3.4文档的位置敏感散列

但是，观察两列越相似，它们在某些频段中相同的可能性就越大。 因此，直观地说，条带策略使得相似的列比不相似的对更可能成为候选对。

3.4.2行条化策略的分析

假定使用b个行条，每个行条由r行组成，并假定某对具体文档之间的Jacc缸d相似度为s。3.3.3节的内容表明，文档的最小哈希签名矩阵中某个具体行中的两个签名相等的概率等于s。接下来我们可以计算这些文档（或其签名）作为候选对的概率，具体计算过程如下：

（1）在某个具体行条中所有行的两个签名相等的概率是sr.

（2）在某个具体行条中至少有一对签名不相等的概率是1 – sr.

（3）在任何行条中的任意一行的签名对都不相等的概率为(1 − sr)b.

（4） 签名至少在一个行条中全部相等的概率，也即成为候选对的概率为1 − (1 − sr)b。

虽然有可能并不特别明显，但是不论常数b和r的取值如何，上述形式的概率函数图像大致为图3-7所给出的S－曲线（S-curve）。 曲线中候选概率1/2处对应的相似度就是所谓的阅值 ( threshold ），它是b和r的函数。 阔值对应的大概是上升最陡峭的地方，对于较大的b和r，相似度在阔值之上的对很可能成为候选对，而在阐值之下的对则不太可能成为候选对，这正是我们想要的结果。 阔值的一个近似估计是(1/b)1/r。例如，如果b=16且r=4， 那么由于16的4次方根为2，阔值的近似值为1 / 2。

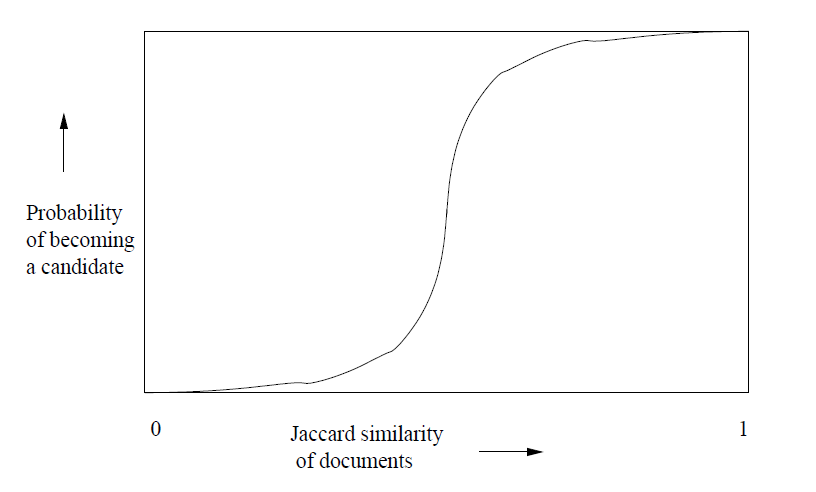


图3-7 S-曲线

例3.11 考虑b=20且严5的情况， 也就是说假定签名的个数为100， 分成20个行条， 每个行条包含5行。 图3-8以表格形式给出了函数1-(1-s5)20的部分值。注意到的是，这里的阐值， 也就是曲线中部上升处的S值，仅仅比0.5稍大一点。另外也注意到，该曲线并非从0到l在阔值处跳跃 的最理想步进函数， 但是曲线中部的斜率十分显著。例如，S从0.4变到0.6， 增加的函数值大于O札因此中间部分的斜率大于3o

又例如，s=0.8时， 1-(0.8)5大约为0.672。 如果再求20次方得到大约0.000 35， 用l减去该值以后得0.999 65。 也就是说， 如果认为两篇文档的相似度为80%， 那么在任意行条中， 5行中签名对全部相等的可能性只有约33%，因而它们会成为候选对。然而，这里有20个行条，因此有20次机会成为一个候选对。3000个对中，大致仅有1个相似度为80%的对不会成为候选对，即成为伪反例。

图3-8 b=20且严5条件下S－曲线的值

3.4.3 上述技术的综合

本节将给出一个完整的相似项发现方法：首先找出可能的候选对相似文档集合，然后基于该集合发现真正的相似文档。必须强调的是，这种方法可能会产生伪反例，即某些相似文档对由于没有进入候选对所以最终没有被识别出来。同样，该方法也可能会产生伪正例， 即在评估了某些候选对后，发现其相似度不 足。

