第 35 卷 第 9 期 计 算 机 工 程 2009 年 5 月
Vol.35 No.9 Computer Engineering May 2009
·软件技术与数据库· 文章编号：1000—3428(2009)09—0028—03 文献标识码：A 中图分类号：TP311

基于语言变量的关系数据库模糊查询

陈逸菲，叶小岭，张颖超
(南京信息工程大学信息与控制学院，南京 210044)

摘 要：在模糊理论的基础上，将权重概念引入关系数据库模糊查询中，以体现用户对查询中各个属性的相对重视程度。记录按匹配度的降序输出，方便用户选择。权重和匹配度都是语言变量，取值为语言值，更加贴近自然。采用模糊集合的 alpha 截集去模糊的思想，将带语言值权重的模糊查询条件转化为精确的 SQL 语句，利用 RDBMS 的机制进行记录的筛选，避免对整个数据库表的扫描，在一定程度上保证查询的效率。

关键词：模糊查询；语言变量；权重

Fuzzy Queries Based on Linguistic Variables in Relational Databases

CHEN Yi-fei, YE Xiao-ling, ZHANG Ying-chao
(College of Information and Control, Nanjing University of Information Sciences & Technology, Nanjing 210044)

【Abstract】The concept of weight is introduced in fuzzy queries based on theory of fuzzy set, which shows the relative importance that users pay to different attributes in query clauses. The records that satisfy queries are output according to the decrease of matching degrees, so users can select what they want more conveniently. Weights and matching degrees are linguistic variables, whose values are linguistic terms, and the queries become more flexible. The alpha-cut of fuzzy set is used to translate the linguistic fuzzy querying conditions into crisp SQL clauses. It is possible to make use of the mechanism of RDBMS to filter the records, which avoids scanning the whole table and assures the efficiency.

【Key words】fuzzy queries; linguistic variables; weight

1 概述

自 Zadeh 提出模糊集合理论以来，出现了许多基于该理论的数据库模糊查询方法。

文献[1]基于模糊数的交、并、补的 α 截集运算，对 SQL 语言进行扩展，提出了将模糊的查询条件转换为精确 SQL 语句的方法。文献[2]给出了 SQL 语言较完整的模糊化版本 SQLf。由于用户在查询时对于每个子条件的倾向、重视程度不同，有必要为各个子查询条件指定权重。文献[3-4]提出将权重表示为[0,1]之间的数值，权重值越大，表示对应的条件在整个查询中起的作用越大。由于权重和模糊集合的隶属度都是[0,1]之间的值，因此在模糊查询中引入权重是很方便的。文献[5]在 SQLf 的基础上将模糊查询扩展为带权重的情况。但是对于非专业用户而言，[0,1]之间的数值权重确定起来并不方便。文献[6]指出模糊集合隶属度具有很强的健壮性，没有必要对隶属度给出精确的估计，表现在模糊隶属度本来就是用于描述不精确的信息，因此，没有必要精确为数值；在绝大多数模糊查询的文献，如文献[1-2,5]中都是使用 min/max 将隶属函数进行结合，此时没有必要知道隶属度的具体值。基于这 2 点可知，只要建立隶属度的全序关系即可。类似，权重也没有必要为[0,1]之间的数值，只要能反映出用户对条件的相对重视程度即可。因此，建立语言值表示形式的隶属度、权重是可行的。文献[7]考虑了在文献检索系统中引入语言值权重，但是这些方法对关系数据库的查询而言并不适用。

本文在文献[8]的基础上，提出了一种新的利用模糊集合截集来计算语言值表示的模糊查询的方法，该方法避免了原方法需扫描整个数据库表的缺点，提高了查询的效率。

2 语言变量及权重、匹配度

语言变量是以自然语言或人工语言中的字或句，而不是以数作为值的变量。语言变量的值被称为语言值，用于描述那些不太精确的现象[7]。文献[9]给出了语言变量的定义。

定义 1 语言变量由一个 5 元组(H, $T(H)$, U, G, M)表示。其中，H 是变量的名称；$T(H)$是 H 的术语集合，即 H 的语言值的集合；H 的每一个值是由 X 表示的模糊变量，并且 X 定义在论域 U 上，与基本变量 u 相关；G 是用来产生 H 的值的语法规则；M 是语义规则，与每个 X 的意义 $M(X)$相关，$M(X)$是 U 上的模糊子集：

