HIT CS&E

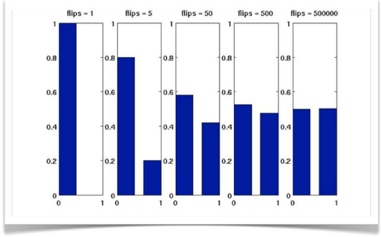
第4章　Chernoff界

骆吉洲
计算机科学与技术学院

HIT CS&E

随机实验结果的集中性

Flip a coin for many times:



- 几何分布(两点分布之和)的集中性
- 泊松实验之和的集中性
  泊松实验—0-1随机变量(不必独立同分布)

HIT CS&E

提纲

4.1 Chernoff界的导出及常用形式
- **4.1.1 矩生成函数**
- **4.1.2 Chernoff界的导出**
- **4.1.3 Chernoff界的常用形式**
- **4.1.4 简单应用**

4.2 特殊情况下更好的*Chernoff*界
4.3 集合平衡配置的随机算法
4.4 超方体上排列行路由的随机算法

HIT CS&E

参考文献

- 《Randomized Algorithms》
  第4章
- 《概率与计算》
  第4章

HIT CS&E

4.1 Chernoff界及其常用形式

- 矩生成函数
- Chernoff界的导出
- Chernoff界的常用形式

HIT CS&E

4.1.1 矩生成函数

HIT CS&E

矩生成函数

定义：随机变量X的矩生成函数指的是

$$M(\lambda)=\mathrm{E}\left[e^{\lambda X}\right]$$

X的各阶矩融入同一函数

由泰勒展开可知

$$\mathrm{E}\left[e^{\lambda X}\right]=\mathrm{E}\left[\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}X^k\right] = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{E}[X^k]$$

取k阶导数，再令$\lambda=0$

$$\mathrm{E}[X^n]=M^{(n)}(0)$$

HIT CS&E

矩生成函数的性质

$$M(\lambda)=\mathrm{E}\left[e^{\lambda X}\right] = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{E}[X^k]$$

性质1：两个随机变量的矩生成函数相同，则这两个随机变量相同

性质2：两个随机变量的各阶矩相同，则这两个随机变量相同

$\mathrm{E}[e^{\lambda(X+Y)}] = \mathrm{E}[e^{\lambda X}e^{\lambda Y}]=\mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\lambda Y}]$ 独立即可

性质3：两个独立随机变量之和的矩生成函数则等于这两个随机变量的矩生成函数之积

HIT CS&E

例

$$M(\lambda)=\mathrm{E}\left[e^{\lambda X}\right]$$

考虑两点分布：$\Pr[X=1]=p \quad \Pr[X=0]=1-p$

$$M_X(t)=\Pr[X=1]\cdot e^t+\Pr[X=0]\cdot e^0$$
$$=p\cdot e^t+(1-p)$$

考虑二项分布Y——n个独立同分布的两点分布之和

$$M_Y(t)=(M_X(t))^n=[1-p+pe^t]^n$$
$$\mathrm{E}[Y]=M_Y^{(1)}(0)=np[1-p+pe^t]^{n-1}e^t|_{t=0}=np$$
$$\mathrm{E}[Y^2]=M_Y^{(2)}(0)=n(n-1)p^2[1-p+pe^t]^{n-2}e^{2t}+np[1-p+pe^t]^{n-1}e^t|_{t=0}$$
$$=n(n-1)p^2+np$$
$$\mathrm{Var}[Y]=(\mathrm{E}[Y^2]-\mathrm{E}[Y])^2)=np-np^2=np(1-p)$$

请你试试用矩生成函数计算几何分布的方差？

HIT CS&E

4.1.2 Chernoff界的导出

HIT CS&E

Chernoff界

Chernoff界

定理：$X_1,\ldots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$，$\mu=\mathrm{E}[X]$，则对任意$\delta>0$有

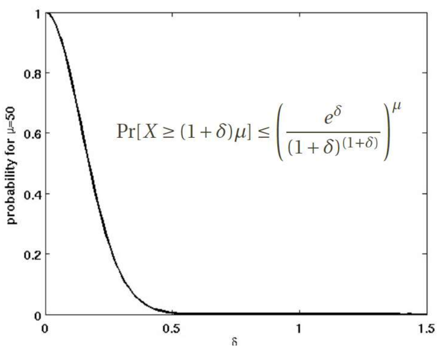
$$\Pr[X\geq(1+\delta)\mu]<\left[\frac{e^\delta}{(1+\delta)^{1+\delta}}\right]^\mu$$

提供了估计尾概率$\Pr[X>t]$的新工具

- Markov不等式也是这样的工具 为啥还要Chernoff界？
- Chebyshev不等式也是这样的工具
- Chernoff界给出的概率界更准

HIT CS&E

概率变化趋势示例



HIT CS&E

导出思路

定理：$X_1,\dots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$，$\mu=E[X]$，则对任意$\delta>0$有

$$\Pr[X\geq(1+\delta)\mu]<\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

第1步：计算X的矩生成函数$E[e^{\lambda X}]$

第2步：给出$\Pr[X>t]$的表达式，其中含有t作为参数

应用Markov不等式 $\Pr[h(X)\geq t]\leq\frac{E[h(X)]}{t}$

$h(X)=E[e^{\lambda X}]$

第3步：针对λ进行优化，得出结论

HIT CS&E

第1步：矩生成函数

$X_1,\dots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$

$(\forall i)$：$\Pr[X_i=1]=p_i$　$\Pr[X_i=0]=1-p_i$

$$E[e^{\lambda X_i}]=\Pr[X_i=1]\cdot e^{\lambda}+\Pr[X_i=0]\cdot e^{0}$$
$$=p_i\cdot e^{\lambda}+(1-p_i)$$
$$=1+p_i\cdot(e^{\lambda}-1)$$
$$\leq e^{p_i(e^{\lambda}-1)}\qquad 1+x\leq e^x\quad x>0$$

$$E[e^{\lambda X}]=E[e^{\lambda X_1}]\cdot E[e^{\lambda X_2}]\cdot\ldots\cdot E[e^{\lambda X_n}]\qquad \text{独立性}$$
$$\leq e^{p_1(e^{\lambda}-1)}\cdot e^{p_2(e^{\lambda}-1)}\cdot\ldots\cdot e^{p_n(e^{\lambda}-1)}$$
$$=e^{\mu(e^{\lambda}-1)}\qquad \mu=E[X]=p_1+\ldots+p_n$$

