目 录

张庆国

第七章　检索效果的改善（续）

在第四章中，我们讨论了检索效果及其一些定性的改善方法。在那之后，我们又几次涉及了检索效果的改善问题。本章我们将再介绍一些改善检索效果的定量的技术方法。

7.1　文献—语词矩阵的若干推论

7.1.1　词联接矩阵

设检索系统中有文献—语词矩阵

$$D=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{pmatrix} \qquad (7.1)$$

其中 a_{ij} 表示第 j 个标引词对第 i 篇文献的重要程度，即权。a_{ij} 的一种有用的取值方法为取离散值 0 和 1。

我们给出矩阵 D 的转置 D^T

$$D^T=\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{N1} \\ a_{12} & a_{22} & \cdots & a_{N2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1M} & a_{2M} & \cdots & a_{NM} \end{pmatrix} \qquad (7.2)$$

再令

$$K=D^TD=(k_{ij})_{MM} \qquad (7.3)$$

则 K 是 M 行 M 列的一个方阵。我们在 5.3 中已经引出了这个矩阵，并且把它称为词联接矩阵。

7.1.2 词结合矩阵和改良型文献—语词矩阵

词联接矩阵可用来确定同一文献的标引词对，而这样的标引词对中的词无须在词义上相关。事实上，意义相近的词一般不全是同一文献的标引词，不然的话，则表明标引中有冗余的成分了。

列检索提问式的人可能使用与文献标引不同、但意义相近的词。为了满足这种要求，我们引入一个词的结合矩阵T。矩阵T的元素 t_{ij} 表示在什么程度上适用于第i个标引词的文献同时也适用于第j个标引词。

$$T=\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1M} \\ t_{21} & t_{22} & \cdots & t_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ t_{M1} & t_{M2} & \cdots & t_{MM} \end{pmatrix} \qquad (7.4)$$

例如，五个标引词

波、速度、海、洋、风

的词结合矩阵可表示如下：

$$T=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 1 \end{pmatrix}$$

这个词结合矩阵表明，第三个标引词“海”和第四个标引词“洋”被认为在标引文献的意义上是等价的（$t_{34}=t_{43}=1$）；$t_{51}=\frac{1}{2}$ 表明用第五个标引词“风”标引的文献中有一些但不是全部也适用于第一个词“波”；$t_{15}=0$ 表示用“波”标引的文献被认为并不一定与“风”相关。

需要指出的是。词结合矩阵不是对称的。如本例中的t_{51}≠t_{15}。

引入词结合矩阵后。我们可以导出改良型的文献—语词矩阵

$$DT=(d_{ij})_{NM} \tag{7.5}$$

DT中的元素

$$d_{ij}=a_{i1}t_{1j}+a_{i2}t_{2j}+\cdots+a_{iM}t_{Mj}$$

$$=\sum_{k=1}^{M}a_{ik}t_{kj} \tag{7.6}$$

我们考察d_{ij}的意义。在定义d_{ij}的式(7.6)中。第一项中的a_{i1}表示文献i与第1个标引词的相关程度。t_{1j}表示第1个标引词与第j个标引词的相近程序。因此。$a_{i1}t_{1j}$表示以第一个标引词作为中介时第i篇文献与第j个标引词的相关程度。其它各项分别是以第二个词。第三个词。…。第M个词作为中介的。于是我们看到。d_{ij}表示在考虑了系统的所有词间关系后。第i篇文献和第j个标引词的相关程度。或权。

例如。用上例中的五个标引词标引了六篇文献。文献—语词矩阵D如下

$$D=\begin{pmatrix}1&1&0&0&0\\1&0&1&0&0\\1&0&0&1&0\\0&1&0&0&1\\0&0&1&0&1\\0&0&0&1&1\end{pmatrix}$$

其改良型文献—语词矩阵

$$DT = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 1 & 1 & 1 \\ \frac{1}{2} & 0 & 1 & 1 & 1 \end{pmatrix}$$

我们通过第五篇文献来比较D和DT。在D中，$a_{54}=0$，这表明第五篇文献与第四个标引词“洋”并没有直接标引关系。但在词结合矩阵T中我们看到，“洋”和第三个标引词“海”在标引文献的意义上是结合的($t_{34}=1$)。而“海”又用来标引了这第五篇文献。因此第四个标引词“洋”也是适用于第五篇文献的($a_{54}=1$)。又，虽然$a_{51}=0$，但$a_{55}=1$，$t_{51}=\frac{1}{2}$，所以$a_{51}=\frac{1}{2}$。即标引词“风”把它对“波”的关系传递给了第五篇文献。$a_{51}=\frac{1}{2}$表示第五篇文献在较小的程度上也适用于“波”。

7.2 与词结合矩阵相关的权和响应向量

设提问向量

$$Q=\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_M \end{pmatrix} \qquad (7.6)$$

