表 3.6 l，k与lf_p的对应关系

l	k ≤	$l \times 0.25^k$
20	8.8	9.53E-05
30	10.5	1.34E-05
40	11.8	3.23E-06
50	12.7	1.06E-06
60	13.5	4.21E-07
80	14.8	9.77E-08
100	15.8	3.13E-08

虽然使用 LSH 不能保证找到解决方案，但找到解决方案的概率较高。这个概率可以通过增加 LSH 函数的数量来提升。

在此基础上，我们引入离散空间上的广义局部敏感哈希（Generalized Locality Sensitive Hashing）的定义。假设有一个在 d 维空间$\mathbb{R}^d$上，包含 n 个数据点$p = (p_1, \dots, p_d)$的集合 P。对于任意两点 p 和 q，它们之间的距离定义为

$$\left|\left|p - q\right|\right|_S = \left(\sum_{i=1}^{d} |p_i - q_i|^S\right)^{\frac{1}{S}} \tag{3.28}$$

对于任意的 S>0，这个距离函数被称为标准化l_S。如果$\left|\left|p - q\right|\right|_S \leq$ R，则称 p 是 q 的 R 近邻（R-near neighbor）。如图 3.3，p_1，p_2，p_3都是 q 的 R 近邻，p_4则不是。

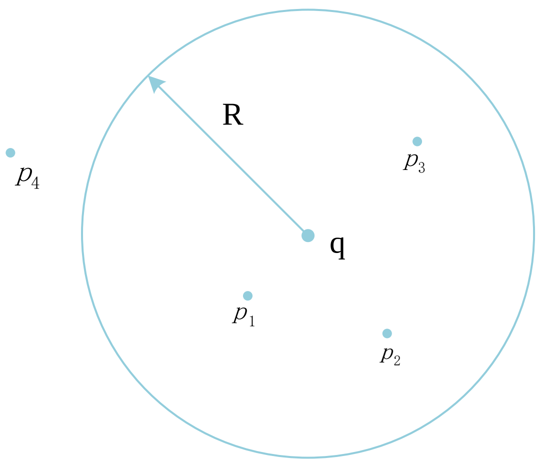

图 3.3 离散空间上的R近邻

从几何学的角度理解 LSH，本质就是一种投影。首先根据散列函数进行散列操作，这个函数在二维空间上可以表示为一条直线，使得空间上相邻的点投影在这条直线的同一段区间里，即散列到同一个桶中。这种 LSH 和连续函数 mod k 的散列方式是有本质区别的，无法进行精准的归类，让所有相邻的点都被散列在一个桶中，而不相邻的点也

无法保证一定不在同一个桶中。但是 LSH 在一定程度上可以区分查询点的相近点和较远点。

上述最近邻 (near neighbor, NN) 问题可以扩展到近似最近邻(Approximate near neighbor, ANN) 问题[22] ，也被称为 Randomized c-approximate R-near neighbor ((c,R)-NN)。

给定点集$P \subseteq \mathbb{R}^d$，查询点$q \in \mathbb{R}^d$，查询范围 R>0 和近似因子 c>1，(c, r)-NN 问题的输出如下：若存在$p \in P$，满足$\|p-q\|_s \le R$，则输出某点$p' \in P$，满足$\|p'-q\|_s \le cR$。

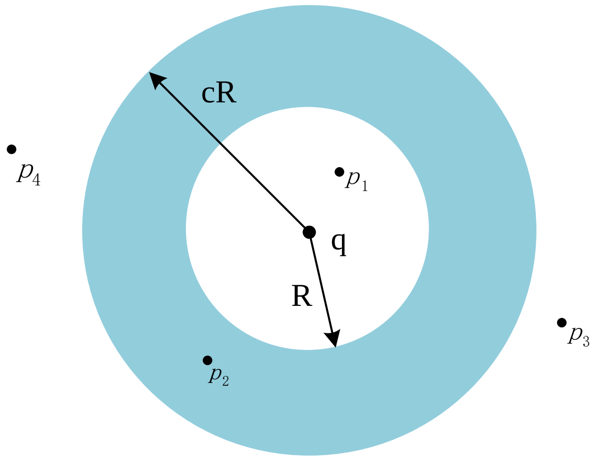

图 3.4　离散空间上的ANN

我们需要让在相近的 R 近邻（图 3.4 中白色部分）进行 hash 冲突的概率更大，而 cR 近邻（图 3.4 中蓝色部分）则越小。但是原始的 LSH 则会使得 R 近邻和 cR 近邻都尽可能的大，会导致查找的结果很多时候并不能得到一个较好的回馈。