HIT CS&E

第2步：Markov不等式

$X_1,\dots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$

$E[e^{\lambda X}]\leq e^{\mu(e^{\lambda}-1)}$

$$\Pr[X\geq t]=\Pr[e^{\lambda X}\geq e^{\lambda t}]\qquad \lambda>0,\ X>0$$
$$\leq\frac{E[e^{\lambda X}]}{e^{\lambda t}}\qquad \text{Markov不等式}$$
$$\leq\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda t}}$$

$$\Pr[X\leq t]=\Pr[e^{\lambda X}\geq e^{\lambda t}]\qquad \lambda<0$$
$$\leq\frac{E[e^{\lambda X}]}{e^{\lambda t}}$$
$$\leq\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda t}}$$

$$\Pr[X\geq t]\leq\min_{\lambda>0}\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda t}}\qquad \Pr[X\leq t]\leq\min_{\lambda<0}\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda t}}$$

HIT CS&E

第3步：优化

$$\Pr[X\geq t]\leq\min_{\lambda>0}\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda t}}\qquad \Pr[X\leq t]\leq\min_{\lambda<0}\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda t}}$$

$$\Pr[X\geq(1+\delta)\mu]\leq\min_{\lambda>0}\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda(1+\delta)\mu}}\qquad \Pr[X\leq(1-\delta)\mu]\leq\min_{\lambda<0}\frac{e^{\mu(e^{\lambda}-1)}}{e^{\lambda(1-\delta)\mu}}$$

令$\lambda=\ln(\delta+1)$　　令$\lambda=\ln(1-\delta)$

$$\Pr[X\geq(1+\delta)\mu]<\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}\qquad \Pr[X\leq(1-\delta)\mu]<\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu}$$

$\forall\delta>0$　　$0<\forall\delta<1$

HIT CS&E

结论

Chernoff界

定理：$X_1,\dots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$，$\mu=E[X]$，则对任意$\delta>0$有

$$\Pr[X\geq(1+\delta)\mu]<\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

对任意$1>\delta>0$有

$$\Pr[X\leq(1-\delta)\mu]<\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu}$$

HIT CS&E

4.1.3 Chernoff界的常用形式

HIT CS&E

Chernoff界给出了下面的尾不等式

$$\Pr[\,X \ge (1+\delta)\mu\,] < \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu} \le e^{-\mu\delta^2/3}$$

$$\Pr[\,X \le (1-\delta)\mu\,] < \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu} \le e^{-\mu\delta^2/2}$$

为了使不等式更方便使用
需要找出右端更简洁的表达形式

HIT CS&E

$$\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu} \le e^{-\mu\delta^2/3} \qquad 0<\delta<1$$

$$\Updownarrow$$

$$\mu\,[\delta - (1+\delta)\ln(1+\delta)] \le -\mu\delta^2/3$$

$$\Updownarrow$$

$f(\delta) = \delta^2/3 + \delta - (1+\delta)\ln(1+\delta) \le 0$

$f(0)=0$ $\quad f(\delta)<0$

$f'(\delta) = 2\delta/3 - \ln(1+\delta)$ $\quad f'(0)=0$ $\quad f'(1)<0$ $\quad f'(\delta)<0$

$f''(\delta) = 2/3 - 1/(1+\delta)$ $\quad f''(\delta)<0 \quad \delta<1/2$

$f''(\delta)>0 \quad \delta>1/2$

HIT CS&E

$$\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu} \le e^{-\mu\delta^2/2} \qquad 0<\delta<1$$

$$\Updownarrow$$

$$\mu\,[-\delta - (1-\delta)\ln(1-\delta)] \le -\mu\delta^2/2$$

$$\Updownarrow$$

$g(\delta) = \delta^2/2 - \delta - (1-\delta)\ln(1-\delta) \le 0$

$g(0)=0$ $\quad g(\delta)<0$

$g'(\delta) = \delta + \ln(1-\delta)$ $\quad g'(0)=0$ $\quad g'(\delta)<0$

$g''(\delta) = 1 - 1/(1-\delta)$ $\quad g''(\delta)<0$

HIT CS&E

常用形式

Chernoff界

定理：$X_1,\ldots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n} X_i$，$\mu=E[X]$，则对任意$1>\delta>0$有

$\Pr[\,X \le (1-\delta)\mu\,] < e^{-\mu\delta^2/2}$

$\Pr[\,X \ge (1+\delta)\mu\,] < e^{-\mu\delta^2/3}$

$\Pr[\,|X-\mu| \ge \delta\mu\,] < 2e^{-\mu\delta^2/3}$

对任意$t>2e\mu$有

$\Pr[\,X>t\,] < 2^{-t}$

HIT CS&E

4.1.4 简单应用

HIT CS&E

二项分布

成功实验的总次数

- 独立同分布重复Bernoulli实验n次
- X_i=1表示第i次试验成功
- X—成功实验的总次数 $\quad X=\sum_i X_i$ $\quad E[X]=n/2$

$$\Pr[\,|X-E[X]| \ge E[X]/2\,]$$
$$= \Pr[\,|X-n/2| \ge n/4\,]$$
$$< 2\exp\{-(n/2)(1/2)^2/3\}$$
$$< 2e^{-n/24}$$

算法重复遍数

LAZYSELECT(S, k)

1. R=独立、均匀、可放回地从S随机选取的$n^{3/4}$元素;
2. 在$O(n)$时间内排序R;
3. $x=(k/n)n^{3/4}$;
4. l=max{ , 0}; h= min{ , $n^{3/4}$};
5. $L=min(R, l)$; $H=min(R, h)$;
6. $L_p=Rank(S,L)$, $H_p=Rank(S,H)$; (参见第2章)
7. $P=\{y\in S \mid L\le y\le H\}$;
8. If $min(S, k)\in P$ and $|P|\le 4n^{3/4}+1$
9. Then 排序P, $min(S, k)=min(P, (k-L_p))$, 算法结束;
10. ELSE goto 1.