其中q_j是用户请求中第j个词所带的权。Q是M维空间中的一个向量，q_j是它的分量。

在文献数据库中，每篇文献均用一文献向量描述。对于以式（7．6）的形式规定的请求，第i篇文献的匹配度值可用公式

$$r_i=\sum_{j=1}^{M}a_{ij}q_j \qquad (7.7)$$

来计算。当i取所有值时，我们得到一个对提问（7．6）的响应向量

$$R=DQ=\begin{pmatrix}r_1\\r_2\\\vdots\\r_N\end{pmatrix} \qquad (7.8)$$

响应向量中的非零分量表示用一个或多个标引词标引的文献，而这些标引词在请求向量中对应非零权。

如果用词结合矩阵T表示标引词之间的联系，那么文献—语词矩阵D应该用其改良型的DT代替，式（7．8）成为

$$R=DTQ \qquad (7.9)$$

词结合矩阵T和提问向量Q的乘积

$$TQ=\begin{pmatrix}t_{11}q_1+t_{12}q_2+\cdots+t_{1M}q_M\\t_{22}q_1+t_{22}q_2+\cdots+t_{2M}q_M\\\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\t_{M1}q_1+t_{M2}q_2+\cdots+t_{MM}q_M\end{pmatrix}$$

是一个向量。我们通过它的一个分量来看这个向量的意义。

TQ的一个分量是

$$t_{11}q_1+t_{12}q_2+\cdots+t_{1M}q_M$$

与原提问向量中的相应分量q_1比较，有

$$t_{11}q_1+t_{12}q_2+\cdots+t_{1M}q_M$$

$$= t_{i1}q_1 + (t_{i1}q_1 + \cdots + t_{i,i-1}q_{i-1} + t_{i,i+1}q_{i+1} + \cdots + t_{iM}q_M)$$

$$= q_i + (t_{i1}q_1 + \cdots + t_{i,i-1}q_{i-1} + t_{i,i+1}q_{i+1} + \cdots + t_{iM}q_M)$$

$$\geqslant q_i$$

当上式中的等号成立时，说明 $t_{ij} = 0\ (i \neq j)$，因而第 i 个标引词与其它标引词无关；若第 i 个标引词与其它标引词相关，则至少有一个 $t_{ij} \neq 0\ (i \neq j)$，这时上式中的不等号成立。于是，TQ 中的各分量是原提问向量 Q 中各相应分量经过词结合矩阵 T 修正后的值。

我们再用 M 表示屏蔽运算，它把小于某给定阈值（如 1）的所有分量都化为零，并用 1 取代其它的所有分量，那么 MTQ 就是一个向量 $Q^{(1)}$，其中值为 1 的分量表示被看作是和提问向量 Q 的分量等价或非常相关的标引词。这样的标引词被称为第一代的词。

例如，设提问向量 Q 和词结合矩阵 T 如下：

$$Q = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \qquad T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 1 \end{pmatrix}$$

于是有

$$TQ = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ \frac{1}{2} \end{pmatrix}$$

指定M的屏蔽级为 1，则有

$$Q^{(1)}=MTQ=\begin{pmatrix}1\\0\\1\\1\\0\end{pmatrix}.$$

其中第一代的词为与分量 $q_1^{(1)}$、$q_3^{(1)}$ 和 $q_4^{(1)}$ 相应的标引词。

仿照从Q中导出 $Q^{(1)}$ 的方式，可以从 $Q^{(1)}$ 中导出

$$Q^{(2)}=MTQ^{(1)}=(MT)^2Q$$

$$Q^{(3)}=(MT)^3Q$$

……

等等。与 $Q^{(2)}$、$Q^{(3)}$、…中非零分量对应的词分别称为第二代词、第三代词、…、等等。如果用 S_1 表示第1代词的集合，则显然有

$$S_0 \subset S_1 \subset S_2 \subset \cdots$$

其中 S_0 表示与原提问向量Q中的非零分量对应的词，它们可称作第零代词，即提问词。

提问词、第一代词、第二代词及更多代词的全体与向量

$$Q+MTQ+(MT)^2Q+\cdots \qquad (7.10)$$

的非零分量相对应。上式也可写成以下形式

$$(I-MT)^{-1}Q \qquad (7.11)$$

其中I代表单位矩阵

第 $1+1$ 代词的个数等于或大于第1代词的个数。多出来的这些词是我们通过词结合矩阵扩展出来的。对这些扩展出来的词应该加以限制，而不能把它们与上一代甚至上几代的词相同看待。限制是通过加权实现的，即连续几代的词的权递减。前面的序列(7.9)

可改写成

$$Q+\lambda MTQ+(\lambda MT)^2Q+\cdots \qquad (7.12)$$

或

$$(I-\lambda MT)^{-1}Q \qquad (7.13)$$

我们把

$$(I-\lambda MT)^{-1}$$

称为扩展因子。使用扩展因子后，提问向量由Q变成$(I-\lambda MT)^{-1}Q$。于是响应向量的形式(7.8)变成

$$R=D(I-\lambda MT)^{-1}Q$$

7.3 通过结合反馈进行的提问自动修正

我们考虑一个检索输出的n篇文献的文献—语句矩阵D和它的转置D^T。

$$D=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}$$

$$D^T=\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}$$

其中m表示这n篇输出文献涉及的标引词个数。在矩阵D^T，每一行表示一个标引词与这n篇输出文献的各自适用程度。

响应向量

$$R=\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

其中分量 r_i 表示第 i 篇输出文献与用户提问的匹配程度。因此，向量 D^TR 的第 i 个分量

$$\sum_{t=1}^{n} a_{ti} r_t = a_{1i} r_1 + a_{2i} r_2 + \cdots + a_{ni} r_n \qquad (7.15)$$

