第 5 章　信息检索的基础数学原理

由于当今信息量呈几何级数膨胀和用户信息需求多样化发展趋势，在检索的实践活动中会涉及大量的信息处理与存储过程。用户信息检索的最终实现必须依靠强有力的计算机应用程序去自动执行或智能信息处理作为支撑，而强有力的计算机应用程序必须依据数学原理及其模型方法的建立为前提，利用数学原理与模型方法来建立检索基础模型是必不可少的工作。运用数学原理不仅能使信息检索作为研究对象的概念含义精确化，而且能够深刻揭示信息检索过程的显性现象与潜在的隐性规律。在信息检索中引入数学原理及其模型方法，将检索过程中的信息及其处理过程加以解释和抽象，表达成某种数学模型，再经演绎与推断，从而指导检索实践和促进检索工作的技术进步。数学原理及其模型的引入使得信息检索有了更加严谨的论证，检索过程和信息需求本质的描述也更为精确。迄今为止，基于集合理论的布尔模型、Salton 模型和模糊集合模型等数学一般原理最为成熟，也在检索实践中得到了普遍应用。

5.1　简单布尔检索

5.1.1　基本原理

布尔模型是一种以经典集合论和布尔代数为理论基础的非常简单的信息检索模型。它采用布尔代数的方法，用布尔逻辑表达式表示用户需求提问，通过对信息标识和提问式的比较来检索信息。对某一特定的信息，通常表示成 $D=(t_1,t_2,\cdots,t_n)$ 的形式。由于布尔逻辑式可以表达成与用户思维习惯相一致的提问要求，因此，用户提问可以表示为由三种逻辑运算符即逻辑与(*)、逻辑或(+)和逻辑非(−)连接起来的布尔表达式，标引词 t_1 和 t_2 之间可能具有的逻辑运算是 $t_1 \wedge t_2$ 和 $t_1 \vee t_2$，而任一标引词的逻辑非运算为 $-t$，这些逻辑运算将作为用户提问的一部分出现在布尔表达式的某个位置上，图 5-1 可以很直观地显示这些逻辑运算的结果。

显然，上述的布尔运算实际上是集合之间的交、并、补运算。也就是说，布尔检索实际上是通过若干个检索词所包含的信息集合的交、并、补运算来响应用户信息需求提问的。

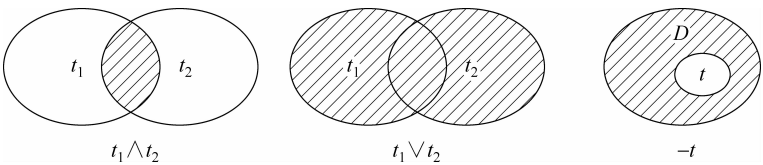

图 5-1　布尔运算逻辑关系图

布尔模型在解释信息检索的数据处理过程时，主要遵循两条基本规则。

系统索引词集合中的每一个索引词在一篇文档中只有两种状态：出现或者不出现。相应地，每个索引词的权值 $w_{ij} \in \{0,1\}$。

检索提问式 $\boldsymbol{q}$ 由三种布尔逻辑运算符“and”、“or”、“not”连接索引词来构成。

根据布尔逻辑的运算规定，提问式 $\boldsymbol{q}$ 可以被表示成由合取子项（conjunctive components）组成的析取范式（disjunctive normal form，dnf 或 DNF）形式。例如，布尔提问式

$$\boldsymbol{q} = k_1 \text{ and}(k_2 \text{ or not } k_3)$$

可以写成如下等价的析取范式形式：

$$\boldsymbol{q}_{\text{dnf}} = (k_1 \text{ and } k_2 \text{ and } k_3)\text{or}(k_1 \text{ and } k_2 \text{ and not } k_3)\text{or}(k_1 \text{ and not } k_2 \text{ and not } k_3)$$

这里，$\boldsymbol{q}_{\text{dnf}}$ 为提问式 q 的主析取范式。进一步地，可以用如下简化形式来表示 $\boldsymbol{q}_{\text{dnf}}$：

$$\boldsymbol{q}_{\text{dnf}} = (1,1,1) \text{ or } (1,1,0) \text{ or } (1,0,0)$$

