# 一种改进的基于语义距离的模糊数据库检索技术

## An Improved Retrieval Technique of Fuzzy Database Based on Semantic Distance

宣军英 （嘉兴学院 数学与信息工程学院 浙江 嘉兴 314001）

**摘 要:** 模糊数据库检索是信息检索技术的一个重要分支。本文将语义距离引入模糊数据库中，并且结合模糊逻辑理论，提出了一种改进的语义距离计算方法，构造了一种新的实现模糊数据库数据检索的框架。

**关键词:** 语义距离 模糊逻辑 模糊数据库 检索技术

数据库检索是信息管理系统的一个重要环节，传统数据库仅允许对精确的数据进行存储和处理，如在人员管理系统中，检索身高为 170cm 的男性，这里的“170”和“男”都为精确值。虽然 SQL 可以使用 like 函数和字符串匹配运算符增加一定的“模糊性”，但还是有其局限性。如在寻找一个人的时候，只知道他的大概情况，我们说“一个比较瘦的中等个子的年轻男子（大约 170cm）”，这里的“瘦”、“中等”、“年轻”都属于模糊数据，“比较”、“大约”是模糊词，这样的数据难以在传统人员数据库中找到。

要实现模糊数据库中数据的检索操作，主要有两种方法：一是依然建立传统关系型数据库，利用模糊理论的方法和工具对检索条件进行扩充，将 SQL 语言进行模糊扩展，形成精确检索所表示的模糊检索含义。这种方法虽然简单易现，但检索结果往往具有较强的主观性。二是利用模糊数学理论体系，用模糊数据来描述模糊事件并进行模糊运算，把不完全性、不确定性引入到数据库系统中形成模糊数据库，并建立相应的模糊检索语言。这种方法虽然相对计算较为复杂，但有实际使用意义。由于模糊数据库中两个模糊数据之间无法定义为相等或不相等，则可利用相似度、贴近度及语义距离来衡量模糊数据间的关系。本文使用语义距离来比较模糊数据间关系，构造了一种改进的语义距离计算公式，并应用到模糊数据库的检索中。

## 1 模糊逻辑理论及模糊数据的表示

### 1.1 模糊语言及其隶属函数

模糊语言 L 可形式化为 $L=\{U,T,E,N\}$，其中 U 为语言主题的全体，也称为模糊论域。模糊论域是一个闭区间表示的一个概念成立的范围，例如：成年人的身高论域 U 为 [100cm – 230cm]；T 是单词、句子的模糊集合，模糊集合由模糊子集构成，是对事物模糊程度标准的一种衡量。例如：描述身高的模糊集合 $T_h=\{$高，中等，矮$\}$；E 是表示术语的字母和符号构成的集合；N 是 T 对 U 的模糊关系，可以将 T 看成语言主题 U 的模糊子集。

模糊集合中的特征函数称为隶属函数，是描述有渐变性事物和现象的“中介过渡性”的关键，令 U 为论域，其上的一个模糊数据可以用模糊集合 A 的隶属函数 $\mu_A: U\rightarrow[0,1]$ 来表示。其中：对每一个论域 U 中的元素 x，$\mu_A(x)$ 表示 x 属于 A 的程度。隶属函数有：正态型、戒上型、戒下型。一般，描述模糊语气算子的隶属函数选用戒上型、戒下型，描述模糊化算子选用正态型。

例：人员信息表中，成年人身高论域 U 为 [100cm – 230cm]，则模糊数据身高的模糊概念“矮”、“中等”、“高个”的隶属函数（正态型）分别假设为：

$$\mu_{矮}(x)=\begin{cases}1 & x\le 150\\ [1+(\frac{x-150}{10})^2]^{-1} & x>150\end{cases} \quad (1)$$

$$\mu_{中等}(x)=\left[1+(\frac{x-170}{10})^2\right]^{-1} \quad x=170 \quad (2)$$

$$\mu_{高个}(x)=\begin{cases}1 & x\ge 180\\ [1+(\frac{x-180}{10})^2]^{-1} & x<180\end{cases} \quad (3)$$

