$\varphi = 2^{-1/2} e^{\gamma} \frac{2}{3} \prod_{p=1}^{\infty} \left[\frac{(4p+1)(4p+2)}{(4p)(4p+3)} \right]^{(-1)^{\nu(p)}}$$

and $P(u)$ *is a periodic and continuous functions of* u *with period* 1 *and amplitude bounded by* 10^{-5}.

该方法也需要用随机平均（stochastic averaging）的技巧提升准确度，即用多个不同的哈希函数或将数据平均分成多个段

Flajolet — Martin算法

论文<Probabilistic counting algorithms for data base applications>实验结果

TABLE III

Sample Executions of Algorithm PCSA on 6 Files with the Same Multiplicative Hashing Function

File	Card.	8	16	32	64	128	256
man 1	16405	17811	16322	14977	15982	16690	17056
		1.08	*0.99*	*0.91*	*0.97*	*1.01*	*1.03*
man 1.w	38846	40145	40566	40145	43290	41230	42592
		0.96	*1.01*	*0.96*	*1.07*	*1.02*	*1.06*
man 2	3149	2427	2887	3015	3015	2840	2982
		0.77	*0.91*	*0.95*	*0.95*	*0.90*	*0.94*
man 2.w	10560	10590	9711	9100	9100	10032	10734
		1.00	*0.91*	*0.86*	*0.86*	*0.95*	*1.01*
man 8	3075	4452	3744	3360	3252	3097	3106
		1.44	*1.21*	*1.09*	*1.05*	*1.00*	*1.01*
man 8.w	11334	10590	10590	10363	10705	10999	10676
		0.93	*0.93*	*0.91*	*0.94*	*0.97*	*0.94*

Note. The figure displays the file name, the exact cardinality, the estimated cardinality for nmap = 8, 16, 32, 64, 128, 256, and the ratio of estimated cardinalities to exact cardinalities (in italics).

数据分成段数

采用0-1数组长度 $\boxed{L > \log_2(n/\text{nmap}) + 4.}$

本讲小结

- 不同元素统计问题
- 最小哈希的方法与应用

主要参考资料

Cameron Musco <COMPSCI 514 - Algorithms for Data Science> Slides

A. Blum, J. Hopcroft, and R. Kannan <Foundations of Data Science> Book

Christopher Musco <NYU CS-GY 6763 (3943) Algorithmic Machine Learning and Data Science> Slides

P. Flajolet et al., <HyperLogLog: The analysis of a near-optimal cardinality estimation algorithm> Paper

P. Flajolet, GN Martin. <Probabilistic counting algorithms for data base applications> Paper

谢谢！