1-9步运行20遍还未找到正确解的概率有多大?

算法重复遍数

算法A运行一遍得到解的概率为p

重复调用算法A直到得到问题的解

将算法运行遍数记为随机变量X ——几何分布

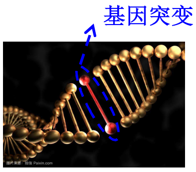
$E[X] = 1/p$

$E[X \ge (1+\delta)/p] < 2\exp\{-\delta^2/(3p)\} = p_0/2 \quad \delta_1=\sqrt{\frac{-3p\ln p_0}{2\ln 2}}$

$E[X \le (1-\delta)/p] < 2\exp\{-\delta^2/(2p)\} = p_0/2 \quad \delta_2=\sqrt{\frac{-2p\ln p_0}{2\ln 2}}$

算法运行遍数以$1-p_0$的概率介于$(1-\delta_2)/p$和$(1+\delta_2)/p$之间

参数估计



DNA序列

设该基因突变的概率为p(未知)

用统计规律来估计p

- 实验测试n次
- 测得突变X次
- $q = X/n$
- 实验测试很昂贵
- $q\approx p$可信不可信?

置信区间

- $\Pr[\, p\notin[q-\delta, q+\delta]\,] < 1-\gamma$
- $[q-\delta, q+\delta]$称为置信区间
- γ称为置信水平

能否将实验次数n和置信水平γ关联起来?

$p\notin[q-\delta, q+\delta]$

$p<q-\delta$	或	$p>q+\delta$
$np<nq-n\delta$		$np>nq+n\delta$
$nq>np+n\delta$		$nq<np-n\delta$
$X>E[X](1+\delta/p)$		$X<E[X](1-\delta/p)$

$\Pr[\, X>E[X](1+\delta/p)\,]$

$< \exp(-np(\delta/p)^2/3)$

$\le \exp(-n\delta^2/3)$

$\Pr[\, X<E[X](1-\delta/p)\,]$

$< \exp(-np(\delta/p)^2/2)$

$\le \exp(-n\delta^2/2)$

$\Pr[\, p\notin[q-\delta, q+\delta]\,] < \exp(-n\delta^2/3) + \exp\{-n\delta^2/2\}$

$\Pr[\, p\notin[q-\delta, q+\delta]\,] < \exp(-n\delta^2/3) + \exp\{-n\delta^2/2\}$

- 已知n,δ, 可以计算置信水平
- 已知n,γ, 可以计算δ，即置信区间
- 已知δ,γ, 可以计算实验次数n

作业

针对第2章的数值随机算法
- 计算π
- 计算定积分

求解精度表示为置信区间

分别用Chernoff界建立抽样次数n与求解精度间的关系

HIT CS&E

重审球和箱子模型

m个球

均匀独立地将球投入箱子：最大负载*whp*?

n个箱子

X_1—第1个箱子的负载

$$x_{ij}=\begin{cases}1 & 第j个球落入第i个箱子\\ 0 & 否则\end{cases}\quad \Pr[X_{ij}=0]=1-1/n$$

$$X_1=\sum_{j=1}^{m}X_{1j}\qquad \mu=\mathbf{E}[X_1]=\frac{m}{n}$$

X_1服从参数为m和$1/n$的二项分布

HIT CS&E

X_1—第1个箱子的负载 $X_1=\sum_{j=1}^{m}X_{1j}$ $\mu=\mathbf{E}[X_1]=\frac{m}{n}$

$$\Pr[X\ge(1+\delta)\mu]<\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

情形1：$m=n$ $\mu=1$

$$\Pr[X\ge L]\le\frac{e^L}{eL^L}\le\frac{1}{n^2}\qquad 取L=\frac{e\ln n}{\ln\ln n}$$

最大负载为$O\left(\frac{\ln n}{\ln\ln n}\right)$的概率为$1-\frac{1}{n}$

HIT CS&E

X_1—第1个箱子的负载 $X_1=\sum_{j=1}^{m}X_{1j}$ $\mu=\mathbf{E}[X_1]=\frac{m}{n}$

$$\Pr[X\ge t]\le 2^{-t}\text{ for }t\ge 2e\mu$$

情形2：$m\ge n\ln n$ $\mu\ge\ln n$

$$\Pr\left[X_1\ge\frac{2em}{n}\right]=\Pr[X_1\ge 2e\mu]\le 2^{-2e\mu}\le 2^{-2e\ln n}<\frac{1}{n^2}$$

最大负载为$O\left(\frac{m}{n}\right)$的概率为$1-\frac{1}{n}$

HIT CS&E

4.2 特殊情况下更好的界

HIT CS&E

Chernoff界

定理：$X_1,\ldots,X_n$是独立随机变量，$\Pr[X_i=1]=\Pr[X_i=-1]=1/2$，$X=\sum_{i=1}^{n}X_i$，$\mu=E[X]$，则对任意$t>0$有

$$\Pr[X\ge t]<e^{-t^2/2n}$$

HIT CS&E

Chernoff界

定理：$X_1,\ldots,X_n$是独立随机变量，$\Pr[X_i=1]=\Pr[X_i=-1]=1/2$，$X=\sum_{i=1}^{n}X_i$，$\mu=E[X]$，则对任意$t>0$有

$$\Pr[|X|\ge t]<2e^{-t^2/2n}$$

HIT CS&E

Chernoff界

定理：$X_1,\ldots,X_n$是独立随机变量，$\Pr[X_i=1]=\Pr[X_i=0]=1/2$，$X=\sum_{i=1}^{n} X_i$，$\mu=E[X]=n/2$，则

