New Directions in Nearest Neighbor Searching with Applications to Lattice Sieving

Anja Becker, Léo Ducas, Nicolas Gama, Thijs Laarhoven

背景介绍

- 最近邻搜索(NNS)问题：给定N个n维向量集合，通过预处理，使得寻找某个向量在集合中的最近邻向量时间在$O(N^\rho)(\rho < 1)$
- 近似NNS：距离给定向量$\boldsymbol{v}$最近的向量距离r_1(角度θ_1)，其余向量大于距离r_2(角度θ_2)
- 高维球面：$\mathcal{S}^{n-1} := \{\boldsymbol{x} \in \mathbb{R}^n, ||\boldsymbol{x}|| = 1\}$

局部敏感哈希(LSH)

- 局部敏感哈希(LSH)：hash函数族$\mathcal{H}$。
  - 其中的hash函数满足：相近的向量大概率hash值相同
- 预处理：取t组k个hash函数$h_{i,j} \in \mathcal{H}$，对每个向量$\boldsymbol{w}$计算t组$h_i(\boldsymbol{w}) = (h_{i,1}(\boldsymbol{w}), \cdots, h_{i,k}(\boldsymbol{w}))$，生成t个hash table
- 查找：对于向量$\boldsymbol{v}$，将$h_i(\boldsymbol{v})$在hash table中对应的所有$\boldsymbol{w}$纳入$\boldsymbol{v}$的近邻向量候选
- 定义hash函数的碰撞概率：

$$p(\theta) := \Pr_{h \sim \mathcal{H}}[h(\boldsymbol{v}) = h(\boldsymbol{w}) | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \cos\theta]$$

- 若向量角度大于θ_2，希望hash值相等的概率低：调整k

$$\begin{aligned} P_2 &= \Pr[h_i(\boldsymbol{v}) = h_i(\boldsymbol{w}) | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \langle \boldsymbol{v}, \boldsymbol{w} \rangle \leq \cos\theta_2] \\ &= \prod_{j=1}^{k} \Pr[h_{i,j}(\boldsymbol{v}) = h_{i,j}(\boldsymbol{w}) | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \langle \boldsymbol{v}, \boldsymbol{w} \rangle \leq \cos\theta_2] \leq p(\theta_2)^k \leq \frac{1}{N} \end{aligned}$$

  - $k \geq \frac{\log N}{-\log p(\theta_2)}$

- 若向量角度小于θ_1，希望至少要有一组相同的hash值：调整t

$$P_1 = \Pr[h_i(\boldsymbol{v}) = h_i(\boldsymbol{w}) | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \langle \boldsymbol{v}, \boldsymbol{w} \rangle \geq \cos\theta_1] \geq p(\theta_1)^k \geq N^{-\frac{\log p(\theta_1)}{\log p(\theta_2)}}$$

  - $t \cdot P_1 \geq 1$，即$t \geq \frac{1}{P_1} = N^{\frac{\log p(\theta_1)}{\log p(\theta_2)}}$

- 时间复杂度：计算$t \cdot k$次hash，$\widetilde{O}(N^{\frac{\log p(\theta_1)}{\log p(\theta_2)}})$

局部敏感过滤(LSF)

- 局部敏感过滤(LSF)：过滤函数族$\mathcal{F}$
- 预处理：取t组k个过滤函数$f_{i,j} \in \mathcal{F}$，对每个向量$\boldsymbol{w}$进行t次组合过滤$f_i(\boldsymbol{w}) = (f_{i,1}(\boldsymbol{w}), \cdots, f_{i,k}(\boldsymbol{w}))$，生成t个过滤后的向量集合$(L_1, \cdots, L_t)$
- 查找：对于向量$\boldsymbol{v}$，将能通过的$f_i(\boldsymbol{v})$对应的过滤集合中对应的所有$\boldsymbol{w}$纳入$\boldsymbol{v}$的近邻向量候选
- 定义过滤函数的碰撞概率：

$$p(\theta) := \Pr_{f\sim\mathcal{F}}[\boldsymbol{v}, \boldsymbol{w} \in L_f | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \cos\theta]$$

  - $p(0)$：随机向量通过过滤函数f的概率$\Pr[\boldsymbol{v} \in L_f | \boldsymbol{v} \in \mathcal{S}^{n-1}]$
  - $t \cdot p(0)^k$：随机向量能通过组合过滤的数量（向量能进到几个组）

- 若角度大于θ_2，希望最终被分在一个组中的概率低：调整k

$$P_2 = \Pr[\boldsymbol{w} \in L_i | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \boldsymbol{v} \in L_i, \langle \boldsymbol{v}, \boldsymbol{w} \rangle \leq \cos\theta_2]$$
$$= \prod_{j=1}^{k} \frac{\Pr[\boldsymbol{v}, \boldsymbol{w} \in L_{i,j} | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \langle \boldsymbol{v}, \boldsymbol{w} \rangle \leq \cos\theta_2]}{\Pr[\boldsymbol{v} \in L_{i,j} | \boldsymbol{v} \in \mathcal{S}^{n-1}]} \leq \left(\frac{p(\theta_2)}{p(0)}\right)^k \leq \frac{1}{N}$$

