

一种基于正态分布密度函数的模糊查询方法

李 雪

(南京航空航天大学 计算机科学与技术学院,江苏 南京 211100)

摘 要:在数据库中进行信息查询时,用户经常需要表示不精确的查询请求。然而传统的数据库无法对这些不精确的查询条件进行匹配,导致查询到空的结果集或者查询结果过多从而难以筛选,难以满足用户实际要求。在关系数据库中进行模糊查询已经进行了大量研究,其中大部分是对不同的模糊集设置不同的隶属函数来进行查询,将其应用于大样本数据时便会遇到很多困难。根据模糊集理论以及正态分布函数适合于大样本数据的特征,文中用正态分布密度函数来一般化隶属函数,使其可以自动对模糊集合进行区间匹配,得到对应的精确的区间,从而实现满足用户需求的模糊查询结果,最后结合实例进行演算并对结果进行分析。结果表明,该方法减少了人们设置隶属函数时的个人主观性,提高了匹配结果的准确性。

关键词:模糊查询;大样本数据;正态分布;密度函数;隶属函数;模糊集合;区间匹配

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2018)09-0001-06

doi:10.3969/j.issn.1673-629X.2018.09.001

A Fuzzy Matching Method Based on Normal Distribution Density Function

LI Xue

(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
Nanjing 211100, China)

Abstract: When people query information in a database, users often need to represent imprecise query requests. However, the traditional database can't process the imprecise query conditions, leading to the empty query result set or numerous query results, which is difficult to filter and meet the user's actual requirements. Many research have been proposed for dealing with fuzzy data and queries in the relational database, most of them are to set different membership functions for different fuzzy sets, and applying it to large sample data will encounter many difficulties. According to fuzzy set theory and characteristics of the normal distribution function suitable for large sample data, we adopt the normal distribution density function to generalize the membership functions, which can perform interval matching for fuzzy sets automatically and obtain the corresponding exact interval of the fuzzy sets. In this way, the fuzzy query result that meets the user's demand is realized, and the result is analyzed by combining with the example. It is showed that the method can reduce the individual subjectivity when setting the membership function, and improve the accuracy of matching results.

Key words: fuzzy query; large sample data; normal distribution; density function; membership function; fuzzy set; interval matching

0 引 言

日常生活中可以见到很多模糊现象,比如有一个古老的希腊悖论是这样说的:一颗豌豆不叫一堆,两颗也不叫,这是大家都公认的,但是可不可以说 10 000 颗豌豆不叫一堆而 10 001 颗叫一堆呢?肯定不能这样定义,但是适当的界限在哪里?这是一个渐进的过程,没有明确的范围,不能用精确的数值进行描述。然而,研究者们一直忽略这些不确定性,并且所有的数据

库管理系统都是基于精确的数据。直到 20 世纪 60 年代,Zadeh 教授发起了关于模糊性的研究。

查询操作是数据库的重要操作之一,而现有的绝大多数查询操作都要求指定某个属性位于某个具体区间的所有记录。但是可以发现,在日常生活中,人们想要得到某种结果时,并不能很明确地表达他们的查询要求。比如某个人需要租房子,需要找到价格便宜并且离市中心近的房子,当在数据库中进行查询时,“离

收稿日期:2017-10-16

修回日期:2018-02-27

网络出版时间:2018-05-16

基金项目:国家自然科学基金(61370075)

作者简介:李 雪(1991-),女,研究生,CCF 会员(E200054521G),研究方向为 NoSQL 图数据库、模糊查询。

网络出版地址:<http://jns.cnki.net/kcms/detail/61.1450.TP.20180515.1655.044.html>

市中心近”以及“价格便宜”这样的查询语句就是模糊的概念。但是一般的数据库管理系统并不能对这样的模糊语句进行查询,在网络规模不断扩大的今天,数据库的查询次数以指数级别在增长,精确的查询则要求数据库的操作者对数据库中的数据有很清楚的了解,例如 DBA(database administrator,数据库管理员)。这直接使得对数据库内容不熟悉的人很难去获得其想要的查询结果,其大概率查询到空的结果集或者过多的数据记录从而难以筛选。而模糊查询则为该问题提供了解决思路。然而不论是传统的关系数据库还是图数据库都只能对精确的条件进行查询,无法查询形如上文提到的“价格便宜并且离市中心近的房子”这样带有模糊语句的查询条件,因此急需一种合适有效的方法将模糊的查询条件转化成精确的查询区间,使其可以在数据库管理系统中执行。