$T=X1+X2+\cdots+Xi+\cdots$

可以将“权重”和“匹配度”作为语言变量，给出如下定义：

$$T(Importance)/T(Matching\ Degree)=absolutely\ high+extremely\ high+very\ high+high+fairly\ high+somewhat\ high+medium+somewhat\ low+fairly\ low+low+very\ low+extremely\ low+absolutely\ low \quad (1)$$

要将上述语言变量形式的权重、匹配度用于模糊查询，就必需在语言变量的各语言值之间建立全序关系，即定义有序语言值。

基金项目：江苏省“六大人才高峰”基金资助项目(06-A-07)；南京信息工程大学科研基金资助项目(y644)
作者简介：陈逸菲(1981－)，女，讲师、博士研究生，主研方向：模糊信息处理，时空数据库；叶小岭，副教授；张颖超，教授
收稿日期：2008-11-28 **E-mail：**ch_yi_f@126.com

定义 2[7] 有序语言值：定义一个有限的基数为奇数的语言值标记集合 $S=\{s_i\}$，$i\in\{0,1,\cdots,T\}$（$s_i\geqslant s_j$ 如果 $i\geqslant j$）。集合中的第 $T/2+1$ 个术语表示“近似为 0.5”，其他的术语关于它对称。每个语言值由定义在[0,1]上的模糊数表示。本文采用梯形隶属函数 (a_i,b_i,c_i,d_i)，其中 a_i，d_i 分别为最左、最右的端点，b_i，c_i 区间范围内的隶属度为 1.0。

此外定义以下运算：

$$Neg(s_i)=s_j;\ j=T-i \tag{2}$$

$$\max(s_i,s_j)=s_i \text{ 如果 } i\geqslant j \tag{3}$$

$$\min(s_i,s_j)=s_i \text{ 如果 } i\leqslant j \tag{4}$$

本文使用 13 个语言值标记构成的集合来运算，见表 1[10]。

表 1　语言值权重/匹配度和其对应的梯形模糊数

语言值符号	语言值权重/匹配度	梯形模糊数
s_{12}	*Absolutely high*	(1.0,1.0,1.0,1.0)
s_{11}	*Extremely high*	(0.901 7,0.949 3,1.0,1.0)
s_{10}	*Very high*	(0.806 5,0.854 1, 0.901 7,0.949 3)
s_9	*high*	(0.711 3,0.758 9, 0.806 5,0.854 1)
s_8	*Fairly high*	(0.616 1,0.663 7,0.711 3,0.758 9)
s_7	*Somewhat high*	(0.520 9,0.568 5,0.616 1,0.663 7)
s_6	*Medium*	(0.425 7,0.473 3,0.520 9,0.568 5)
s_5	*Somewhat low*	(0.330 5,0.378 1, 0.425 7,0.473 3)
s_4	*Fairly low*	(0.235 3,0.282 9, 0.330 5,0.378 1)
s_3	*Low*	(0.140 1,0.187 7, 0.235 3,0.282 9)
s_2	*Very low*	(0.047 6,0.093 5,0.140 1,0.187 7)
s_1	*Extremely low*	(0.0,0.0, 0.047 6,0.093 5)
s_0	*Absolutely low*	(0.0,0.0,0.0,0.0)

定义 3[7] *NTL*(Numeric To Linguistic label)是一个将[0,1]区间的数映射到有序模糊语言标记集合 S 上的函数：

$$NTL:\ [0,1]\to S;$$

$$NTL(x)=Sup_l\{s_l\in S:\mu_{s_l}=Sup_n\{\mu_{s_n}(x)\}\} \tag{5}$$

例如某人的年龄为 30 岁，其关于年轻的数值隶属度 $\mu_{\text{young}}(30)=0.5$，根据式(5)得其对应的语言值隶属度为 $NTL(0.5)=s_6$。此时不再用[0,1]之间的点，而是用[0,1]区间上的一个(梯形)模糊数 s_6 来表示隶属度。前者称为Ⅰ型模糊数，后者称为Ⅱ型模糊数，即Ⅱ型模糊数的隶属度是一个Ⅰ型模糊数[9]。