其中，(1,1,1)、(1,1,0)和(1,0,0)是 $\boldsymbol{q}_{\text{dnf}}$ 的三个合取子项（合取子项可用符号 $\boldsymbol{q}_{\text{cc}}$ 表示），它们是一组向量，由对应三元组 (k_1,k_2,k_3) 的每一分量取 0 或 1 值而得到。

基于上述规则与假定，布尔模型对于任一篇文档 $\boldsymbol{d}_j \in D$，定义 $\boldsymbol{d}_j$ 与用户提问 $\boldsymbol{q}$ 的匹配函数为

$$\text{sim}(\boldsymbol{d}_i,\boldsymbol{q}) = \begin{cases} 1, \text{如果存在 } \boldsymbol{q}_{\text{cc}} \mid (\boldsymbol{q}_{\text{cc}} \in \boldsymbol{q}_{\text{dnf}}) \text{ 且对于任意 } k_i, \text{有 } g_i(\boldsymbol{d}_j) = g_i(\boldsymbol{q}_{\text{cc}}) \\ 1, \text{其他} \end{cases} \tag{5-1}$$

式(5-1)中，函数 g_i 定义为 $g_i(\boldsymbol{d}_j) = w_{ij}$。现在，假设文档集合 D 中存在两篇文档 $\boldsymbol{d}_1$ 和 $\boldsymbol{d}_2$，其中，$\boldsymbol{d}_1$ 含有索引词 k_1 和 k_2，$\boldsymbol{d}_2$ 含有索引词 k_1 和 k_3，则它们的文档向量分别为

$$\boldsymbol{d}_1 = (1,1,0)$$

$$\boldsymbol{d}_2 = (1,0,1)$$

根据匹配函数 $\text{sim}(\boldsymbol{d}_j,\boldsymbol{q})$ 的定义，很显然文档 $\boldsymbol{d}_1$ 与提问式 $\boldsymbol{q} = k_1 \text{and}(k_2 \text{ or not } k_3)$ 的匹配函数值为 1，即文档 $\boldsymbol{d}_1$ 与提问 $\boldsymbol{q}$ 是相关的；而文档 $\boldsymbol{d}_2$ 与提问 $\boldsymbol{q}$ 的匹配函数值为 0，表

明文档 d_2 与提问 q 是不相关的。

5.1.2　布尔检索模型的特点

布尔模型是最早提出的一种信息检索一般数学模型。1957 年，巴·希列尔(Y. Bar-Hille)就对布尔逻辑应用于计算机信息检索的可能性进行了探讨；20 世纪 60 年代末期，布尔检索模型正式被大型文献检索系统所采用；70 年代时逐渐成为各种商业性联机检索服务系统的标准检索模式。目前，基于布尔检索框架的各类检索系统仍具有顽强的生命力，并在信息搜索与信息服务领域占据重要地位。

在布尔检索中，用户的查询要求用普通的语言叙述，即用户可完全按照自己的思维习惯提问。其中查询要求(条件)A、B、C、D 等可以分别用若干个标引词来表示，然后可以用布尔逻辑运算符“∨”、“∧”、“—”将用户的提问“解析”成信息服务系统可以接受的形式。这种结构化的提问方式与用户的思维习惯相一致，所以成为布尔逻辑检索的一个突出优点。布尔检索的一个用户界面实例如图 5-2 所示。

图 5-2　布尔检索实例图(以 ProQuest 为例)

以 ProQuest 为例，图 5-2 布尔检索实例图中的“并且”、“或者”与“排除”运算，就是典型的布尔检索应用。这种模型把复杂的检索过程简单化，能够将比较复杂的信息提问按其概念组配的逻辑关系描述出来，从而变成可以由计算机执行的逻辑运算，变成机器根据