构造一个数据结构 LSH，对于任意查询$q \in \mathbb{R}^d$，如果在 P 中存在 q 的 R 近邻，能以 1-δ (δ>0) 的概率给出 P 中 q 的 cR 近邻。跟连续空间上的 LSH 相似，如果 LSH 函数的规模扩大一倍，这个概率将从 1-δ上升到 $1-\delta^2$。当我们对离散空间进行放缩，使 R=1 时，ANN 问题就变成了 c 近似最近邻问题 (c-approximate near neighbor problem, c-NN) 。

由此，我们可以正式地定义局部敏感哈希。如果对于任意两点p, q，从 H 中均匀随机选择一个函数 g，有

① 如果$\|p-q\| \le R$，$\Pr\big(g(p)=g(q)\big) \ge P_1$；

② 如果$\|p-q\| \ge cR$，$\Pr\big(g(p)=g(q)\big) \le P_2$；

③ $0 < P_2 < P_1 < 1$，

则称哈希函数族 H 是局部敏感的，也称 Gapped LSH。P_1和P_2间的距离可能非常小，需要额外添加步骤放大两者的间隙。

我们选择l个函数$h_1, h_2, \ldots, h_l$，其中$h_{\mathrm{j}}(q) = \left(g_{\mathrm{j1}}(q), \ldots, g_{\mathrm{jk}}(q)\right)$，而$g_{\mathrm{ji}}$是从 H 中均匀随机选择的函数（$1 \leq \mathrm{j} \leq l,\ 1 \leq \mathrm{i} \leq k$）。使用这些函数将 P 转换到$l$个散列表中，之后对于每一组询问$q$，查找这些散列表以提取数据点进行验证。如果$\left||p-q|\right| \leq R$，$\Pr\left(h_{\mathrm{j}}(p) = h_{\mathrm{j}}(q)\right) \geq {P_1}^k$；如果$\left||p-q|\right| \geq cR$，$\Pr\left(h_{\mathrm{j}}(p) = h_{\mathrm{j}}(q)\right) \leq {P_2}^k$。

对于一个 c-NN 问题，我们先找到 $\mathrm{L} = tl$ 个点验证，然后中止。定义事件 E1 为假阳性的数量少于 $\mathrm{L} = tl$ 个数据点（即总哈希冲突次数$< tl$），E2 为以 1-θ的概率找到一个解决方案。我们需要根据给定的t和θ，找到最佳参数l和k。

$\left||p-q|\right| \geq cR$碰撞的概率$\Pr\left(h_{\mathrm{j}}(p) = h_{\mathrm{j}}(q)\right) \leq {P_2}^k = \frac{1}{n}$（n 表示数据点的数量），因此可以推导得$k = -\frac{\ln n}{\ln P_2}$。

q 在散列表 a 中发生冲突的期望 $\mathrm{E}(\#\mathrm{collisions\ with\ q\ in\ a\ table}) \leq 1$，所以$l$个散列表中 q 发生冲突的期望 $\mathrm{E}(\mathrm{total\ \#collisions\ with\ q\ in}\ l\ \mathrm{tables}) \leq l$。根据马尔可夫不等式 (Markov Inequality) 可知，Y 为仅假设在非负值上的随机变量，对于任意$\mathrm{t} \in \mathbb{R}^+$，都有$\Pr(\mathrm{Y} \geq t) \leq \frac{E(Y)}{t}$。因此，

$$\Pr(\mathrm{total\ \ \#collisions} \geq tl) \leq \frac{l}{tl} = \frac{1}{t} \tag{3.29}$$

$$\Pr(< tl\ \ \mathrm{collisions}) \geq 1 - \frac{1}{t} \tag{3.30}$$

如果存在 NN（即$\left||p-q|\right| \leq R$），找到一个 ANN 的概率为

$$\begin{aligned}&\Pr\left(h_1(p) = h_1(q) \vee \ldots \vee h_l(p) = h_l(q)\right)\\&= 1 - \Pr\left(h_1(p) \neq h_1(q) \wedge \ldots \wedge h_l(p) \neq h_l(q)\right)\\&\geq 1 - \left(1 - {P_1}^k\right)^l \approx 1 - e^{-l{P_1}^k} = 1 - \theta\end{aligned} \tag{3.31}$$

在$\theta = e^{-l{P_1}^k}$，${P_2}^k = \frac{1}{n}$的条件下，

$$l = -\ln\theta / {P_1}^k = -\ln\theta / n^{-\frac{\ln P_1}{\ln P_2}} = -\ln\theta \times n^{\frac{\ln P_2}{\ln P_1}} = \mathrm{O}(n^\rho)(\rho = \frac{\ln P_2}{\ln P_1} < 1) \tag{3.32}$$

