中文全文检索算法研究

杨一平 （首都经济贸易大学信息管理系 100026）

一、全文检索系统概况

1. 全文检索系统应具备的功能

一个全文检索系统至少要具备两个功能：

(1)文章中任何有意义的词、字都可被检索。

(2)能对检索词之间的关系进行位置和逻辑操作。

另外，全文检索的响应时间应在秒级以内。

2. 中文全文检索的基本技术

目前，已开发出来的中文全文检索系统，其基本技术可归纳为三种类型：

(1)主题词索引。建立主题词索引。根据主题词典，对检索条件中切分后相邻自由词组合与主题词典匹配，得出检索结果。

(2)词索引。对源文献进行分词，抽词，用切分获得的词的全体作为标引词，据此建立索引文件。检索时将检索条件(提出输入)进行分词，对每个词进行检索，并根据检索条件中的位置、逻辑关系找到符合条件的文章。

(3)单汉字索引。以汉字单字为单位对源文献进行分词，据此自动标引，避免了分词组合造成的歧义性。检索时将检索条件(提问输入)切分成单个汉字，通过索引对汉字进行多次匹配组合成词。

二、全文检索算法研究

1. 基于关系代数理论的单汉字索引算法

设 D 是检索范围内的所有文章的集合。全文检索可以表述为：给定任一个含有通配符的字符串 φ 为检索条件，以 D 中满足 φ 的文章为输出结果，若一篇文章中至少含有一个与 φ 匹配的字符串，则称该文章是满足的。

检索范围内的文章具有以下内容：

(1)有唯一的标识符(＃ID)来表示一篇文章。这里设所有文章的＃ID 按自然数升序排列。

(2)标题包括正标题和副标题。

(3)作者有一至多名。

(4)主题词，由作者、编辑等给出，它能反映文章的主题、重要时间、重要人物等。

(5)时间，文章的出版时间。

(6)正文，文章的内容。正文可以看成是由若干个检索字按一定顺序排列组成的。每个检索字在文章中占据一定的位置。用关系 E(＃ID, C)表示所有文章的集合 D。C 是文章正文，＃ID 是该文章的唯一标识符，以自然数来表示。

检索字及建立单汉字索引：

每个字(不论字母，数字，汉字均占两个字节)在文章中可能出现零次或多次。文章中每一个位置都有唯一的一个字符相对应。

字：W{字，该字在文章中的位置，文章标识＃ID}，一个字在文章中可以出现多次，即有多个位置。

字与文章的关系如下图：1<=i<=m，m=文章篇数，1<=j<=n，n=检索字数。

	T_1	T_2	…	T_n
#ID$_1$	P_{11}	P_{12}	…	P_{1n}
#ID$_2$	P_{21}	P_{22}	…	P_{2n}
…	…	…	…	…
#ID$_m$	P_{m1}	P_{m2}	…	P_{mn}

T 为检索字，＃ID 为文章的标识符，P 为字 T 在文章＃ID 中的位置集合，若字 Tj 不在文章＃IDi 中出现，则 Pij 为 Tj 在该文章中的位置集合。

用关系 I(B, ＃ID, P)表示索引，B 是被检索字，＃ID 是该检索字所在文章的标识符，P 是 B 在文章＃ID 中的位置的集合。

对于 E 中每个元组 e，将其分量 e[C](文章正文)所包含的每个检索字作为索引，并将结果作为多个元组插入到 I(B, ＃ID, P)中。

检索条件 φ 是起止于检索字的含有通配符的字符串，通配符 $ 表示一个任意字中含有的 m 个检索字为 T1, T2, ... Tm。

匹配的定义：设 e 是 E 的任意元组，β 是 e[C]中的一

个字符串，它所含检索字集合用 δ 表示。φ(1,i)是 φ 的一个子串(<=i<=m)。i=1 时，若 δ1 与 φ(1,1)满足=T1，称 β 与 φ(1,i)匹配，i>1 时，设 β 中有关的 i 个检索字按出现顺序为 δ1,δ2…δi，它们在文章中的位置分别为 P1,P2,…,Pi，若它们和 φ(1,i)有关系 δk=TK(K=1,2,…,i),Pk,Pk+1，φ(k,k+1) (K=1,2,…,i-1)满足(*)式，则 β 与 φ(1,i)匹配。

$$P_kP_{k+1}=1 \quad \text{若}\varphi(k,k+1) \quad \text{不含通配符}$$
$$P_kP_{k+1}=1+N \quad \text{若}\varphi(k,k+1) \quad \text{含 N 个通配符 \$} \qquad (*)\text{式}$$