事先确定的程序进行自动匹配的过程，这种运算上的简单易行是布尔逻辑检索系统的突出优势。

布尔模型具有简单性(simplicity)、容易理解性(easy understanding)、简洁形式化(clean formalism)等突出优点。布尔模型的简单性、易理解性与易实现等特点为其在检索系统和检索工具中的广泛应用奠定了良好基础。尽管布尔模型有着种种优点，但它还是存在明显的局限性。

(1) 布尔模型是基于二值判定为标准的，信息对象要么相关，要么不相关，并没有一个相关信息级别的概念，例如符合信息需要的相关性程度大小，因此很难有好的检索效果。

(2) 构造布尔逻辑式不是一件轻松的事情，对于普通信息用户，很难用 AND(逻辑与)、OR(逻辑或)、NOT(逻辑非)运算的结合来准确地表达自己的信息需求，并且检索词的简单组配也不能完全反映实际需要。

(3) 检索结果输出完全依赖于布尔提问与检索系统中信息的匹配情况，很难控制输出量的大小。

(4) 布尔提问表示存在某些不合理的地方。对于“∨”提问，包含一个在提问中出现的检索词的信息与包含几个在提问中出现的标引词的信息被认为是一样的重要；对于“∧”提问，包含多个标引词的信息与不包含任何标引词的信息被看成是一样不相关。

(5) 检索结果不能按用户定义的重要性排序输出，用户只能从头到尾浏览输出结果才能知道哪些信息更适合自己的需要。

鉴于布尔模型的这些不足，人们提出用语词加权和部分匹配的功能来扩展经典的布尔模型，将向量模型和布尔模型融为一体，来克服传统布尔模型的一些缺陷，这就是扩展布尔模型。

5.2 信息检索模糊集合论

信息检索模糊集合模型是建立在模糊集合论基础上的，模糊集合论可以看做是经典集合论的推广。1965 年美国加州大学伯克利分校的札德(LA. Zadeh)教授发表了一篇关于“模糊集合”的著名论文，由此奠定了模糊理论的研究与发展。

模糊集合论对经典集合论的推广主要表现在：它把元素属于集合的概念模糊化，承认集合论范围内存在既不完全属于某集合，又不完全不属于某集合的元素，即变经典集合论“绝对的属于”概念为“相对的属于”概念；同时，又进一步把属于概念数量化，承认论域

上的不同元素对于同一集合具有不同的隶属程度，因此引入了隶属度(membership)的概念。

模糊集合理论处理的是边界不明确的集合表示，其中心思想是把集合中的元素和隶属函数结合在一起。隶属函数的取值在[0,1]上，0 表示元素不隶属于该集合，1 表示完全隶属于该集合，值在 0 和 1 之间表示元素为该集合的边际元素。

定义：给定论域 U，U 的模糊子集 A 可以定义为 U 到闭区间[0,1]上的一个映射：LA：$U \rightarrow [0,1]$，LA 为 A 的隶属度。正如经典集合论是传统精确数学的基础一样，模糊子集论是模糊理论的基础，同样也可以定义模糊子集上的运算。常见的三种运算分别是模糊集合的补运算、两个或多个集合的并、交运算。

定义：给定论域 U，A 和 B 分别为 U 的两个模糊子集，$A-$是 A 关于 U 的补集，u 为 U 中的元素，则

$$\mu_{\overline{A}}(u) = 1 - \mu_{\overline{A}}(u)$$

$$\mu_A \bigcup B(u) = \max(\mu_A(u) \cdot \mu_B(u))$$

$$\mu_A \bigcap B(u) = \min(\mu_A(u) \cdot \mu_B(u))$$

5.2.1　模糊检索的数学描述

模糊检索是将信息文档看成是与提问在一定程度上相关，对于每一个标引词，都存在一个模糊的信息集合与之相关；对于某一给定的标引词，用隶属函数表示每一则信息文档与该词相关的程度，即隶属度，其取值在[0,1]上，则有信息文档 d 和标引词 t，d 对于 t 的隶属度可以定义为

$$\mu_F : D \times T \rightarrow [0,1],$$

$$(d,t) \rightarrow \mu_F(d,t)\ \forall (d,t) \in D \times T$$

则在信息检索系统中文档 d 与标引词 t 的二元模糊关系 F 可以描述为

$$F = \{[(d,t), \mu_F(d,t)] d \in D, t \in T\} \tag{5-2}$$

由于用户通常希望检索出的信息能较高地满足其需求主题，因此，这里所定义的 $\mu_F(d,t)$表示文献 d 涉及标引词 t 所达到的程度，而不是标引词 t 反映文献 d 的主题内容的程度。