1 相关研究工作

1.1 模糊查询

1.1.1 关系数据库模糊查询

国内外对模糊查询进行了大量研究,例如,文献[1-4]均说明了如何使用正向法和反向法对查询条件进行处理,将其转化成精确的查询区间;文献[5]定义了一种模糊数据库查询语言,它可以处理数据库中遇到的模糊的或者精确的查询条件;文献[6]设计了一个样本数据库,可以在上面进行模糊查询或者精确查询,对比传统数据库,在样本数据库上进行模糊查询的时间花费大大降低;文献[7-8]在现有关系数据库上对查询语言进行模糊扩展,同时引入了权重的概念。

1.1.2 云数据库模糊查询

关于 NoSQL 模糊查询,国内并没有相关研究,国际上此类研究也并不是很多。文献[9]对 NoSQL 图数据库的 Cypher 语言进行扩展得到 Cypherf 语言使其可以进行模糊查询;文献[10]提出了一个基于图的模糊 NoSQL 模型,称 FNoSQL,用来处理大数据的同时也扩展了 NoSQL;文献[11]提出使用合适的工具用来在 Neo4j 的 cypher 语言中表达模糊查询,用模糊性来帮助解决模糊模式匹配。

文献[12]实现了可视化模糊查询生成器 Ask-Fuzzy,它是前端的可视化界面和后端的数据库之间的一个中间层,可以在 where 中基于概念而不是数据来使用模糊谓词进行查询。

文献[13-14]提出了一个可以灵活查询模糊数据库的系统 SUGAR,该系统建立在 Neo4j 图数据库管理系统中,支持 Cypher 查询,由 transcriptor 模块和分数计算器模块组成,并且引进了 Fudge 语言来实施,该语言可以灵活处理模糊的或者精确的数据,可以用来处

理模糊图数据库中的偏好查询。

1.2 相关理论知识

生活中经常会有很多无法用精确的数值表达的不确定性,比如人们交谈时的口语,各种形容词以及副词,各个领域主观的意见以及判断等等,这些都是模糊性表达。所有的模糊性都是一种主观的衡量,需要用隶属函数对这些模糊性进行转换,明确这些模糊性的隶属规律,找到合适的隶属函数,这样就可以在充满不确定性的现实生活中和必须使用精确数据进行计算的数学中找到一个很好的平衡点^[15-16]。

1.2.1 模糊集理论

(1) 模糊集。

模糊集 A 可以用元素 x 和它的隶属函数 $\mu_A(x)$ 的集合表示,其中 x 是论域 U 中的一个元素,也就是:

$$A = \{ (x, \mu_A(x)) \mid x \in U \}$$

一般可以通过模糊集合所对应的隶属函数来对模糊集合进行划分,因此如果知道论域 U 上的模糊集合,就能通过模糊集合中的元素 x 来确定元素 x 所对应的隶属函数 $\mu_A(x)$ 。

(2) α 截集。

α 截集可以认为是论域 U 中所有元素 x 所对应的隶属函数 $\mu_A(x)$ 的值都不小于 α 的一个集合:

$$A_\alpha = \{ x \mid \mu_A(x) \geq \alpha, \forall x \in U \}$$

其中, α 为置信水平(阈值)。

1.2.2 正态分布

如果有随机变量 X 服从参数为 $\mu, \sigma (\sigma > 0)$ 的概率分布,那就称随机变量 X 服从正态分布,记为 $X \sim N(\mu, \sigma^2)$,它的概率密度函数表示为:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

