第 16 卷第 4 期,增刊　　广西科学院学报　　Vol. 16, No. 4 & Supplement
2000 年 11 月　　Journal of Guangxi Academy of Sciences　　November 2000

扩展布尔检索模型——Salton 模型
Expanded Bull Retrieval Model——Salton Model

李广原
Li Guangyuan

(广西师范学院计算中心　南宁　530001)
(Computer Centre, Guangxi Teacher's College, Nanning, 530001)

摘要　介绍一种扩展的布尔检索模型——Salton 模型，该模型通过对标引词加进权值，将向量检索与布尔检索融为一体，利用矢量方法对扩展布尔检索进行讨论，得到一个计算相似度较好的式子，它在一定程度上克服布尔检索的某些缺点。
关键词　布尔检索　Salton 模型　相似度
中图法分类号　G 354.4
Abstract　An expanded Bull retrieval model——Salton model is introduced. The model units the vector retrieval with the Bull retrieval by applying weight value to the indexing terms, discussing Bull retrieval by utiliting the vector method to obtain a better formula which calculates the similarity. This model overcomes drawbacks of Bull retrieval in a certain extent.
Key words　Bull retrieval, Salton model, similarity

在信息检索中，大多是基于关键词一类的检索，即在待检索的文本集中，每篇文本用若干个标引词加以标引，标引词用来反映文本的内容，用户检索时，给出检索标引词，检索系统根据用户给定的标引词和文本集中每一文本进行匹配，根据匹配的大小来决定是否输出该文本，匹配的程度常用相似度来表示，相似度越大，表示文本越符合用户的需要，反之，越不符合用户的需要。目前，常用的文本信息检索方法有：布尔检索、向量空间检索、概率检索和全文检索。在这些方法当中，尤以布尔检索方法最为成熟。

1　布尔检索

所谓布尔检索，就是采用布尔代数的方法，用布尔表达式表示用户提问，通过对文本标识与用户给出的检索式进行逻辑比较来检索文本。

设文本集 D 中某一文本 i，该文本可表示为：

$$D_i = (t_1, t_2, \cdots, t_m),$$

其中，t_1，t_2，…，t_m 为标引词，用以反映 i 的内容。

另设用户某一检索式如下：

2000-08-04 收稿。

$Q_j = (t_1 \wedge t_2) \vee (t_3 \wedge (t_4))$.

对于该检索式，系统响应并输出的一组文本应为：它们都含有标引词 t_1 和 t_2，或者含有标引词 t_3，但不含有标引词 t_4。

布尔检索具有简单、易理解、易实现等优点，故得到广泛的应用。1967 年后，布尔检索模型正式被大型文献检索系统采用，并渐成为各种商业性联机检索系统的标准检索模式，服务信息情报界 30 多年，直到现在，大多数商用检索系统仍采用布尔检索。

尽管布尔检索有着种种的优点，但是它的缺点仍然是明显的，它存在的主要缺陷有以下几点。

（1）布尔逻辑式的构造不易全面反映用户的需求。

用标引词的简单组配不能完全反映用户的实际需要，用户需要那一方面内容的文本，需要到多大程度，这是检索式无法表达清楚的，如对上述检索式，t_1 和 t_2，究竟用户希望能得到更多地反映 t_1 内容的文本还是反映 t_2 内容的文本，传统的布尔检索无法解决此问题。

（2）匹配标准存在某些不合理的地方。

例如，在响应某个用“∧”连接的检索时，系统把只含有其中一个或数个但非全部检索词的文本看作与那些根本不含有其中一个检索词的文本一样差，同样加以排除；另一方面，用响应某个用“∨”连接的检索式时，系统都不能把含有所有这些检索词的文本看作比那些只含有其中一个检索词的文本更好一些。

（3）检索结果不能按照用户定义的重要性排序输出。

系统检索输出的文本中，排在第一位的文本不一定是文本集中最适合用户需要的文本，用户只能从头到尾浏览才能知道输出文本中那些更适合自己的需要。

2 Salton 模型

Salton 模型是由 Salton 于 1983 年提出的一种所谓的扩展布尔检索模型，它是将向量检索模型与布尔检索模型融为一体，并克服了传统希尔模型的一些缺陷，下面我们用矢量的方法来讨论布尔检索。

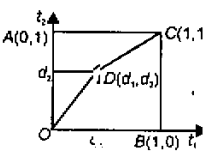

图 1 布尔检索矢量表示法

设文本集中每篇文本仅由两个标引词 t_1 和 t_2 标引，并且 t_1、t_2 允许赋以权值，其权值范围为 $[0, 1]$，权值越接近 1，说明该词越能反映文本的内容，反之，越不能反映文本的内容，在 Salton 模型中，上述情形用平面坐标系上某点代表某一文本和用户给出的检索式，见图 1。图 1 中的横、纵坐标用 t_1、t_2 表示，其中 $A(0, 1)$ 表示词 t_1 权值为 0，词 t_2 权值为 1 的文本，$B(1, 0)$ 表示词 t_1 权值为 1，词 t_2 权值为 0 的文本，$C(1,1)$ 表示词 t_1、t_2 的权值均为 1 的文本，文本集 D 中凡是可以用 t_1、t_2 标引的文本可以用四边形 OACB 中某一点表示，同样，用户给出检索式后，也可用四边形 OACB 中某一点表示。