对任意$t>0$有

$$\Pr[\,X\ge \mu+t\,]<e^{-2t^2/n}$$

对任意$\mu>t>0$有

$$\Pr[\,X\le \mu-t\,]<e^{-2t^2/n}$$

对任意$\delta>0$有

$$\Pr[\,X\ge(1+\delta)\mu\,]<e^{-\delta^2\mu}$$

对任意$1>\delta>0$有

$$\Pr[\,X\le(1-\delta)\mu\,]<e^{-\delta^2\mu}$$

HIT CS&E

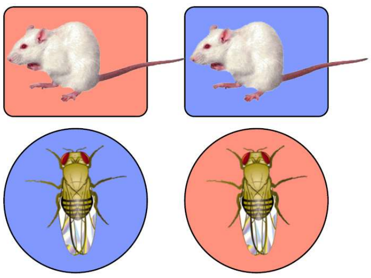
4.2 集合平衡配置

- 应用背景
- 集合平衡配置的随机算法

HIT CS&E

应用背景

科学控制

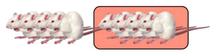


HIT CS&E

相关性 ≠ 因果性

探寻单个因素与某种现象之间的相关程度

将同样的实验对象分为两组
在其中一组对象上开展实验



如果实验对象是不同的，该怎么处理呢？

HIT CS&E

实验对象分组

- 将实验对象按统计规律分为两
- 每个实验对象均有一系列特征$f_1,f_2,\ldots,f_m$
- 对每个特征($\forall f_i$)

	subject 1	subject 2	subject 3	subject 4
gender	boy	girl	boy	girl
age	≥3	≥3	<3	<3
teeth	un-healthy	un-healthy	healthy	un-healthy
bodyfat	normal	over-weight	normal	normal
allergy	no	no	yes	yes

第一组中具有特征f_i的实验对象总数 ≈ 第二组中具有特征f_i的实验对象总数

HIT CS&E

实验对象分组

- 将实验对象按统计规律分为两组
- 每个实验对象均有一系列特征$f_1,f_2,\ldots,f_m$
- 对每个特征($\forall f_i$)

	$subject_1$	$subject_2$	$subject_3$	$subject_4$
$feature_1$	1	0	1	0
$feature_2$	1	1	0	0
$feature_3$	0	0	1	0
$feature_4$	1	0	1	1
$feature_5$	0	1	0	1

第一组中具有特征f_i的实验对象总数 ≈ 第二组中具有特征f_i的实验对象总数

集合平衡配置问题

问题定义

输入：$n\times m$的0-1矩阵A

$$\begin{array}{c|c} & \text{对象1} \quad \text{对象2} \quad \text{对象3} \quad \cdots \quad \text{对象}m \qquad \in\{1,-1\}^m \\ \begin{array}{c}\text{特征}f_1: \\ \text{特征}f_2: \\ \text{特征}f_3: \\ \\ \text{特征}f_n:\end{array} & \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix} \end{array}$$

输出：将A的m列分为两组使得每组对应元素之和的差最小

$$c_i = \sum_{x_j=1} a_{ij} - \sum_{x_j=-1} a_{ij}$$

$$\max_{1\le i\le n} |c_i| = |c|_\infty$$

问题定义

输入：$n\times m$的0-1矩阵A

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix}$$

输出：向量$x=(x_1,\ldots,x_m)^T\in\{1,-1\}^m$使得 $|Ax|_\infty$最小

怎么解呢？

平衡配置的随机算法

输入：0-1矩阵$A_{n\times m}$

输出：$x\in\{-1,+1\}^m$使 $\min |Ax|_\infty$

$(\forall 1\le j\le m)$随机独立取 $x_j\in\{-1,+1\}$

$$x_j = \begin{cases} +1 & \Pr[x_j=+1] = 1/2 \\ -1 & \Pr[x_j=-1] = 1/2 \end{cases}$$

随机配置算法

- m个实验对象，两个组
- 为每个实验对象均匀随机、独立地指定一个分组
- 不考虑输入矩阵

这也行？

随机配置算法的性能

定理：对于任意0-1矩阵$A_{n\times m}$和任意均匀随机独立选取的向量$x\in\{-1, +1\}^m$, 有

$$\Pr[\ |Ax|_\infty > \sqrt{12m\ln n}\ ] < \frac{2}{n}$$

分析思路

$(\forall 1\le j\le m)$随机独立取 $x_j\in\{-1,+1\}^m$

$$x_j = \begin{cases} +1 & \Pr[x_j=+1] = 1/2 \\ -1 & \Pr[x_j=-1] = 1/2 \end{cases}$$

$$\|Ax\|_\infty = \max_{1\le i\le n} |(Ax)_i| = \max_{1\le i\le n} \left|\sum_{j=1}^m a_{ij}x_j\right|$$

先固定第1行，分析 $\left|\sum_{j=1}^m a_{1j}x_j\right|$ 超过一定阈值的概率

然后，用Union Bound得出$|Ax|_\infty$超过一定阈值的概率

第1行

$(\forall 1\le j\le m)$随机独立取 $x_j\in\{-1,+1\}^m$

$$x_j = \begin{cases} +1 & \Pr[x_j=+1] = 1/2 \\ -1 & \Pr[x_j=-1] = 1/2 \end{cases}$$

概率上界 $\Pr\left[\left|\sum_{j=1}^m a_{1j}x_j\right| > t\right]$

- 随机变量之和超过阈值的概率
- 复杂之处在于：
  - 求和项为1,-1 (而非随机实验的直接结果)
  - 系数为0,1

HIT CS&E

$a_1.$

1	1	0	1	0	0	1	1	0	0
+1	-1	-1	+1	+1	-1	-1	+1	-1	+1

x

坏事件：$\left|\sum_{j=1}^{m} a_{1j}x_j\right| > t$

矩阵第1行：含有k个1

- 情形1：$k < t$ 坏事件不会发生
- 情形2：$k \geq t$ 坏事件发生 $\Leftrightarrow$ 投掷均匀硬币k次 $|\text{HEADs} - \text{TAILs}| > t$

HIT CS&E

$a_1.$

1	1	0	1	0	0	1	1	0	0
H	T		H			T	H		

x

投掷均匀硬币k次 ($t \leq k \leq m$)