标引词的模糊集合是在标引过程中建立的，标引人员不是简单地把标引词赋予信息文档，还要指出标引词与信息文档的相关程度。如 $d=\{(t_1,0.5),(t_2,0.8)\}$，数字 0.5 和 0.8 表示信息文档对于标引词 t_1，t_2 的隶属度，数值越大表示隶属度越大。当全部信息文

档标引完毕，也就为每个标引词定义了一种隶属函数，指明了每一信息文档对于每个标引词的相关程度。

隶属函数是模糊集合论乃至整个模糊学的最基本概念之一，正确构造隶属函数是应用模糊学方法的关键。由于隶属度的确定，既有客观性的一面，也有主观性的一面，因此，在解决实际问题时，构造切合实际的隶属函数至今还没有非常满意的解决方法。

5.2.2 信息文档对标引词的隶属度

在标引词集合中，由于概念相关的模糊性，两个标引词在不同程度上总是存在着语义上的关联，因此，信息文档对标引词的隶属度是通过标引词表来计算的。标引词表可以通过词-词关联矩阵来建立，这个矩阵的行和列分别对应于集合中的标引词，矩阵中词 t_i 和 t_j 的关联因子可以定义为

$$C_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \tag{5-3}$$

式中 n_i 表示包含标引词 t_i 的信息文档的数目，n_j 表示包含标引词 t_j 的信息文档的数目，则标引词 t 的模糊集合中，文献 d 的隶属度：

$$\mu_F = 1 - \Pi(1 - C_{i,j}) \tag{5-4}$$

5.2.3 提问检索词的相关性描述

用户提问通常是由布尔逻辑式表达的，即用布尔逻辑运算符将标引词连接起来。布尔逻辑的常用运算符有“与”、“或”、“非”，即 $\wedge$，$\vee$，$\rightarrow$。提问匹配以通过引入模糊算符来确定信息文档对于提问的相关程度。设 D 为信息文档集，Q 为提问集，$\forall d \in D, q \in Q$，$Q \times D$ 上的模糊关系 R：

$$R = \{(q,d,\mu(q,d)) \mid q \in Q, d \in D\}$$

式中 $\mu(q,d)$ 表示信息文档 d 对于提问 q 的相关程度。

根据模糊集合的运算规则，将三个基本的模糊运算符分别定义如下。

(1) 若 $q=a \vee b$，则 $\mu(q,d)=\max(\mu(d,a),\mu(d,b))$，这里 $a,b \in T$，$\mu(d,a)$，$\mu(d,b)$ 分别表示信息文档 d 论述标引词 a 和 b 所达到的程度。

(2) 若 $q=a \wedge b$，则 $\mu(q,d)=\min(\mu(d,a),\mu(d,b))$。

(3) 若 $q=-a$，则 $\mu(q,d)=1-\mu(d,a)$。

在模糊集合检索中，对于布尔模型的用户信息需求的处理通常是把表达用户需求的布尔逻辑式转换成析取范式的形式。例如，$q=t_a \wedge (t_b \vee -t_c)$，可以写成与之等价的析取

范式：$q_{\text{dnf}}=(1,1,1)\vee(1,1,0)\vee(1,0,0)$，其中的每个分量都是$(t_a,t_b,t_c)$的一个二值加权向量，它们构成了 $\bar{q}_{\text{dnf}}$的合取分量，用 CC_i 表示第 i 个合取分量，则提问可以推广为 p 个合取分量的形式：

$$\bar{q}_{\text{dnf}} = \text{CC}_1 \vee \text{CC}_2 \vee \cdots \vee \text{CC}_p \tag{5-5}$$