3　语言值权重和匹配度的计算

令 A，A_i 为关系数据库中关系 T 的属性。假设用户的查询如下：

SELECT　A
FROM　　T
WHERE　A_1 is C_1 AND A_2 is $C_2\cdots$AND A_k is C_k
WEIGHT　A_1 is w_{A_1}; A_2 is w_{A_2};$\cdots$; A_k is w_{A_k}
WITH　α

其中，C_i 是模糊概念，如“年轻”，用Ⅰ型模糊数表示；w_{A_i} 是 A_i 的语言值权重；α 是语言值匹配度的阈值。

构成了有序语言值形式的条件集合：

$$\{(w_{A_1},m_{C_1}(A_1)),(w_{A_2},m_{C_2}(A_2)),\cdots,(w_{A_k},m_{C_k}(A_k))\}$$

其中，$m_{Ck}(A_k)$ 表示属性 A_k 关于“A_k is C_k”的语言值隶属度。

记录 R 关于上文的语言值匹配度计算如下：

$$MD(R)=LWC[(w_{A_1},m_{C_1}(A_1)),(w_{A_2},m_{C_2}(A_2)),\cdots,(w_{A_k},m_{C_k}(A_k))]=\min_i\{\max(Neg(w_{A_i}),m_{C_i}(A_i))\} \tag{6}$$

类似若将上文中的“AND”换成“OR”，则有

$$MD(R)=LWD[(w_{A_1},m_{C_1}(A_1)),(w_{A_2},m_{C_2}(A_2)),\cdots,(w_{A_k},m_{C_k}(A_k))]=\max_i\{\min((w_{A_i},m_{C_i}(A_i))\} \tag{7}$$

其中，*LWC* 为 Linguistic Weighted Conjunction；*LWD* 为 Linguistic Weighted Disjunction。

例 1　对表 2 中的关系 Person 执行下列查询

SELECT　name
FROM　　Person
WHERE　age is young AND height is tall
WEIGHT　age is medium; height is very high
WITH　medium

其中，young, tall 的隶属函数为

$$\mu_{\text{young}}(x)=\begin{cases}1 & x\leqslant 25\\ \dfrac{1}{1+\left(\dfrac{x-25}{5}\right)^2} & x>25\end{cases} \tag{8}$$

$$\mu_{\text{tall}}(x)=\begin{cases}1 & x\geqslant 190\\ \dfrac{1}{1+\left(\dfrac{x-190}{10}\right)^2} & x<190\end{cases} \tag{9}$$

表 2　Person

编号	姓名	年龄	身高/cm
*R*1	王刚	20	173
*R*2	吴平	31	165
*R*3	夏明	33	181
*R*4	张昊	40	175
*R*5	毛建	25	183
*R*6	苏宁	28	177
*R*7	范冰	39	170
*R*8	丁俊	22	185
*R*9	贾涛	26	172
*R*10	陈强	31	183

以 *R*1 为例说明计算方法：

$$\mu_{\text{young}}(R1)=1,\mu_{\text{tall}}(R1)=0.257$$

根据式(5)得到对应的语言值隶属度为 s_{12}, s_3，代入式(6)：

$$MD(R1)=\min(\max(Neg(s_6),s_{12}),\max(Neg(s_{12}),s_3))=s_3$$

4　基于 α 截集的改进算法

显然直接根据式(6)、式(7)计算匹配度进行查询，要计算数据库表中每条记录相关属性的隶属度，效率很低。这里给出利用 α 截集，将带有语言值权重的模糊查询条件转化为精确 SQL 语句的新方法。

4.1　DLWC(Derivation of LWC)算法

对式(6)进行分析，当 $MD(R)=\min\{\max(Neg(w_{A_1}),m_{C_1}(A_1)),\cdots,\max(Neg(w_{A_k}),m_{C_k}(A_k))\}\geqslant\alpha$ 时，也就是说 $\forall i,Neg(w_{A_i})\geqslant\alpha$ 或 $m_{C_i}(A_i)\geqslant\alpha$，$R$ 才会出现在结果中。若 $Neg(w_{A_i})\geqslant\alpha$，则不管 $m_{C_i}(A_i)$ 为何值，$\max(Neg(w_{A_i}),m_{C_i}(A_i))$ 都必定大等于 α。w_{A_i} 由用户在查询时给出，对于每条记录而言都相同，因此，$Neg(w_{A_i})$ 是已知的；而 $m_{Ci}(A_i)$ 则由每条记录的属性值决定，往往不相同。所以，先判断 $Neg(w_{A_i})$ 可以进行初步筛选，减少不必要的计算。只有当 $Neg(w_{A_i})<\alpha$ 时才需要对每条记录计算 $m_{C_i}(A_i)$。