$|\text{HEADs} - \text{TAILs}| > t$ $\Leftrightarrow |Y-\mu| > t/2$

Y_j—第j次投掷硬币的结果

$$Y_j = \begin{cases} 1 & \Pr[Y_j=1] = 1/2 \\ 0 & \Pr[Y_j=0] = 1/2 \end{cases}$$

$Y=\sum_{j=1}^{k} Y_j$

$\mu = E[Y] = k/2$

HEADs – TAILs > t $Y < \frac{k}{2} - \frac{t}{2}$

TAILs – HEADs > t $Y > \frac{k}{2} + \frac{t}{2}$

HIT CS&E

Chernoff界：$Y_1,\ldots,Y_k$是独立泊松实验，$Y=\sum_{i=1}^{k} Y_i$, $\mu=E[Y]$，则 $\Pr[\,|Y-\mu| > \delta\mu\,] < 2e^{-\mu\delta^2/3}$ 对$0<\delta<1$成立

$$\Pr\left[\,\left|\sum_{j=1}^{m} a_{1j}x_j\right| > t\,\right]$$

$\leq \Pr[\,|\text{HEADs} - \text{TAILs}| > t\,]$

$= \Pr[|Y-\mu| > t/2]$

$= \Pr[|Y-\mu| > \delta\mu]$ $\delta = t/k$

$< 2\cdot \exp\{-(k/2)(t/k)^2/3\}$

$= 2\cdot \exp\{-t^2/6k\}$ $k \leq m$ 取$t=\sqrt{12m\ln n}$

$= \frac{2}{n^2}$

HIT CS&E

利用Union Bound

$\Pr[\,|Ax|_\infty > \sqrt{12m\ln n}\,]$

$= \Pr\left[\,\exists i: \left|\sum_{j=1}^{m} a_{ij}x_j\right| > \sqrt{12m\ln n}\,\right]$

$\leq n\Pr\left[\,\left|\sum_{j=1}^{m} a_{1j}x_j\right| > \sqrt{12m\ln n}\,\right]$

$= \frac{2}{n}$

HIT CS&E

结论

定理：对于任意0-1矩阵$A_{n\times m}$和任意均匀随机独立选取的向量$x\in\{-1,+1\}^m$，有

$$\Pr[\,|Ax|_\infty > \sqrt{12m\ln n}\,] < \frac{2}{n}$$

练习：存在0-1矩阵$A_{n\times n}$使得

$$|Ax|_\infty = \Omega(\sqrt{n})$$

对任意向量$x\in\{-1,+1\}^n$成立

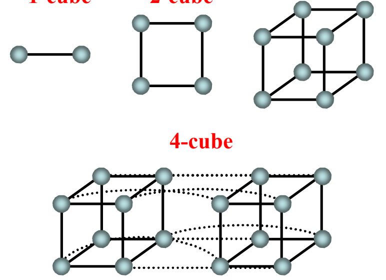
HIT CS&E

4.3 超方体上的随机路由算法

- 超方体拓扑结构
- Bit Fixing路由算法
- 随机路由算法

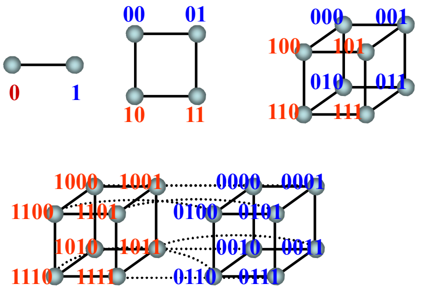
HIT CS&E

n-Cube



HIT CS&E

n-超方体的顶点标记



HIT CS&E

n-超方体的形式定义

n-超方体

- 顶点：$N=2^n$个
  - 每个顶点i唯一地标记为二进制串$(i_1,\ldots,i_n)\in\{0,1\}^n$
- 边：
  - $i\leftrightarrow j \quad \Leftrightarrow (i_1,\ldots,i_n)$与$(j_1,\ldots,j_n)$间的Haming距离等于1
- 度：每个顶点有n条邻接边
- 边数： $n2^{n-1} = (\log N)\cdot(N^{1/2})$

HIT CS&E

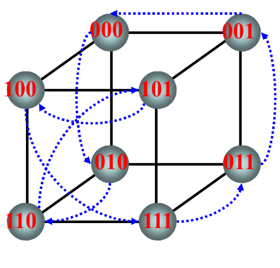
排列路由问题

n-超方体上的排列路由问题

- 输入
  - *n*-超方体
  - $(\forall i = 0,1,\ldots,2^n-1)$消息$m(i)$,源顶点$i$, 目的顶点$d(i)$
  - $d(0),\ldots,d(2^n-1)$是$0,1,\ldots,2^n-1$的一个排列
- 输出：
  （$\forall i$）消息$m(i)$从i到$d(i)$的路由
- 路由规则
  - 同步路由(分为若干步，每个步骤通信同步进行)
  - 一个同步步骤内
    - ✓ 每边每向仅有一个消息通过
    - ✓ 每个顶点至多向其每个相邻顶点发送一个消息（每个顶点每步至多发送*n*个消息）

HIT CS&E

实例

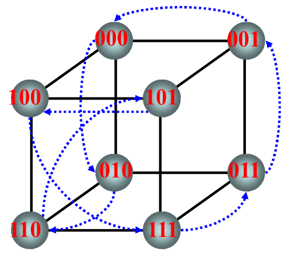


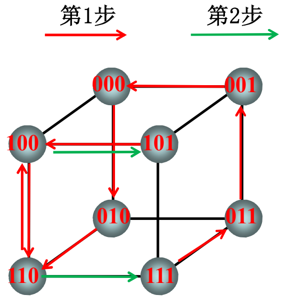
$d(000)=010$
$d(001)=000$
$d(010)=110$
$d(011)=001$
$d(100)=111$
$d(101)=100$
$d(110)=101$
$d(111)=011$

HIT CS&E

实例

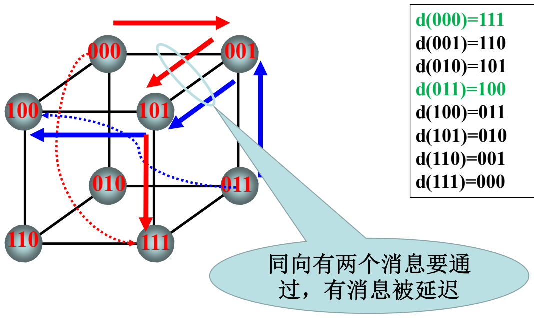
第1步　　第2步





HIT CS&E

流量控制



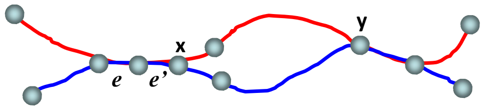
d(000)=111
d(001)=110
d(010)=101
d(011)=100
d(100)=011
d(101)=010
d(110)=001
d(111)=000