表示就这 n 篇输出文献而言，第 i 个标引词与用户提问的适用程度。这个值应该被加进原提问 Q 中相应的分量值 q_i 中去。即

$$q_i' = q_i + \sum_{t=1}^{n} a_{ti} r_t \qquad (7.16)$$

当 i 取遍所有 m 个标引词时，我们便有修改后的提问向量

$$Q' = Q + D^TR = \begin{pmatrix} q_1 + \sum_{t=1}^{n} a_{t1} r_t \\ q_2 + \sum_{t=1}^{n} a_{t2} r_t \\ \cdots\cdots\cdots\cdots \\ q_m + \sum_{t=1}^{n} a_{tm} r_t \end{pmatrix} \qquad (7.17)$$

这个新的提问向量不仅取决于原来的请求 Q，而且还取决于所产生的输出 R。修正后的请求取决于输出的程度可以通过引入参数 λ 控制，即令

$$Q' = Q + \lambda D^TR \qquad (7.18)$$

下图是根据上面的反馈形式处理向量 Q 的情况：

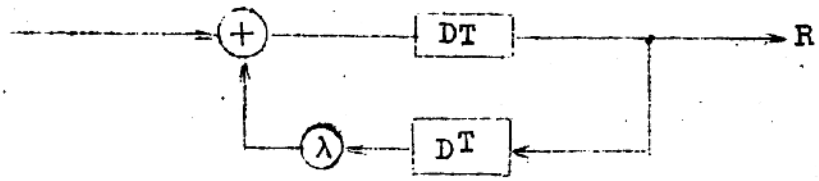

图 7.1 通过输出反馈自动调节用户请求

用修改后的提问向量 $Q+\lambda D^{T}R$ 代替 Q 之后，式（7.9）便成为

$$R = DT(Q+\lambda D^{T}R) \qquad (7.19)$$

解出 R

$$\begin{aligned} R &= (I-\lambda DTD^{T})^{-1}DTQ \\ &= [I+\lambda DTD^{T}+(\lambda DTD^{T})^{2}+(\lambda DTD^{T})^{3}+\cdots]DTQ \end{aligned} \qquad (7.20)$$

将上式代入式（7.18），则我们得到反馈检索的提问形式

$$\begin{aligned} Q' &= Q+\lambda D^{T}R \\ &= Q+\lambda D^{T}[I+\lambda DTD^{T}+(\lambda DTD^{T})^{2}+\cdots]DTQ \\ &= Q+\lambda D^{T}DTQ+\lambda^{2}D^{T}DTD^{T}DTQ+\cdots \\ &= [I+\lambda D^{T}DT+(\lambda D^{T}DT)^{2}+\cdots]Q \\ &= (I-\lambda D^{T}DT)^{-1}Q \\ &= (I-\lambda KT)^{-1}Q \end{aligned} \qquad (7.21)$$

式中 K 是词联接矩阵

检索的实际过程如下：设检索从一个请求向量 Q 开始，并指定需要一个给定值数量的命中文献，如 n_0 篇。开始令 $\lambda=0$，按式（7.21）给出的提问向量检索。每次检索后，得到一响应向量，其中的非零分量表示输出的命中文献。如果检出文献少于 n_0 篇，或感觉到查全率过低，则给定一个较小的正数 λ，并按式（7.21）重新确定提问向量并再次检索。如检索结果仍小于 n_0 篇或查全率还是低，则逐渐增大 λ 的值反复进行上述过程，直至输出的文献数达到用户要求的 n_0 篇，或用户对查全率基本满意为止。上述增大 λ 值的过程可以是自动的。

每次反馈时，提问向量可能增加一些非零分量，这意味着下次检索将增加一些检索词，即扩大了请求。我们注意到，扩大请求的

方式是由内部标引词的词间关系而不是由用户的兴趣控制的。如果用户参预控制扩大请求的过程，则效果会更好一些。用户参预控制的方法是：每次λ值变化后，打印输出与新的提问向量中非零分量相应的词，以及由式（7．15）确定的权。这样，用户便能知道提问向量Q的扩展情况。如果他认为新增加的词与原提问是无关的，那么可删掉这个词。

在本节的讨论中我们可以看到，提问的修正方法并非直接出于优化查全率和查准率的企图，它受数据库特性的影响要大于受用户要求的影响。在下一节中，我们将介绍一种直接从提高查全率和查准率的角度考虑的检索提问式修正方法。

7．4 检索策略的最优化

设用户的提问向量为

$$Q=\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{pmatrix}$$

我们已有响应向量

$$R=DQ=\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

其中

$$r_i=\sum_{j=1}^{m} a_{ij}q_j \qquad (i=1, 2, \cdots, n)$$

我们称r_i为文献i的响应函数。