计算信息文档与提问相关的过程类似于经典布尔模型中的计算，只不过在模糊检索中处理的对象是模糊集合而不是普通的集合。

对于上述的提问 $q=t_a\wedge(t_a\vee t_b)$，D_a 表示标引词 t_a 在文献集上的模糊子集，它由隶属度大于既定阈值的文献所组成。同理，可以定义标引词 t_a 和 t_c 的模糊子集 D_b、D_c，由于所有的集合都是模糊不确定的，即使信息文档 d 不包括标引词 t_a，该信息文档也有可能属于集合 D_a（见图 5-3）。

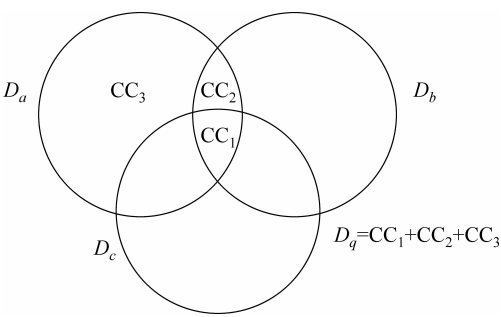

图 5-3　提问 $q=t_a\wedge(t_b\vee t_c)$的模糊文献集

提问模糊集合 D_q 是 q_{dnf}的三个合取分量的模糊集合的并运算，则 D_q 中信息文档 d 的隶属度：

$$\begin{aligned}\mu(q,d) &= \mu\text{cc}_1+\text{cc}_2+\text{cc}_3\cdot d = 1-\prod_{i=1}^{3}(1-\mu\text{cc}_i\cdot d)\\ &=1-\{\mu(d,a)\mu(d,b)\mu(d,c)\}\times\{1-\mu(d,a)\mu(d,b)(1-\mu(d,c))\}\\ &\quad\times\{1-\mu(d,a)(1-\mu(d,b))(1-\mu(d,c))\}\end{aligned}$$

计算得出 $\mu(q,d)$，它所反映的正是信息文档 d 对于提问 q 的相关程度。所以，提问 q 可以定义为信息文档集合 D 上的一个模糊子集：$q=\{(d,\mu(q.d))\mid d\in D\}$。用户给定一个阈值 $\lambda(0\leqslant\lambda\leqslant1)$，将小于 λ 的项去掉。当 $\mu(q,d)\geqslant\lambda$ 时，d 作为命中的信息文档输出，输出可以采取按照对提问的相关程度的大小形式排序输出。通过控制 λ 的取值，可以输出合适的文献。

基于模糊集合模型的检索结果是建立在信息文档集上的，且其隶属度就是信息文档集对用户提问的相关程度的模糊子集。就目前的水平而言，还无法十分精确、有效地确定这个隶属函数：在提问匹配中引入的 max 和 min 算符不能很好地反映真实的匹配过程，而把提问的布尔逻辑表达式转换成析取范式，用代数和、代数积分计算析取模糊集合以获取模糊集合中信息文档的隶属度，更加适合于模糊信息检索应用。

模糊检索模型与经典布尔模型关系密切，它基本保留了布尔检索功能，但是更为灵活，对那些既想利用布尔检索长处，又想避免其二值相关性测度局限性的人们来说，能够较好地满足信息检索需求。模糊检索模型还支持对命中文档按相关度大小的排序输出。

5.3 扩展布尔检索

1983 年信息检索专家萨尔顿(G. Salton)及其博士生福克斯(E. A. Fox)等人提出的一种基于布尔逻辑框架的混合布尔与向量特性的混合检索模型，即扩展布尔模型。扩展的布尔检索模型是基于布尔逻辑基本假设的改进，下面采用矢量的方法来讨论布尔信息检索。