根据上述分析，可知

$$\Pr(E1 \cap E2) \geq 1 - (1 - \Pr(E1)) - (1 - \Pr(E2)) = 1 - (\frac{1}{t} + \theta) \tag{3.33}$$

令$\delta = (\frac{1}{t} + \theta)$，事件 E1 和 E2 同时为真的概率就是$\Pr(E1 \cap E2) \geq 1 - \delta$。$\Pr(E1 \cap E2) = 1 - \delta$当且仅当所有假阳性情况都被找到，$\Pr(E1) = 1$。算法空间复杂度$\mathrm{O}(dn + nl) = \mathrm{O}(dn + n^{1+\rho})$，搜索部分的时间复杂度$\mathrm{O}(dl) = \mathrm{O}(dn^\rho)$。

对于一个 R 近邻问题，我们在l个散列表中搜索询问串 q 的哈希值，合并所有产生哈希冲突的项，对它们进行验证。关于假阳性和敏感度部分的分析与 c-NN 相同，空间复杂度也是一样的。而在搜索部分的时间复杂度上，需要额外考虑解个数的期望，$\mathrm{E}(\#\text{false positives} + \#\text{occurrences of solutions}) = \mathrm{O}(dl) + \mathrm{O}\left(d \times {P_1}^k \times l \times \#solutions\right)$。由于${P_1}^k \times l = n^{-\rho} \times n^{\rho} = 1$，所以化简得$\mathrm{O}(dn^{\rho} + d \times \#solutions)$。

对于一个长度为 d 的子串，

① 如果$\left||p-q|\right| \le R$，$\Pr\left(g(p) = g(q)\right) \ge \frac{d-r}{d} = P_1$；

② 如果$\left||p-q|\right| \ge cR$，$\Pr\left(g(p) = g(q)\right) \le \frac{d-cr}{d} = P_2$；

所以$\rho = \frac{\ln P_2}{\ln P_1} = \frac{\ln 1-\frac{r}{d}}{\ln 1-\frac{cr}{d}} \le \frac{1}{c}$，$l = \mathrm{O}(n^{\rho}) = \mathrm{O}\left(n^{\frac{1}{c}}\right)$。

3.3.2 Order Min Hash 模型

设 H 是定义在集合$\mathcal{U}$（全集）上的散列函数族。当

$$\mathrm{s}(\mathrm{x},\mathrm{y}) \ge s_1 \Rightarrow Pr_{\mathfrak{h}\in\mathcal{H}}[\mathfrak{h}(x) = \mathfrak{h}(y)] \ge p_1, \tag{3.34}$$

$$\mathrm{s}(\mathrm{x},\mathrm{y}) \le s_2 \Rightarrow Pr_{\mathfrak{h}\in\mathcal{H}}[\mathfrak{h}(x) = \mathfrak{h}(y)] \le p_2, \tag{3.35}$$

则称集合$\mathcal{H}$上的概率分布对相似度$\mathscr{s}$关于(s_1, s_2, p_1, p_2)敏感的。其中$s_1 \ge s_2$，$p_1 \ge p_2$。如果存在一组散列函数的分布关于(s_1, s_2, p_1, p_2)敏感，则允许使用 Gapped LSH 对相似序列进行聚类。在上面的定义中，具体概率取决于$\mathcal{H}$中对任意 x, y∈ $\mathcal{U}$构造的哈希函数的选择。在有间隙的 LSH 中，相似元素之间哈希冲突的概率上升($\ge p_1$)，而对于不同元素，哈希冲突的概率较小($\le p_2$)。

编辑相似度上的 LSH 要求必须对字符串的 k-mer 内容和相对顺序都敏感，但对于 k-mers 在字符串中的绝对位置相对不敏感。这引出了下面的定义。与 minHash 类似，k-mers 是通过在 k-mers 上使用置换来随机选择的。此外，为了保留有关的相对顺序的信息，ℓ 个 k-mers 被随即选中，并按照它们在序列中出现的顺序（而不是随机置换所定义的顺序）记录。

此外，该方法必须处理重复的 k-mers。同一 k-mer 的两个副本出现在序列中的不同位置，区分这两个副本对于 k-mer 之间的相对顺序很重要。我们通过在 k-mers 后附加“出现次数”，使其唯一。

更准确地说，对于长度$|\mathrm{S}| = \mathrm{n}$的字符串 S，考虑 k-mers 对及其出现次数的集合$\mathcal{M}_k^w(S)$。如果序列 S 中有 x 个 m 的副本，那么集合$\mathcal{M}_k^w(S)$中有 x 对元素，形如