计算 $m_{C_i}(A_i)$ 时也不必扫描所有的记录。因为只有 $m_{C_i}(A_i)\geqslant\alpha$ 的记录才可能出现在结果中，其余的记录是无关的。但是 $m_{C_i}(A_i)$ 和 α 都是语言值，不能直接去模糊。假设 $\alpha=s_p,p\in\{1,2,\cdots,T\}$，可知语言值标记 s_{p-1} 和 s_p 的交点为 λ_p（见图 1，其中横坐标 μ 为Ⅰ型模糊数隶属度，纵坐标 $m(\mu)$ 是 μ 关于语言标记对应的梯形模糊数的隶属数）。只有当记录

的属性 A_i 关于模糊概念 C_i 的隶属度不小于 λ_p 时，其语言值隶属度才可能为 s_p。对于那些关于 C_i 的数值隶属度大等于 λ_p 的记录，其由式(5)计算得到的语言值隶属度必定大等于 s_p。

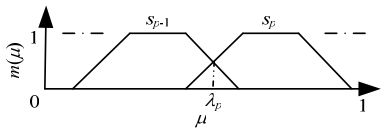

图 1　语言值标记 s_{p-1} 和 s_p 的交点

因此，可以利用 λ_p 对 C_i 的隶属函数去模糊，得到其 λ_p 截集 $[\underline{\beta_i},\overline{\beta_i}]$，只有在此范围内的记录才可能出现在结果中。根据上述讨论得到 DLWC 算法：

(1)设 $\alpha = s_p, p \in \{1,2,\cdots,T\}$，求出 s_p 与 s_{p-1} 的交点 λ_p。

(2)判断查询中的子条件“A_i is C_i”的权重 w_{A_i} 是否满足 $Neg(w_{A_i}) \geqslant \alpha$；不满足则转(3)，满足则转(4)。

(3)求 C_i 的 λ_p 截集 $[\underline{\beta_i},\overline{\beta_i}]$。

(4)若 $i < k$，i+1，转(2)，否则转(5)。

(5)将(2)～(4)中得到的 $[\underline{\beta_i},\overline{\beta_i}]$ 组合成精确的 SQL 语句。

(6)利用 SQL 语句筛选出记录集合 $set(R)$，对此集合中的每一条记录利用式(6)计算语言值匹配度。

(7)将匹配度大于阈值 α 的记录按降序输出。

注意当 $p=0$ 时，DLWC 算法退化成文献[8]中的算法，即只能扫描整个表来计算匹配度，得到查询结果。

例 2 用 DLWC 算法重新计算例 1

$\alpha = s_6$，根据表 1 可以计算出 s_5 和 s_6 的交点 $\lambda_6 = 0.4495$。年龄的权重为 s_6，而 $Neg(s_6) = s_{12-6} = s_6 \geqslant s_6$，所以不必计算各记录关于“age is young”的隶属度；height 的权重为 s_{12}，而 $Neg(s_{12}) = s_2 < s_6$，只有当 $m_{\text{tall}}(height) \geqslant s_6$ 时，记录才可能满足条件。根据式(9)求出“height is tall”的隶属函数的 λ_6 截集 [178.93,250] (设正常人的身高上限为 250 cm)，得到精确的 SQL 语句：

```
SELECT   name
FROM     Person
WHERE    height>=178.93
```

表 2 中在此范围内的记录为 $R3$, $R5$, $R8$, $R10$。根据式(6)对这 4 条记录计算匹配度，并按降序排列，结果见表 3。

表 3　例 2 的查询结果

记录号	age 的语言值隶属度	height 的语言值隶属度	语言值匹配度
$R8$	s_{12}	s_9	s_9
$R5$	s_{12}	s_8	s_8
$R3$	s_4	s_7	s_6
$R10$	s_5	s_8	s_6