下面我们来看看 Salton 模型中是如何构造相似度计算式的。

对于由 t_1 和 t_2 构成的检索式 $q = t_1 \vee t_2$，在图 1 中只有 A、B、C 3 点所代表的各文本才是最理想的文本，对于某一文本 D 来说，当 D 点离 A、B、C 3 点越接近时说明相似度越大，或者说，当 D 点离 O 点越远时，相似度越大。因而 D 与 O 的距离

$$|DO| = \sqrt{(d_1 - 0)^2 + (d_2 - 0)^2} = \sqrt{d_1^2 + d_2^2}$$

可以作为我们衡量一文本与查询 q 的相关程度的一个尺度，显然 $0 \leqslant |DO| \leqslant \sqrt{2}$，为了使相似度控制在 0 与 1 之间，将相似度定义为：

$$\mathrm{sim}(D,Q(t_1 \vee t_2)) = \sqrt{\frac{d_1^2 + d_2^2}{2}}, \tag{1}$$

对于由 t_1 和 t_2 构成的查询 $q = t_1 \wedge t_2$，只有 C 点才是最理想的文本，用 D 与 C 的距离

$$|DC| = \sqrt{(1-d_1)^2 + (1-d_2)^2}$$

作为我们衡量一文本与查询 q 的相关程度的一个尺度，于是，把相似度定义为：

$$\mathrm{sim}(D,Q(t_1 \wedge t_2)) = 1 - \sqrt{\frac{(1-d_1)^2 + (1-d_2)^2}{2}}, \tag{2}$$

(1)、(2) 式还可推广到对检索标引词进行加权的情形，设检索标引词 t_1、t_2 的权值分别为 a，b，$0 \leqslant a,b \leqslant 1$，则(1) 式、(2) 式可进一步推广为：

$$\mathrm{sim}(d,Q(t_1,a) \vee (t_2,b)) = \sqrt{\frac{a^2 d_1^2 + b^2 d_2^2}{a^2 + b^2}} \tag{3}$$

$$\mathrm{sim}(d,Q(t_1,a) \wedge (t_2,b)) = 1 - \sqrt{\frac{a^2(1-d_1)^2 + b^2(1-d_2)^2}{a^2 + b^2}}, \tag{4}$$

Salton 模型还给出了把标引词推广到 n 个时的相似度计算公式。

设 $d = (d_1, d_2, \cdots, d_n)$，

其中 d_i 表示第 i 个标引词 t_i 的权值，$0 \leqslant d_i \leqslant 1$。

由布尔运算符“$\vee$”及“$\wedge$”所确定的检索式分别为：

$Q_{\vee(p)} = (t_1,a_1) \vee (t_2,a_2) \vee \cdots \vee (t_n,a_n)$，

$Q_{\wedge(p)} = (t_1,a_1) \wedge (t_2,a_2) \wedge \cdots \wedge (t_n,a_n)$，

其中 a_i 表示第 i 个检索标引词 t_i 的权值，$0 \leqslant a_i \leqslant 1$，这里，$p$ 是一个可变的量，$1 \leqslant q \leqslant \infty$，在 Salton 模型中，以(3)、(4) 式作为基本的出发点，在 n 个标引词生成的 n 维欧氏空间中应用 L_P 矢量模公式进行欧氏模的计算，将文本和查询的相似度定义为：

$$\mathrm{sim}(d,Q_{\vee(p)}) = \left[\frac{a_1^p d_1^p + a_2^p d_2^p + \cdots + a_n^p d_n^p}{a_1^p + a_2^p + \cdots + a_n^p}\right]^{\frac{1}{p}},$$

$$\mathrm{sim}(d,Q_{\wedge(p)}) = 1 - \left[\frac{a_1^p (1-d_1)^p + a_2^p (1-d_2)^p + \cdots + a_n^p (1-d_n)^p}{a_1^p + a_2^p + \cdots + a_n^p}\right]。$$

3　结语

在文本信息检索中，布尔检索不仅具有简单、易理解等特点，而且易于在计算机中加以实现，是一种最为常用的检索方法。扩展的布尔模索模型——Salton 模型克服了传统布尔模型的一些缺陷，更符合了用户的需要。

参考文献

1　赖茂生等编著．计算机情报检索．北京：北京大学出版社，1993.

2　康耀红著．现代情报检索理论．北京：科学技术文献出版社，1990.

（责任编辑：黎贞崇）