若 E 中共有 Q 个元组与 φ(1,i)匹配，其中第 K 个元组 ek[C]共有 L_K 个匹配串，位置用集合 P_K 来表示，则<e_k[#ID],P_K>称为 φ(1,i)的一个匹配组．所有 Q 个匹配组集合称为 φ(1,i)的全部匹配组。

根据给定的 φ 和 I 导出关系 R(#ID,P)，即检索结果。

φ(1,M)的全部匹配组经过 M 次迭代得到．若第 i 次(1<=i<=m-1)迭代的结果是求得((1,i)的全部匹配组，则第 i+1 次迭代便可获得 φ(1,i+1)的全部匹配组，用迭代的原理来导出检索结果。

算法如下：

(1)R_1 或 R_2 不空，取 $r_1(R_1, r_2 \in R_2$，且 r_1[#ID]=r_2[#ID]。当 R_1 或 R_2 为空时停止。

(2)将 r_1[P]，r_2[P]中可能满足(*)式的位置及 φ(i,i+1)代入(*)式计算，将 P={$P_1,P_2,\cdots,P_L$}插入 Z，转(1)。若令∈R_2 中的#ID{R1 中的#ID}，则可去掉 R_2 中的多余元组，提高执行效率。

2.基于模糊数学的主题词检索算法

主题词一般由作者，编者，出版者提供。它包括该文章的中心思想，重要人物，重要事件等等。一篇文章中有若干个主题词，每个主题词在文章中出现一次或多次。这种关系可以表示为：

G(#ID,W)．#ID 是文章标识符，W 是词集合。

词：(主题词 Wi，该主题词在文章中出现的次数 $wcount_i$)

可根据作者，编者，出版者提供的主题词，提高查准率。因此，单汉字索引+主题词控制不失为一种好的检索方法。

这里我们应用模糊数学的有关概念来研究如何确定查准率的问题。即在给定论域 U 中的元素两两之间都赋于区间[0,1]内的一个数，叫相似系数。它的大小表征两个元素彼此接近或相似的程度。

在这里，我们可以把查准率看成检索条件和文章之间的相似度。因为检索条件越符合文章的主题(二者越相似)，查准率越高，也就是文章和检索条件之间的相似度越高，查准率越高。

模型如下：

用户的检索条件可以切分为零到多个检索词，它可用矢量形式表示。一个典型的检索条件可以表示为：

$Q=(q_1,q_2,\cdots q_m)$

其中每一个 q 表示赋给检索条件 Q 的一个检索词。

文章主题词可以用矢量形式表示为：

$D=(t_1,t_2,\cdots,t_n)$

其中，t_i 是文章的主题词。

令特征词集合=检索词集合∩文章主题词集合，设特征词集合中特征词的个数为 p。这样可去掉没用的词，缩小词的范围，提高检索速度。这样，Q=D。

引入权的概念，将 Q 和 D 重新表示。

令 W_{dk},W_{qk}表示文章 D 中的主题词 t_K、检索条件 Q 里检索词 q_K 的权，用它们来表示文章和检索条件的特征值。

$W_{dk}=t_k$ 在文章中出现的次数/所有特征词在文章中出现的次数之和。

$W_{qk}=q_k$ 在检索条件中出现的次数/所有特征词在检索条件中出现的次数之和。

查准率是基于此检索条件和文章相似度的大小。

一个检索条件和文章相似度可以通过传统的矢量积公式表示：

$$SIMILARITY(Q,D)=\frac{\sum_{k=1}^{m} W_{qk}W_{tk}}{\sum_{k=1}^{m} W_{qk}^2 \sum_{k=1}^{m} W_{tk}^2}$$

一个在检索条件矢量和文章矢量之间进行匹配的矢量匹配系统提供了以文章 D 和检索条件 Q 的相似度值的递减输出．由于这里的检索是对单汉检索结果上的文章集而言，故这种排序输出保证查全率的基础上，表示了检索条件反映主题的程度，给检索者提供了极大的方便。

(来稿时间:1997 年 6 月)