HIT CS&E

遗忘性路由算法

遗忘性路由算法
- 消息$m(i)$从i路由到j所经过的路径仅依赖于i和j的标记
- “路由路径无记忆性”便于路由器硬件的实现

问题
- 两条路由路径在e的端点汇合，在e'的端点x分离
- 那么这两条路由路径还能再次汇合到某个顶点y吗？



答：不能
- x,y之间只能有一条路由路径
- x,y之间的路由路径是由x,y的标记唯一确定的

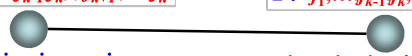
HIT CS&E

Bit Fixting路由算法

输入：超方体拓扑结构，消息$m(i)$及目的地址$d(i)$
输出：$(\forall i)$将$m(i)$从顶点i路由到$d(i)$
算法： 路由的一个同步步骤如下构成
1. $(\forall k)$顶点k检查消息队列Q_k中是否有消息要送出
2. 若有，则执行如下的消息传送并从队列中删除
3. 监听并接受来自邻居的消息放入队列

$m(i)$
S: $i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_n$
C: $j_1,\ldots,j_{k-1},i_k,i_{k+1},\ldots,i_n$
D: $j_1,\ldots,j_{k-1},j_k,j_{k+1},\ldots,j_n$

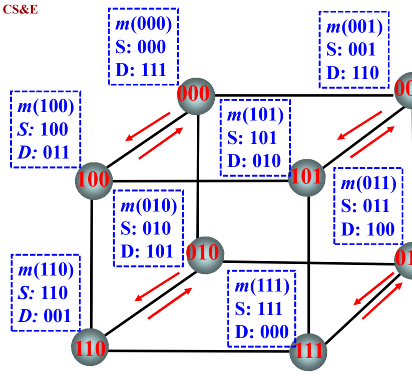
$m(i)$
S: $i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_n$
C: $j_1,\ldots,j_{k-1},j_k,i_{k+1},\ldots,i_n$
D: $j_1,\ldots,j_{k-1},j_k,j_{k+1},\ldots,j_n$



$j_1,\ldots,j_{k-1},i_k,i_{k+1},\ldots,i_n$ $j_1,\ldots,j_{k-1},j_k,i_{k+1},\ldots,i_n$

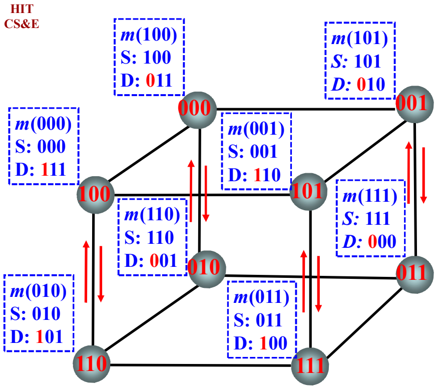
HIT CS&E

运行实例



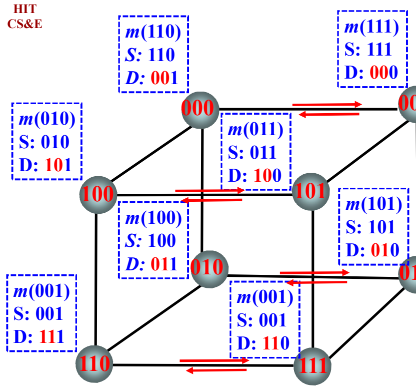
第1步

HIT CS&E



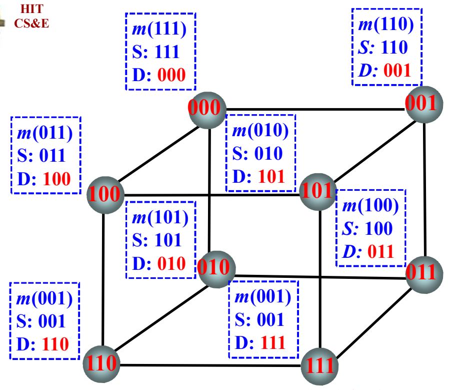
第2步

HIT CS&E



第3步

HIT CS&E



3步足亦！

在其他实例上也很高效吗？

HIT CS&E

Bit Fixing的缺陷

Bit Fixing算法
- 确定型路由算法
- 遗忘性路由算法

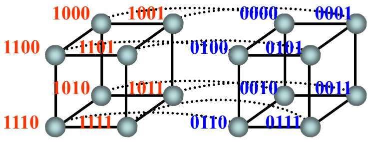
确定型遗忘性路由算法的时间复杂度下界
- N-顶点网络拓扑，每个顶点的度为n (不必是超方体)
- 排列型路由任务($d(1),\dots,d(N)$是$1,\dots,N$的排列)
- [*Kaklamanis, Krizanc, Tsantilas, SPAA'90*]
  确定型遗忘性路由算法在特殊实例上需要$\Omega(N^{1/2}/n^{1/2})$步

Bit Fixing算法的下界更糟糕
- Bit Fixing算法在如下实例上需要$\Omega(N^{1/2})$步

HIT CS&E

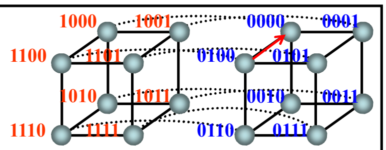
Bit Fixing下界实例

Bit Fixing算法的下界实例族
- $n=2t$
- 顶点i的标记为$(i_1,\dots,i_t,i_{t+1},\dots,i_n)$
- $d(i)$的标记为$(i_{t+1},\dots,i_n,i_1,\dots,i_t)$



d(0000)=0000
d(0001)=0100
d(0010)=1000
d(0011)=1100
d(0100)=0001
d(0101)=0101
d(0110)=1001
d(0111)=1101
d(1000)=0010
d(1001)=0110
d(1010)=1010
d(1011)=1110
d(1100)=0011
d(1101)=0111
d(1110)=1011
d(1111)=1111