1. 选择某个k，并对每篇文档构建其k-shingle集合。将这些k-shingle映射成更短的桶编号 （ 后一步可选）。
2. (2） 将文档－shinglexf按照shingle排序。
3. (d丘） 选择最小哈希签名的长度n。将（2）中排好序的表传递给3.3.5节中的算法来计算所有文档的最小晗希签名。
4. (4）选择阁值t来定义应该达到的相似程度使之被看做是预期的 “相似对”。选择行条数b和每个行条中的行数r， 使得b严n， 而阔值d丘似等于(1/b)llr。如果避免伪反例的产生很重要，那么选 择合适的b和r以产生小于t的阔值。 而如果速度相当重要并且希望限制伪正例的数目，那么选择合适的b和F来获得更高的阔值。
5. (5） 应用3.4.1节中的LSH技术来构建候选对。
6. (6） 检查每个候选对的签名，确定 它们 一致性的比例是否大于fa
7. (7) （该步可选）如果签名足够相似，则直接检查文档本身看它们是否真正 相似。 不相似的文档有时碰巧会具有相似的签名。

3.4.4 习题

习题3.4.1 对s=0.1, 0.2，…，0.9，在下列r、 b取值情况下求S－曲线上l一（ 1-s'l的值：

r=3，b=10;

r=6,，b=20;

r=5,，b=50;

习题3.4.2 对习题3.4.1中的每个（r，的对，计算阁值，即使得1一（1-s'l 等于1/2日如的值，并 将该值与3.4.2节（lib）＂＇的估计值进行比较。

习题3.4.3 利用1.3.5节介绍的技术来计算当s＇非常小时5曲线1 (1 sγ的近似值。

习题3.4.4 假定我们希望采用MapReduce方式来 实现LSH。具体地，这里假设签名矩阵的组块由列构成，每个元素为键4直对，其中键是列号，而值则是签名本身 （也就说，这里的值实际上是一个值向量）。

(a）说明如何通过单个MapReduce过程来生成所有行条的桶。提示：记住Map函数能够基于单个元素生成多个键4直对；

(b）说明如何通过另一个MapReduce过程来将（a）的输出结果转换成需要比较的对列表。具体地，对于每列i，应该存在一个要与i比较且满足与. > i的列组成的列表。

3.5 距离测度

接下来我们稍做偏离一点主题来研究距离测度的一般性概念。尽管Jaccard相似度可以度量两个集合的相似程度，但是它并不是一个真正意义上 的距离测度。也就是说，集合越接近，Jaccard相似度却越大而不是像距离一样应该越短。 与之相反 ，正如我们将看到的那样 ，l减去Jaecard相似度却是一个距离测度， 该测度称为Jaccard距离（Jaccard distance ）。

然而，Jaecard距离并不是唯一有意义的能够度量相近程度的测度。接下来我们要介绍一些在实际中使用的其他距离测度。然后，在3.6节我们会介绍在有些距离测度上也可以应用LSH技术， 从而只允许我们关注相近的点而不用比较所有点。在第7章讨论聚类时我们还会介绍距离测度的其他应用。

3.5.1 距离测度的定义

假定有一些点组成的集合，我们称这个集合为空间（space ）。该空间下的距离测度（similarity measure）是一个函数d巾，y），以空间中的两个点作为参数，输出是一个实数值。该函数必须满足下列准则：

（1）d(x,y） 二三 0（距离非负）；

（2）当且仅当 x =y 时， d(x,y) = 0 （只有点到自身的距离为0，其他距离都大于0 );

(3) d(x, y) = d(y, x) （距离具有对称性）；

(4) d(x,y） 运 d(x, z) + d(z, y) （三角不等式）。

三角不等式是上述条件中最复杂的条件。它的直观意义是，如果从x点行进到y点，那么如果一定要求经过某个特定的第三点z则不会有任何好处。三角不等式准则使得所有的距离测度表现 得如同其描述的是从一个点到另一个点的最短路径的长度。

3.5.2 欧氏距离

欧氏距离是最为人熟知的距离测度，也就是我们通常所想象的 “距离”。在n维欧氏空间中， 每个点是一个n维实数向量。 该空间中的传统距离测度，即我们常说的L2范式（Li-norm）定义如下：

d([x1，码，…，Xn ],[yi ,Y2 ,"',Yn ]) = Jf,(x，，一九)2

也就是说，首先计算每一维上的距离，然后求它们的平方和，最后求算术平方根。

对于上述距离公式，很容易验证它满足上述距离四准则中的前三条。由于上面计算的是算术平方根，因此两点之间的欧氏距离不可能是负数。由于所有实数的平方非负，所以如果存在某个i满足Xi并Yi，那么整个计算结果就严格大于0。而另一方面，只有当所有i都满足XFY，时，距离才为O。由于（xi一Yi = (yi一Xi)，所以对称性准则显然满足。至于欧式距离是否满足三角不等式准则，则需要较多的代数学知识才能证明。但是欧氏空间有一个众所周知的特性，即一个三角形的两边之和不小于第三边。