隶属函数描述了一个特征隶属于某一个模糊概念的程度，隶属度越大，表明该特征隶属于该模糊概念的

程度越大。

例:当一个人的身高为 175cm 时,对应于个子“高”的隶属函数为 0.8,表示比较接近“高”;当身高为 180cm 时,对应个子高的隶属函数为 1,表示这个人是高个的可能性是 100%。

### 1.2 模糊语气算子

语气算子是模糊化程度的描述,如“极”,“相当”,“很”等,放在基本单词前面,可以修饰这些词的肯定程度。如“个子比较高”,“身高大概 170cm”等。其隶属函数可以采用以下形式:

$$\mu(x)=\begin{cases}1 & x\le y\\ [1+(\frac{x-y}{\alpha})^2]^{-\lambda} & x>y\end{cases} \qquad (4)$$

其中,$\alpha$ 为阀值,$\lambda$ 为语气算子对应的取值。用 $H_\lambda$ 作为语气算子来定量描述模糊值,设模糊值为 A,则定义 $H^\lambda$ 为 $H_\lambda=A_\lambda$,$\lambda$ 值的对应语义为“极”,$\lambda=4$;“很”,$\lambda=2$;“比较”,$\lambda=0.5$;“稍微”,$\lambda=0.25$。

例:我们说一个人个子很矮,其隶属函数表示为:

$$\mu_{很矮}(x)=\begin{cases}1 & x\le 150\\ [1+(\frac{x-150}{10})^2]^{-2} & x>150\end{cases} \qquad (5)$$

“大概”,“大约”,“近乎”等词称为模糊化算子,放在一个单词前面,而这个单词一般为一个明确的数值,一般用于修正确定词的模糊范围。当 U=(-∞,+∞)时,通常令:

$$\mu_E(x,y)=\begin{cases}e^{-(x-y)^2} & |x-y|\le\delta\\ 0 & |x-y|\ge\delta\end{cases} \qquad (6)$$

由于上面的公式并无反映参数 $\delta$ 的作用,参考中间型隶属函数的构造方法,对其进行修正,使其更加合理,调整后的隶属函数表示为:

$$\mu_E(x,y)=\begin{cases}e^{-(\frac{x-y}{\delta})^2} & |x-y|\le\delta\\ 0 & |x-y|\ge\delta\end{cases} \qquad (7)$$

### 1.3 模糊数据的表示方法

语义距离是指计算两个模糊数据间距离时,考虑模糊数据在客观世界中所表示的具体语义,由于模糊数据表示方法有多种形式,为保证描述同一属性的模糊数据间求得距离的一致性,可以用模糊区间数、模糊中心数、隶属函数等不同的方法来定义语义距离,本文用隶属函数来表示。

## 2 模糊数据库检索的框架构建

### 2.1 用隶属函数表示的模糊数据间的语义距离

定义 1. 当模糊数据表示为论域 U 上的隶属函数时,两个模糊数据 A 和 B 之间的语义距离 SD(A,B)定义为两个隶属函数 $\mu_A(x)$ 和 $\mu_B(x)$ 之差的切氏范数 $\|\mu_A(x)-\mu_B(x)\|$,即:

$$SD(A,B)=\max_{x\in U}|\mu_A(x)-\mu_B(x)| \qquad (8)$$

假设有两个模糊数据 A 和 B,分别用隶属函数表示为:

$$A=\frac{0.1}{10}+\frac{0.5}{11}+\frac{0.8}{12}+\frac{1}{13}+\frac{0.8}{14}+\frac{0.5}{15}+\frac{0.1}{16}$$

$$B=\frac{0.2}{9}+\frac{0.6}{10}+\frac{0.9}{11}+\frac{1}{12}+\frac{0.9}{13}+\frac{0.6}{14}+\frac{0.2}{15}$$