5.3.1 基于两个标引词的情形

假定信息文档集合中的信息 $\boldsymbol{d}_j$ 仅用两个标引词 t_x 和 t_y 标引，并且 t_x，t_y 允许被赋予一定的权值，其权值分别为 $W_{x,j}$、$W_{y,j}$，权值的取值范围为[0,1]，权值越接近于 1，说明该词越能反映文本的内容，反之，反映文本的内容较差。给标引词加权通常采用的是著名的 tf-idf 加权方案：

$$W_{x,j} = f_{x,j} \times \frac{\mathrm{idf}_x}{\max x_i \times \mathrm{idf}_x} \tag{5-6}$$

式中 $f_{x,j}$ 为标引词 t_x 在文献 $\boldsymbol{d}_j$ 中出现的频率，idf_x 为逆信息文档词频。为了简单起见，用 x，y 分别表示权值 $W_{x,j}$、$W_{y,j}$。我们采用二维图来表示信息文档的提问，用距离的概念表示信息文档与提问的相似度。见图 5-4。

对于析取提问 $\boldsymbol{q}=t_x \vee t_y$，只有 A、B、C 三点所代表的信息文档才是最理想的，对于任一信息文档 $\boldsymbol{D}_j$ 而言，当它离 A、B、C 三点越接近时，说明相似度越大，因而 $\boldsymbol{D}_j$ 到点(0,0)的矢量距离可以用来度量与提问 $\boldsymbol{q}_{or}$ 的相似度，则

$$|\boldsymbol{D}_j| = \overline{x^2 + y^2} \tag{5-7}$$

显然，$0 \leqslant |\boldsymbol{D}_j| \leqslant 1$，为了使相似度控制在 0 和 1 之间，相似度可以规范化为

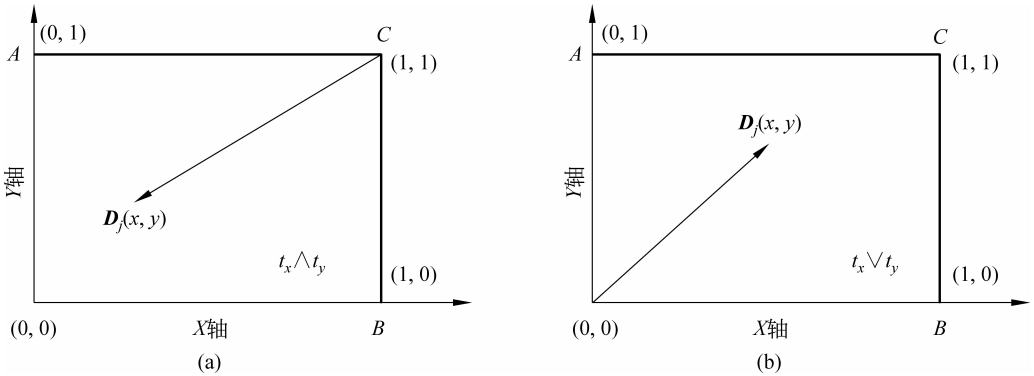

图 5-4　扩展布尔逻辑的矢量表示

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}}, \boldsymbol{d}_j) = \frac{\overline{x^2 + y^2}}{2} \tag{5-8}$$

对于合取提问 $\boldsymbol{q} = t_x \wedge t_y$，只有 C 点才是最理想的文献，则 $\boldsymbol{D}_j$ 到 C 点的矢量距离为

$$| \boldsymbol{D}_j | = \overline{(1 - x^2) + (1 - y^2)} \tag{5-9}$$

它可以作为衡量文献与提问之间相似度的一个尺度，则相似度可以规范化为

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}}, \boldsymbol{d}_j) = 1 - \frac{\overline{(1 - x^2) + (1 - y^2)}}{2} \tag{5-10}$$