  - $k \geq \frac{\log N}{\log p(0) - \log p(\theta_2)}$

- 若角度小于θ_1，希望最终被分在一个组中：调整t

$$P_1 = \Pr[\boldsymbol{w} \in L_i | \boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}^{n-1}, \boldsymbol{v} \in L_i, \langle \boldsymbol{v}, \boldsymbol{w} \rangle \geq \cos\theta_1] \geq \left(\frac{p(\theta_1)}{p(0)}\right)^k \geq N^{-\frac{\log p(0) - \log p(\theta_1)}{\log p(0) - \log p(\theta_2)}}$$

  - $t \cdot p(0)^k \cdot P_1 \geq 1$，即$t \geq \frac{1}{p(0)^k \cdot P_1} = N^{\frac{-\log p(\theta_1)}{\log p(0) - \log p(\theta_2)}}$

- 时间复杂度：计算$t \cdot p(0)^k$次过滤函数，$\widetilde{O}(N^{\frac{\log p(0) - \log p(\theta_1)}{\log p(0) - \log p(\theta_2)}})$

LSH和LSF实例

- Spherical LSH：单个hash函数通过采样$U = 2^{\Theta(\sqrt{n})}$个单位向量$\boldsymbol{s_1}, \boldsymbol{s_2}, \cdots, \boldsymbol{s_U}$，取$\alpha = n^{-\frac{1}{4}}$，构造

$$H_{\boldsymbol{s}_i} := \mathcal{C}_{\boldsymbol{v},\alpha} \backslash \bigcup_{j=1}^{i-1} H_{\boldsymbol{s}_j} \qquad p(\theta) = \exp\left[-\frac{\sqrt{n}}{2}\tan^2\left(\frac{\theta}{2}\right)(1+o(1))\right]$$

$$\rho_{LSH} = \frac{\log p(\theta_1)}{\log p(\theta_2)} = \frac{\tan^2(\theta_1/2)}{\tan^2(\theta_2/2)}(1+o(1))$$

- Spherical LSF：单个过滤函数通过一个随机单位向量$\boldsymbol{s}$，和一个角度α，构造

$$F_{\boldsymbol{s}} := \mathcal{C}_{\boldsymbol{v},\alpha} \qquad p(\theta) = \exp\left[\frac{n}{2}\ln\left(1 - \frac{2\alpha^2}{1+\cos\theta}\right)(1+o(1))\right]$$

$$\rho_{LSF} = \frac{\log(1-\alpha^2) - \log\left(1 - \frac{2\alpha^2}{1+\cos\theta_1}\right)}{\log(1-\alpha^2) - \log\left(1 - \frac{2\alpha^2}{1+\cos\theta_2}\right)}(1+o(1))$$

α的取值

$$\rho_{LSF} \overset{\alpha=0}{\sim} \frac{\log p(\theta_1)}{\log p(\theta_2)} = \frac{\tan^2(\theta_1/2)}{\tan^2(\theta_2/2)} = \rho_{LSH}$$

- 当$\alpha = 0$时，从理论时间复杂度来看，Spherical LSF和Spherical LSH相同

$$k = \frac{\log N}{\log p(0) - \log p(\theta_2)} \geq 1 \Rightarrow \alpha \leq \alpha_0 := \sqrt{1 + \frac{N^{2/n}(\cos\theta_2 - 1)}{2N^{2/n} - \cos\theta_2 - 1}}$$

- 为了使ρ小，α尽可能大，当$k = 1$时，α取到最大值

BDGL构造

随机乘积编码(Random product codes)：使用直积进行Filter的构造

$$C = Q \cdot (C_1 \times C_2 \times \cdots \times C_m)$$

- Q代表$\mathbb{R}^n$上的均匀随机旋转，$n = m \cdot b$，$C_i \subset \sqrt{\frac{1}{m}}\mathcal{S}^{n-1}$，$C_i = \{\boldsymbol{c_{i,1}}, \boldsymbol{c_{i,2}}, \cdots, \boldsymbol{c_{i,B}}\}$
- $m = O(\log n)$ 时理论最优
- 可以使用$B \cdot m$个长度为b的随机向量表示$M = B^m$个长度为n的随机向量
- 等价组合过滤器数量：$t = B^m$

解码：对于一个向量$\boldsymbol{w} \in \mathcal{S}^{n-1}$，寻找所有$\boldsymbol{s} = (\boldsymbol{c}_{1,j_1}|\boldsymbol{c}_{2,j_2}|\cdots|\boldsymbol{c}_{m,j_m})$，满足$\langle \boldsymbol{w}, Q\boldsymbol{s}\rangle \geq \alpha$

算法：点乘可以简单的拆分

- 计算$\boldsymbol{v} = Q^{-1}\boldsymbol{w}$，并拆分$\boldsymbol{v} = (\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_m)$
- 对所有$\boldsymbol{v}_i$，计算$\langle \boldsymbol{v}_i, \boldsymbol{c}_{i,j}\rangle$，并排序
- 深度有限搜索，通过改变阈值剪枝