则 SD(A,B)=0.5。

### 2.2 基于区域包含关系的语义距离计算公式

现实问题中,由于模糊数据之间的语义关系较为复杂,在处理问题时需要根据模糊数据所要表达的语义关系来设计更为自然、恰当的语义距离计算公式。下面通过实例给出一种借助图形手段描述模糊数据表示的语义区域,并利用之间的包含关系获得语义距离计算公式的方法。

假设在人员模糊数据库中,有两个表示身高的模糊数据 S1,S2,值分别为“大约 177cm 到 180cm 之间”,“约 183cm 左右”,求两个数据的隶属函数。

根据隶属函数

$$\mu_{高个}(x)=\begin{cases}1 & x\ge 180\\ [1+(\frac{x-180}{10})^2]^{-1} & x<180\end{cases}$$

表示模糊数据“高个”,若取有限论域 U={175,…185},模糊数据“高个”又可表示为:

$$\mu_{高个}(x)=\frac{0.80}{175}+\frac{0.86}{176}+\frac{0.92}{177}+\frac{0.96}{178}+\frac{0.99}{179}+\frac{1.00}{180}+\frac{1.00}{181}+\frac{1.00}{182}+\frac{1.00}{183}+\frac{1.00}{184}+\frac{1.00}{185};$$

同理模糊数据 S1 的隶属函数表示为:

$$\mu_{S1}(x)=\frac{0.96}{175}+\frac{0.99}{176}+\frac{1.00}{177}+\frac{1.00}{178}+\frac{1.00}{179}+\frac{1.00}{180}+\frac{0.99}{181}+\frac{0.96}{182}+\frac{0.92}{183}+\frac{0.86}{184}+\frac{0.80}{185};$$

模糊数据 S2 的隶属函数为：

$$\mu_{S2}(x)=\frac{0.61}{175}+\frac{0.67}{176}+\frac{0.74}{177}+\frac{0.80}{178}+\frac{0.86}{179}+\frac{0.93}{180}+\frac{0.96}{181}+\frac{0.99}{182}+\frac{1.00}{183}+\frac{0.99}{184}+\frac{0.96}{185};$$

若将模糊数据 S1、S2 和“高个”放入以论域 U 为横坐标，以隶属度 μ 为纵坐标的坐标轴中，可以得到如下图所示的结果。

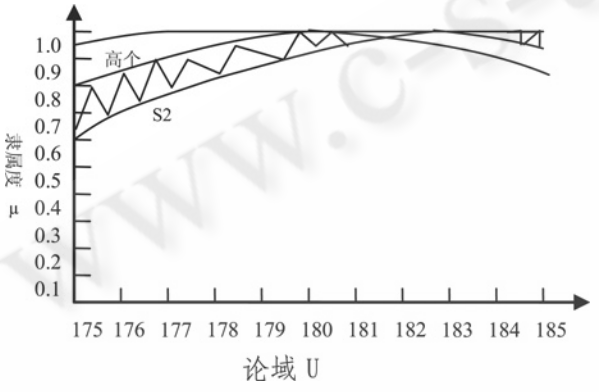


图 1　基于区域包含关系的语义距离计算

由图 1 加曲线型部分为模糊数据 S2 与模糊数据“高个”交叉区域，“高个”与模糊数据 S1 之间的空白部分为 S1 与“高个”交叉部分，从交叉部分面积大小可知模糊数据 S2 比 S1 表示的身高范围更接近个子高。显然，这种利用图形区域包含关系来获得语义距离计算公式的方法非常直观、自然，由图可以推断出：模糊数据图形表示区域间包含程度越大，相应的语义距离越小。