5.3.2　推广到 n 个标引词空间

以上讨论的是两个标引词的情况，信息文档集合中的标引词的数目为 n 时，模型可以推广到 n 维空间的欧几里得距离。根据线性向量模型理论，广义的析取提问和合取提问可以分别表示为

$$\boldsymbol{q}_{\mathrm{or}} = t_1 \vee^p t_2 \vee^p \cdots \vee^p t_n$$

$$\boldsymbol{q}_{\mathrm{and}} = t_1 \vee^p t_2 \vee^p \cdots \vee^p t_n$$

这里，p 是一个可变的量，$1 \leqslant p \leqslant \infty$ 的值在提问时就应当确定。则这两种文献-提问的相似度为

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}}, \boldsymbol{d}_j) = \left[\frac{x_1^p + x_2^p + \cdots + x_n^p}{n}\right]^{\frac{1}{p}}$$

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{and}}, \boldsymbol{d}_j) = 1 - \left[\frac{(1 - x_1)^p + (1 - x)^p + \cdots + (1 - x_n)^p}{n}\right]^{\frac{1}{p}}$$

式中的 x_i 表示信息文档 $\boldsymbol{d}_j$ 中的第 i 个标引词的权值 $W_{i,j}$。由于 p 是一个变量，下面分析 p 的取值对相似度的影响。

(1) 当 $p=1$ 时，

$$\begin{aligned}\mathrm{sim}(\boldsymbol{q}_{\mathrm{and}},\boldsymbol{d}_j) &= 1-\frac{n-(x_1+x_2+\cdots+x_n)}{n}=\frac{x_1+x_2+\cdots+x_n}{n}\\ &=\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j)\end{aligned} \tag{5-11}$$

则布尔逻辑表达式中的布尔逻辑运算符“∧”、“∨”已毫无区别，两者的功能都减退为0，相似度的计算采取简单的向量空间模型余弦函数法，即

$$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q})=\frac{\overline{\boldsymbol{d}_j}\cdot\boldsymbol{q}}{|\overline{\boldsymbol{d}_j}|\times|\boldsymbol{q}|}=\frac{\sum_{i=1}^{t}W_{i,j}\times W_{i,q}}{\sqrt{\sum_{i=1}^{t}(W_{i,j})^2}\times\sqrt{\sum_{i=1}^{t}(W_{i,q})^2}} \tag{5-12}$$

(2) 当 $p=\infty$ 时，标引词的权值在[0,1]上，扩展布尔模型就变成建立在模糊逻辑上的布尔检索模型，则“信息文档-提问”之间的相似度为

$$\begin{aligned}\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j) &= \lim_{p\to\infty}\left[\frac{x_1^p+x_2^p+\cdots+x_n^p}{n}\right]^{\frac{1}{p}}=\max(x_1,x_2,\cdots x_n)\\ \mathrm{sim}(\boldsymbol{q}_{\mathrm{and}},\boldsymbol{d}_j) &= \lim_{p\to\infty}\left\{1-\left[\frac{(1-x_1)^p+(1-x)^p+\cdots+(1-x_n)^p}{n}\right]^{\frac{1}{p}}\right\}\\ &= 1-\max(1-x_1),(1-x_2),\cdots,(1-x_n)\\ &= \min(x_1,x_2,\cdots,x_n)\end{aligned} \tag{5-13}$$

(3) 当 p 值在 1 与∞之间时，扩展布尔模型就介于向量模型和布尔模型之间，p 值越大，∧和∨的功能就越强；p 值越小，∧和∨的功能就越弱，直至 $p=1$，其功能完全消失。见图 5-5。

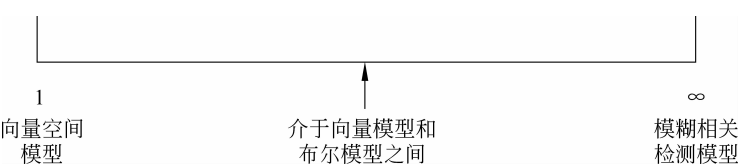

图 5-5 p 值的变化范围

对于提问语言的处理一般是按预先定义的次序对运算符进行分组而展开的，比如对于提问 $\boldsymbol{q}=(t_1\wedge^p t_2)\vee^p t_3$，信息文档 $\boldsymbol{d}_j$ 与提问 $\boldsymbol{q}$ 的相似度通常计算为