HIT CS&E



考虑边e上通过的所有消息
- $e=(0\dots01\ 0\dots00)\rightarrow(0\dots00\ 0\dots00)$
- 消息$m(?\dots?1\ 0\dots00)$的目标地址$d(i)=(0\dots00\ ?\dots?1)$
- 用Bit Fixing算法路由时
  消息$m(?\dots?1\ 0\dots00)$到达$d(i)$前必须通过边e
- e上至少要通过2^{t-1}个消息
- 由t的任意性知
  Bit Fixing存在需要$\Omega(N^{1/2})$步的实例族

HIT CS&E

用随机性打破时间复杂度下界

如何在路由算法中引入随机性呢？

HIT CS&E

随机路由算法

输入：超方体拓扑结构，消息$m(i)$及目的地址$d(i)$

输出：$(\forall i)$将$m(i)$从顶点i路由到$d(i)$

算法：0. $(\forall i)$均匀随机地产生中间地址$t(i)\in[N]$
   //$t(0),t(1),\dots,t(N-1)$不必是$0,1,\dots,N-1$的排列
  1. 调用Bit Fixing将$(\forall i)m(i)$从i路由到$t(i)$
   //第1步结束后才进入第2步
  2.调用Bit Fixing将$(\forall i)m(i)$从$t(i)$路由到$d(i)$

数据结构：
- $(\forall i)$顶点i维护n个队列$Q_{i1}, Q_{i2},\dots, Q_{in}$
- Q_{ik}缓存来自i的第k个邻居的消息
- 顶点的消息队列执行FIFO策略
- 同时到达的消息可按任意顺序进入队列
- 每步每个队列可达到一个消息，可发出一个消息

随机路由算法性能

定理：随机路由算法仅用$8n=O(\log N)$步在n-超方体上完成任意排列型路由任务的概率至少为$1-2/N^2$

注：
- $8n$远远小于Bit Fixing算法的时间复杂度下界
- 即使$d(i)=i$，随机路由算法仍需要$O(\log N)$步

性能分析（关键引理）

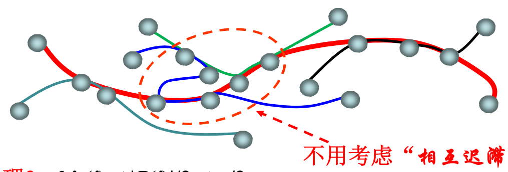
记号：

$P(i)$— 消息$m(i)$从i路由到$t(i)$所经过的路径　　$|P(i)| \le n$

$\text{hit}(i) = |\{ j \mid j\neq i$ 且 $P(j)$与$P(i)$有公共边 $\}|$

引理1：消息$m(i)$从i路由到$t(i)$至多需要$|P(i)|+\text{hit}(i)$步

直观含义：$m(i)$被每个“同路消息”$m(j)$至多迟滞一步



不用考虑“相互迟滞”关系？

引理2：$\text{hit}(i) \le |P(i)|/2 \le n/2$

由引理得定理

$$\text{hit}(i,j) = \begin{cases} 1 & \text{若}P(j)\text{与}P(i)\text{有公共边} \\ 0 & \text{否则} \end{cases}$$

$t(0), t(1),\ldots, t(N-1)$是均匀随机地选取的

$\text{hit}(i, 0)$，$\text{hit}(i, 1)$，…，$\text{hit}(i, N-1)$是独立的泊松实验

$\text{hit}(i) = \text{hit}(i, 0) + \text{hit}(i, 1) +\ldots+ \text{hit}(i, N-1)$

由Chernoff界，对于任意满足$1+\delta>2e$的δ均有

$$\Pr[\text{hit}(i) > (1+\delta) E[\text{hit}(i)]] < \left(\frac{1}{2}\right)^{(1+\delta)E[\text{hit}(i)]}$$

$|P(i)| \le n$, $E[\text{hit}(i)] \le n/2$ (引理2). 取δ使得$(1+\delta) E[\text{hit}(i)]=3n$

$\delta > 5$

$$\Pr[|P(i)|+\text{hit}(i) > 4n] \le \Pr[\text{hit}(i) > 3n] < \left(\frac{1}{2}\right)^{3n}$$

$\Pr[|P(i)|+\text{hit}(i) > 4n] < 1/2^{3n}$　　$\forall i=0,1,\ldots,2^n-1$

$m(i): i\to t(i)$　　对称性　　$m(i): t(i)\to d(i)$

$\Pr[m(i)$的路由步数 $> 8n]$

$\le \Pr[m(i)$：$i\to t(i)$步数 $> 4n$ 或 $m(i)$：$t(i)\to d(i)$步数 $> 4n]$

$\le \Pr[m(i)$：$i\to t(i)$步数 $> 4n] + \Pr[m(i)$：$t(i)\to d(i)$步数 $> 4n]$

$\le 2\cdot\Pr[m(i)$：$i\to t(i)$步数 $> 4n]$

$< 2/2^{3n}$

$\Pr[$算法路由步数 $> 8n]$ $=\Pr[\exists i$: $m(i)$路由步数 $> 8n]$

$\le 2^n \cdot \Pr[m(1)$路由步数 $> 8n]$

$< 2/2^{2n}$

$= 2/N^2$

引理1：消息$m(i)$从i路由到$t(i)$至多需要$|P(i)|+\text{hit}(i)$步

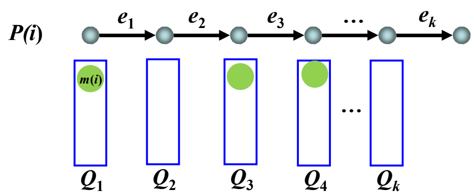
引理2：$\text{hit}(i) \le |P(i)|/2 \le n/2$

由引理1和引理2，可以证得

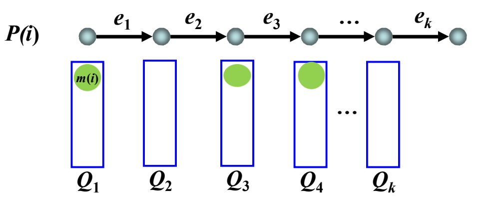
定理：随机路由算法仅用$8n=O(\log N)$步在n-超方体上完成任意排列型路由任务的概率至少为$1-2/N^2$