欧氏空间下还有一些其他的距离测度方法。对于任意常数r, L，范式的定义如下：

当r=2时，就是刚才提到的L2范式距离。另一个常用的距离测度是L1范式距离或者称为曼哈顿距离（ Manhattan distance ）。即两个点的距离是每维距离的绝对值之和。之所以称为 “曼哈顿 距离”，是因为这里在两个点之间行进时必须要沿着网格线前进，就如同沿着城市（如曼哈顿） 的街道行进一样。

另一个有趣的距离测度是L范式，也就是当F趋向于无穷大时L 范式的极限值。当r增大时，只有那个具有最大距离的维度才真 正起作用，因此，正式来讲，ι范式定义为在所有维度t下Ix厂Y;I中的最大值。

例3.12 考虑二维欧氏空间（即通常所说的平面） 上的两个点（2, 7）和（6, 4）。它们的L2范式

距离为」（2-6)2 +(7 -4)2 =14可)2 =5, Li范式距离为12-61 + 17-41 = 4 + 3 = 7，而L”范式距离为max(l2-6J, J7-41) = max(4, 3) = 4。

3.5.3 Jaccard距离

正如本节一开始提到的那样，集合的Jaecard距离可以定义为d(x,y) = 1-SIM(x, y），也就是说

Jae card距离等于l减去x、y的交集与并集的比率。我们必须要验证该函数是一个距离测度。

（1）由于交集的大小不可能大于并集的大小，因此d(x,y）不可能为负值。

（2）若x = y则d巾， y) = O，这是因为xUx =xnx = x。然而，如果x乒y，那么 xny的大小严格小于xUy的大小，因此，d(x,y）严格为 正。

（3）由于交集和并集运算都是对称的，即xUy= yUxfllx 门y=ynx，因此d(x,y) = d(y, x）。

(4）至于三角不等式，回顾一下3.3.3节我们就知道，SIM巾， y）是一个随机最小哈希函数将x和 y映射为相同值的概率。因此，Jaccard距离d巾，y）为一个随机最小晗希函数将x和y映射为不同值的概率。所以，三角不等式条件d(x,y） 运 d(x,z) + d(z, y）可以变换为命题：如果h是一个随机的最小 哈希函数，那么 h(x) -:ft h(y）的概率不高于h(x)-:ft h(z）的概率与h(z) f; h(y）的概率之和。然而，因为只要有h(x) f; h(y） 那么至少h(x）和h(y）中的一个一定与h(z）不同。即它们俩不可能都是h(z），否则两者显然相等。因此上述命题为真。

3.5.4 余弦距离

在有维度的空间下余弦距离（consine distance）才有意义，这些空间包括欧式空间及离散欧式空间，而后者包括坐标只采用整数值或布尔值（0或1）来表示的空间。在上述空间下，点可以代表方向。后者包括坐标只采用整数值或布尔值（0或1）来表示的空间。在上述空间下，点可以代表方向。这里我们并不区分一个向量及其多倍向量①。因此，两个点的余弦距离实际上是点所代

表的向量之间的夹角。不管空间有多少维，该夹角的范围是0 ～ 180度。

于是，我们首先计算夹角的余弦，然后应用反余弦函数（ arc-cosine ）将结果转化为0 ～ 180度之间的角度，从而最终得到余弦距离。给定向量x和y，其夹角余弦等于它们的内积 x.y 除以两个向量的L2范式（即它们到原点的欧式距离）乘积。记得向量的内积[x1, x2,…..,xn].[]y1,y2,……yn]为

侣。3.13 两个向量分别为x= [l, 2，一1］及.y= [2, 1, I］，则内积 x.y= 1 x 2+ 2×1 + ( I）×I = 3o

两个向量的L2范式均为品，比如x的L2范式为􀀐I2+22 +(-1)2 ＝品。因此，x和y的夹角余弦为

3/( J6 J6) =1/2，而余弦值为1/2的角大小为60度。因此，x和y的余弦距离为60度。

我们必须证明余弦距离也是一个距离测度。由于余弦距离定义在0到180之间，因此，余弦距

离非负。当且仅当两个向量表示同一方向时向量的夹角为0。＇ll余弦距离的对称性非常明显：x和y 1111111

…

的夹角显然与y和x的夹角相等。至于三角不等式则能够通过物理含义来最好地诠释，如要将向量

x旋转到y，可以先从x旋转型􀀟z，然后再从z旋转到y。两次旋转经过的夹角之和不会小于直接旋转

所得到的夹角。

3.5.5 编辑距离

编