下面给出语义距离计算公式。

定义 2.　当模糊数据表示为论域 U 上的隶属函数时，两个模糊数据 A 和 B 之间的语义距离 SD(A,B) 定义为两个隶属函数 $\mu_A(x)$ 和 $\mu_B(x)$ 在以论域 U 为横坐标，以隶属度 μ 为纵坐标的坐标轴中图形区域交叉部分的比值，即

$$SD(A,B)=1-\frac{\sum_{i=1}^{n}|\mu_A(x_i)-\mu_B(x_i)|}{\sum_{i=1}^{n}\mu_B(x_i)} \quad (9)$$

根据公式，可以分别计算模糊数据 S1 和 S2 与模糊数据“高个”间的语义距离。

SD(S1,高个) = 0.98；SD(S2,高个) = 0.92

### 2.3　模糊数据检索命令 FSELECT

可以利用上述语义距离计算公式定义模糊数据关系运算符“模糊等于”和“模糊不等于”。

定义 3　已知 X 和 Y 是论域 U 上的两个模糊数据，SD(X,Y) 是两者间的语义距离，ε (ε > 0) 是一个给定的实数小量，则 X 与 Y 之间的模糊关系运算“模糊等于”和“模糊不等于”定义为：

当 SD(X,Y) ≤ ε 时，认为 X 模糊等于 Y，记为 $X \overset{\varepsilon}{=} Y$；当 SD(X,Y) > ε 时，认为 X 模糊不等于 Y，记为 $X \overset{\varepsilon}{\neq} Y$。

由此可得模糊检索 FSELECT 语句。

FSELECT〈结果表〉

FROM〈模糊关系表〉

WHERE [(模糊检索条件)，阈值 α] [AND /OR] [一般检索条件]

其中：<检索条件>是由模糊表达式和条件表达式构成的混合表达式。模糊表达式由模糊关系运算符（如“模糊等于” $\overset{\varepsilon}{=}$、“模糊不等于” $\overset{\varepsilon}{\neq}$）和模糊逻辑运算符（如“模糊与” ∩、“模糊或” ∪）组成，用于模糊数据库中属性值模糊的字段；条件表达式由传统关系运算符（如“等于” =、“不等于” ≠）和逻辑运算符（如：“与” ∧、“或” ∨）组成，用于模糊数据库中属性值精确的字段。

## 3　模糊数据检索实例

### 3.1　传统模糊检索

目前常用的数据库是一种支持关系模型的数据库系统。在关系型数据库中，数据信息被组织成若干张关系二维表的形式，信息检索大多采用结构化查询语言 SQL。如某市暂住人口管理中，部分数据如下表所示：

表 1　人员基本信息表

| 身份证号 | 姓名 | 暂住地 | 性别 | 身高 | 体重 |
|---|---|---|---|---|---|
| 330402198604210612 | 李小迪 | 建设小区 1－101# | 男 | 166cm | 约 66 公斤 |
| 440401197007120214 | 李小童 | 文韦路 275# | 男 | 169cm | 70 公斤 |
| 440304197612030618 | 张丽利 | 文韦路 270# | 女 | 160cm | 52 公斤 |
| 440304198210030613 | 李利平 | 文韦路 370# | 男 | 167cm | 70 公斤 |
| 320402196802160215 | 钟平 | 文秀里 3－120# | 男 | 185cm | 80 公斤 |
| 330302196709121616 | 李强 | 望江门 4－105# | 男 | 170cm 到 175cm | 74 公斤 |
| 330402197004210618 | 赵晓欧 | 建设小区 3－103# | 男 | 中等 | 约 55 公斤 |

假如要检索身高不超过 175 cm，体重为 70 公斤姓“李”的男性的身份证号，姓名，暂住地，可以使用标准 SQL 语句实现传统模糊检索：SELECT 身份证号，姓名，暂住地 FROM RYXXB WHERE 姓名 LIKE‘李＊’AND 身高＜＝175 AND 体重＝70 AND 性别＝’男’。笔者在 Visual Foxpro6.0 环境中编程实现了此方法，检索结果界面如图 2。