4.2　DLWD(Derivation of LWD)算法

对式(7)进一步分析，当 $MD(R) = \max\{\min(w_{A_1}, m_{C_1}(A_1)),\cdots, \min(w_{A_k}, m_{C_k}(A_k))\} \geqslant \alpha$ 时，等价于 $\exists i$ 使得 $\min(w_{A_i}, m_{C_i}(A_i))$, $i \in \{1,2,\cdots,T\}$，R 才会出现在结果中。即 $\exists i$ 使得 $w_{A_i} \geqslant \alpha$ 且 $m_{C_i}(A_i) \geqslant \alpha$。因为 w_{A_i} 是已知的，若 $w_{A_i} < \alpha$，则无论 $m_{C_i}(A_i)$ 为何值，$\min(w_{A_i}, m_{C_i}(A_i))$ 都小于 α，此时不必再计算 $m_{C_i}(A_i)$；只有当 $w_{A_i} \geqslant \alpha$ 时，才需要计算 $m_{C_i}(A_i)$。与 4.1 节类似，采用去模糊化思想可以得到 DLWD 算法：

(1)设 $\alpha = s_p, p \in \{1,2,\cdots,T\}$，求出 s_p 与 s_{p-1} 的交点 λ_p。

(2)判断查询中的子条件“A_i is C_i”的权重 w_{A_i} 是否满足 $w_{A_i} \geqslant \alpha$；不满足则转(3)，满足则转(4)。

(3)求 C_i 的 λ_p 截集 $[\underline{\beta_i},\overline{\beta_i}]$。

(4)若 $i < k$，i+1，转(2)，否则转(5)。

(5)将(2)～(4)中得到的 $[\underline{\beta_i},\overline{\beta_i}]$ 组合成精确的 SQL 语句。

(6)利用 SQL 语句筛选出记录集合 $set(R)$，对此集合中的每一条记录利用式(7)计算语言值匹配度。

(7)将匹配度大于阈值 α 的记录输出。

注意当 $p=0$ 时，DLWD 算法退化成文献[8]中的算法。

5　结束语

本文在语言变量的基础上，将查询中的权重、记录的匹配度都用语言值表示，使得查询更加方便直观、贴近自然语言。并且利用模糊集合 α 截集去模糊的思想，将带有语言值权重的模糊查询语句转化为 RDBMS 能理解的精确 SQL 语句，从而利用 RDBMS 本身的机制进行记录的筛选，在一定程度上提高了查询的效率。今后将考虑对较复杂的模糊查询语句进行语言值表示的扩展。

参考文献

[1] Chen Shyiming, Jong W. Fuzzy Query Translation for Relational Database[J]. IEEE Transactions on Systems, 1997, 27(4): 714-721.

[2] Bosc P, Pivert O. SQLf: A Relational Database Language for Fuzzy Querying[J]. IEEE Transactions on Fuzzy Systems, 1995, 3(1): 1-17.

[3] Kantor P B. The Logic of Weighted Queries[J]. IEEE Transactions on Systems, 1981, 11(12): 816-821.

[4] Sanchez E. Importance in Knowledge Systems[J]. Information Systems, 1989, 14(6): 455-464.

[5] Zhang Yingchao, Chen Yifei, Ye Xiaoling, et al. Weighted Fuzzy Queries in Relational Database[C]//Proc. of the 2nd International Conference on Fuzzy Systems and Knowledge Discovery. Changsha, China: [s. n.], 2005: 430-440.

[6] 欧阳继红. 时空推理中一些问题的研究[D]. 吉林: 吉林大学, 2005.

[7] Viedma E H. An Information Retrieval Model with Ordinal Linguistic Weighted Queries Based on Two Weighting Elements[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based System, 2001, 9(9): 77-87.

[8] 陈逸菲, 张颖超, 叶小岭. 带语言值权重的模糊查询[J]. 计算机应用研究, 2005, 22(6): 73-75.

[9] Zadeh L A. The Concept of a Linguistic Variable and Its Application to Approximate Reasoning[J]. Information Science, 1975, 8(3): 199-249

[10] Chen Shyiming, Lin Yunshyang. A New Method for Fuzzy Query Processing in Relational Database System[J]. Cybernetics and Systems, 2002, 33(1): 447-482.

编辑　任吉慧