一次查询时间复杂度：

$$\mathcal{T}_{LD}(t,\alpha) = O(nB + mB\log B + mt\mathcal{C}_n(\alpha))$$

符号介绍

- 高维球面：$\mathcal{S}^{n-1} := \{\boldsymbol{x} \in \mathbb{R}^n, ||\boldsymbol{x}|| = 1\}$
- 高维圆锥：$\mathcal{H}_{\boldsymbol{v},\alpha} := \{\boldsymbol{x} \in \mathbb{R}^n, \langle \boldsymbol{v}, \boldsymbol{x} \rangle \geq \alpha\}$
- 高维球冠：$\mathcal{C}_{\boldsymbol{v},\alpha} := \mathcal{S}^{n-1} \cap \mathcal{H}_{\boldsymbol{v},\alpha}$

$$\mathcal{C}_n(\alpha) = \frac{\mu(\mathcal{C}_{\boldsymbol{v},\alpha})}{\mu(\mathcal{S}^{n-1})} = \mathrm{poly}(n) \cdot (1 - \alpha^2)^{\frac{n}{2}}$$

- 高维球面楔：$\mathcal{W}_{\boldsymbol{v},\alpha,\boldsymbol{w},\beta} := \mathcal{S}^{n-1} \cap \mathcal{H}_{\boldsymbol{v},\alpha} \cap \mathcal{H}_{\boldsymbol{w},\beta}$，令 $\cos\theta = \langle \boldsymbol{v}, \boldsymbol{w} \rangle$

$$\mathcal{W}_n(\alpha, \beta, \theta) = \frac{\mu(\mathcal{W}_{\boldsymbol{v},\alpha,\boldsymbol{w},\beta})}{\mu(\mathcal{S}^{n-1})} = \mathrm{poly}(n) \cdot \left(1 - \frac{\alpha^2 + \beta^2 - 2\alpha\beta\cos\theta}{\sin^2\theta}\right)^{\frac{n}{2}}$$

- GeoGebra ↗

应用到筛法中

- 目标：找到所有角度小于$\pi/3$的向量
- 询问α：α越大，角度越小，查询代价越小
- 插入(预处理)β：β越大，角度越小，需要构建更多的过滤器以保证查询成功率
- 期望过滤器数量：$t = \tilde{O}(1/\mathcal{W}_n(\alpha, \beta, \pi/3))$
- 询问期望能通过过滤器的数量：$t \cdot \mathcal{C}_n(\alpha)$
- 每个过滤器中的期望向量数量：$N \cdot \mathcal{C}_n(\beta)$

向量数量：$N = \left(\frac{4}{3}\right)^{\frac{n}{2}+o(n)}$

查询时间复杂度：

$$\mathcal{T}_1 = \tilde{\mathcal{O}}\left(\frac{N \cdot \mathcal{C}_n(\alpha) \cdot (1 + N \cdot \mathcal{C}_n(\beta))}{\mathcal{W}_n(\alpha, \beta, \pi/3)}\right) = \tilde{\mathcal{O}}\left(\left(\frac{4(1-\alpha^2)}{3 - 4(\alpha^2 + \beta^2 - \alpha\beta)}\right)^{\frac{n}{2}} \left[1 + \left(\frac{4(1-\beta^2)}{3}\right)^{\frac{n}{2}}\right]\right)$$

预处理时间复杂度：

$$\mathcal{T}_2 = \tilde{\mathcal{O}}\left(\frac{N \cdot \mathcal{C}_n(\beta)}{\mathcal{W}_n(\alpha, \beta, \pi/3)}\right) = \tilde{\mathcal{O}}\left(\left(\frac{4(1-\beta^2)}{3 - 4(\alpha^2 + \beta^2 - \alpha\beta)}\right)^{\frac{n}{2}}\right)$$

空间复杂度：

$$\mathcal{S} = \tilde{\mathcal{O}}\left(N + \frac{N \cdot \mathcal{C}_n(\beta)}{\mathcal{W}_n(\alpha, \beta, \pi/3)}\right) = \tilde{\mathcal{O}}\left(\left(\frac{4}{3}\right)^{\frac{n}{2}} + \left(\frac{4(1-\beta^2)}{3 - 4(\alpha^2 + \beta^2 - \alpha\beta)}\right)^{\frac{n}{2}}\right)$$

- $(\alpha,\beta)=(\frac{1}{2},\frac{1}{2})$

$$\mathcal{T}=(3/2)^{n/2+o(n)}\approx 2^{0.292n+o(n)}$$

$$\mathcal{S}=(3/2)^{n/2+o(n)}\approx 2^{0.292n+o(n)}$$

- $(\alpha,\beta)=(\frac{1}{4},\frac{1}{2})$

$$\mathcal{T}=(5/3)^{n/2+o(n)}\approx 2^{0.368n+o(n)}$$

$$\mathcal{S}=(4/3)^{n/2+o(n)}\approx 2^{0.208n+o(n)}$$

左下角的点只能使用NV筛