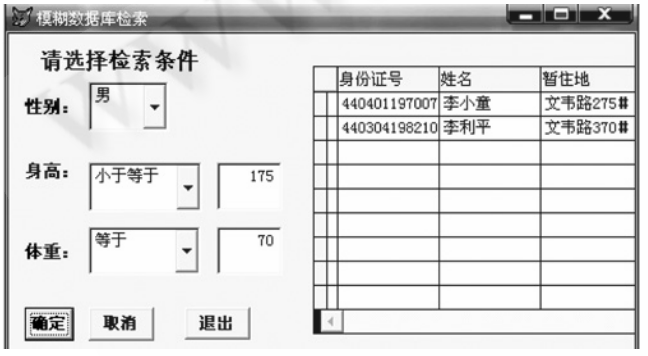


图 2　传统模糊检索界面

## 3.2　基于语义距离的模糊检索

基于模糊语义距离的模糊数据库检索的基本步骤包括：

(1) 检索条件的模糊化处理，将用户给定的检索条件依据模糊隶属函数进行模糊化。

(2) 计算数据库中记录的相关属性与检索条件的模糊语义距离。

(3) 根据比较的模糊相似程度大小，将结果返回给用户。

假设现要检索身高大约在 170cm 到 175cm，体重大约为 70 公斤的男性的身份证号，姓名，暂住地。令实数小量 $\varepsilon$，阀值 $\alpha$ 均为 0.5，可以用模糊检索：

FSELECT 身份证号，姓名，暂住地

FROM RYXXB

WHERE 身高$^{0.5}$＝170 到 175 ∩体重$^{0.5}$＝约 70 左右

AND 性别＝’男’，0.5

要实现模糊检索，必须将模糊条件转换为精确条件。为实现这一功能，本文在数据表中根据语义距离的计算增加一个模糊总隶属度字段。当模糊字段不止一个时，通过计算各自的模糊语义距离，再根据模糊集合中的交运算得到总模糊隶属度，上述模糊 SQL 就可以转化为：

SELECT 身份证号，姓名，暂住地

FROM RYXXB

WHERE 性别＝’男’AND 模糊总隶属度＞＝0.5

通过在程序中增加模糊处理，得到了改进的模糊信息检索结果，实现界面如图 3。


图 3　改进的模糊信息检索界面

可见，传统模糊检索思路简单，检索结果有较大的局限性。而基于语义距离的模糊检索虽然在检索时增加了一定的复杂度，但结果更符合现实世界的客观要求，更有实际使用意义。

（下转第 26 页）

(上接第36页)

## 4 结束语

**在模糊数据库中实现数据检索是一个较为复杂的问题,本文在充分考虑模糊数据的具体语义基础上,通过模糊语义距离的计算,构造了模糊检索语句 FSELECT 的框架,并给出了一个在模糊数据库中实现数据检索的实例。通过本文的研究力求为下一步模糊数据挖掘技术的研究做好准备。**

## 参考文献


1 何新贵.特种数据库技术.北京:科学出版社,2000.

2 窦振中.模糊控制技术及应用.北京:北京航空航天大学出版社,1995.

3 Liu Wei - Yi. The fuzzy functional dependency on the basis of the semantic. Fuzzy sets and Systems, 1993, 59:12 - 15.

4 朱蓉.基于模糊理论的查询技术研究.计算机应用研究,2003,(5):8 - 10.

5 B P Buckles, F E Peter. Fuzzy Databases in the New Era. Proceedings of the 1995 ACM Symposium, February 1995.

6 朱蓉.基于区域包含的语义距离构造及其在模糊数据库中的应用.吉林师范大学学报(自然科学版),2005,(2):22 - 24.

7 张李义.基于模糊语义距离的多媒体信息检索方法研究.情报学报,2003,22(2):133 - 135.

8 郭富强,鱼滨.基于模糊数据库的数据查询研究.微电子学与计算机,2005(9):123 - 126.

9 邓方安,刘晓冀.基于语义距离的模糊信息分类方法.苏州科技学院学报,2004,(9):5 - 9.