证引理1

$P(i)$上各顶点的消息队列
- 仅列出各个顶点与路径对应的队列$Q_1,\ldots,Q_k$
- 初始时，各个队列可能有一个消息待转发



HIT CS&E



只要$m(i)$在某个队列中滞留，必有$m(j)$（$j\in$hit(i)）离开$P(i)$

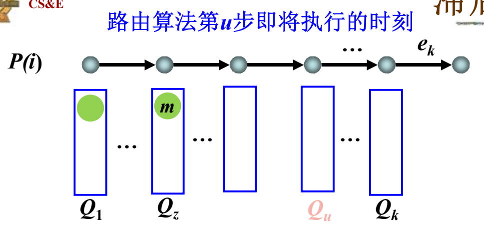
HIT CS&E

论断：只要$m(i)$在某个队列中滞留
必有$m(j)$（$j\in$hit(i)）离开$P(i)$

注意：只要上述论断成立，则引理1成立
$m(i)$至多被滞留 hit(i)次

HIT CS&E

滞后指数



- 对于任意消息m(u表示算法第u个同步即将开始的时刻)
- 在时刻u，m位于在$P(i)$路径的第z个消息队列Q_z
- ($1\le z\le |P(i)|=k$)
- 消息的滞后指数定义为 lag(m) = u-z
- m离开$P(i)$之后，lag(m)无定义
- m进入$P(i)$前，lag(m)无定义

HIT CS&E

滞后指数变化规律

从时刻u到时刻u+1

- m滞留于同一个队列，则lag(m)增大1
- m从Q_z进入Q_{z+1}，　　则lag(m)不变
- m首次进入$P(i)$，　　则lag(m)被定义
- m离开$P(i)$，　　则lag(m)无定义

HIT CS&E

论断证明思路

论断：只要$m(i)$在某个队列中滞留
必有$m(j)$（$j\in$hit(i)）离开$P(i)$

lag($m(i)$)	hit(i)
0	
…	
0→1 -------------------	j_0
…	
1→2 ------- 单射 -------	j_1
…	
g→g+1 -------------------	j_g
…	

max lag($m(i)$) $\le$ hit(i)

HIT CS&E

lag($m(i)$): $g\to g+1$

$time(g)$ = min u: 在u时刻开始时$P(i)$上仍存在消息m
使得lag(m)=g

算法执行第$time(g)$个同步步骤时：

- 必有消息m (lag(m)=g) 通过$P(i)$的一条边e (***)
- 消息m必然离开$P(i)$，否则在time(g)准备执行时仍有lag(m)=g，这与time(g)的定义矛盾
- 任取一个离开$P(i)$的消息$m(j_g)$

由遗忘性路由的性质，离开$P(i)$的消息不会再回到$P(i)$

$g\to j_g$是单射

HIT CS&E

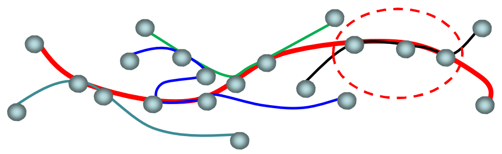
证引理2

记号Route(e)

- e是n-超方体中的任意一条边
- $m(k)$是顶点k要送出的消息, $k\in\{0,1,\ldots,N-1\}$
- $P(k)$是$m(k)$从k路由到中间顶点$t(k)$经过的路径

- Route(e) = | {$P(k)$ | $e\in P(k)$} |

$\text{hit}(i) \le \sum_{e\in P(i)} \text{Route}(e)$



HIT CS&E

E[Route(e)] = E[Route(e')]

$$\text{Route}(e,j) = \begin{cases} 1 & e\in P(j) \\ 0 & 否则 \end{cases}$$

Route(e) = Route(e,0)+ Route(e,1)+ …+ Route(e,N-1)

Route(e')= Route(e',0)+ Route(e',1)+ …+ Route(e',N-1)

n-超方体是对称结构

- 存在顶点的排列π将e映射为e' 并且
- e与0,1,2,…,N-1的关系 正好与 e'与0,1,…,N-1的关系对应

E[Route(e)] = E[Route(e')]

HIT CS&E

E[Route(e)] = 1/2

E[Route(e)] = E[Route(e')]= μ $\forall e,e'$

n-超方体有$n2^{n-1}$条边

$P(i)$路径平均长度

- $t(i)$ 是均匀随机选取的
- $|P(i)|$= 0, 1,2,…,n的概率分别为$\binom{n}{k}/2^n$
- E[$|P(i)|$] = $n/2$

$n2^{n-1}\mu$ = E[$|P(0)|$] + E[$|P(1)|$+…+ E[$|P(2^n-1)|$]] = $n2^{n-1}/2$

$\mu = 1/2$

HIT CS&E

得到引理2

E[Route(e)] = ½

$\text{hit}(i) \le \sum_{e\in P(i)} \text{Route}(e)$

$$\text{E[hit}(i)] \le \sum_{e\in P(i)} \text{E[Route}(e)]$$

$$\le |P(i)|/2$$

HIT CS&E

结论

确定型遗忘性路由算法在某些实例上需要很高的复杂性

引入随机性，借助Chernoff界，可以证明

随机路由算法以很高概率确保路由过程仅需线性步骤

HIT CS&E

补充结论

Maurer不等式

定理：$X_1,\ldots,X_n$是独立的非负随机变量且$E[X_i^2]<\infty$

令$X=\sum_i X_i$, 则对任意$t>0$有

$$\Pr[\ E[X] - X \ge t\ ] < exp\{-\frac{t^2}{2\sum_i E[X_i^2]}\}$$

[1]Maurer. A. A Bound on the Deviation Probabilities for Sums of non-negative Random Variables. Journal of Inequalities in Pure and Applied Mathematics, 4, 2003.

Bernstein不等式

定理：$X_1,\ldots,X_n$是独立随机变量且X_i-E[X_i]≤M对任意i成立

σ_i^2=E^2[X_i]-E[X_i^2].令 $X=\sum_i X_i$ 则对任意$t>0$有

$$\Pr[\ X \ge E[X]+t\ ] < exp\{-\frac{t^2}{2\sum_i \sigma_i^2 + 2Mt/3}\}$$

[2]Bernstain. S. Theory of Probability. Moscow, 1927