图 1 是一个正态分布概率密度函数。从图中可以得知,平均数周围的数的取值概率明显大于两边的,通常生活中很多的大样本数据都服从正态分布,比如学生的成绩大都比较集中,中等的偏多,成绩特别好或者特别差的人总是占少数的;人的身高在某个中等范围的占大多数,特别高的人以及特别矮的属于少数等。基于这个特点,考虑使用正态分布密度函数作为隶属函数对模糊属性进行区间匹配。

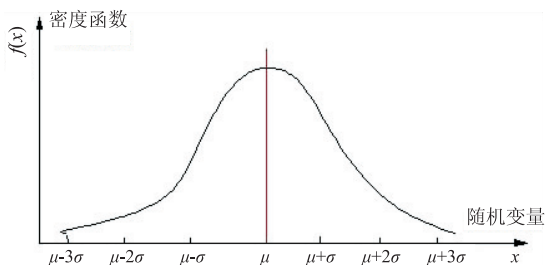


图 1 正态分布密度函数

2 基于正态分布密度函数的模糊化匹配方法

传统的模糊化查询主要是采用 Zadeh 在 1965 年提出的模糊数学理论,通过对论域中不同模糊集设置相应隶属函数来计算某具体数值的隶属度,从而进行查询。然而在实际的图像数据库中,采用预设的隶属函数会面临诸多问题,如:

- (1) 很难对所有的模糊集都设置一个相对较为准确的隶属函数;
- (2) 忽视数据库中已有样本的信息利用;
- (3) 对数据库中节点作属性增加的代价较大;
- (4) 很难确定最终的匹配区间。

为了将模糊化查询应用于图形数据库,文中提出了一种基于正态分布密度函数的模糊化自动匹配方法,克服了上述问题,并结合实例作了性能分析。

设某一类节点 N 拥有 k 个实体,即: $\{ N \} = \{ n_1, n_2, \dots, n_k \}$ 。

其某个属性 $attr$ 是连续的,第 i 个实体对应 $attr$ 的值用 $n_i \cdot attr$ 表示,一般地,可以预计算 $\{ N \}$ 在属性 $attr$ 的均值 $E(N)$ 与方差 $D^2(N)$,即

$$E(N, attr) = (n_1 \cdot attr + n_2 \cdot attr + \dots + n_k \cdot attr) / k$$
$$D^2(N, attr) = ((n_1 \cdot attr - E(N, attr))^2 + (n_2 \cdot attr - E(N, attr))^2 + \dots + (n_k \cdot attr - E(N, attr))^2) / k$$

则假设 $\{ N \}$ 的 $attr$ 属性值服从正态分布,即 $N.attr \sim N(E(N, attr), D^2(N, attr))$ 。为了简便,所有的 $E(N, attr)$ 用 E 表示, $D(N, attr)$ 用 D 表示。其概率密度函数为:

$$f(x) = \frac{1}{\sqrt{2\pi}D} \exp\left(-\frac{(x-E)^2}{2D^2}\right)$$

由概率论可知,函数 $f(x)$ 的线下面积为 1,其定义域为 $(-\infty, +\infty)$ 。

根据具体模糊集的内容对定义域进行划分。一般地,对于模糊集 A 其均为论域 U 的子集,根据具体的 A 可将论域进行 t 个划分,即:

$$d(U) = \{d_1, d_2, \dots, d_t\}$$

且 A 属于第 s 个划分,即 $A = ds$ 。例如:

如果 U 是年龄 $[0, 100]$,可将其分为“年幼”、“年轻”、“中年”、“老年”4 个划分,“年轻”为第 2 个划分;

如果 U 是考试成绩 $[0, 100]$,可将其分为“差”、“普通”、“优异”3 个划分,“普通”是第 2 个划分。

论域的每个划分 d_i 都可以赋之一个权重 w_i ,满足: $w_1 + w_2 + \dots + w_t = 1$ 。

为了描述方便,设所有的权重是一样的,即

$$w_i = \frac{1}{t} \quad (1 \leq i \leq t)$$

给出模糊区间的定义:

若模糊集 A 可将论域进行 t 个划分,且 A 属于第 s 个划分,则 A 的模糊区间表示为:

$$f_{\text{area}}(A) = [f_{\min}(A), f_{\max}(A)]$$

若 $s = 1$,则 $f_{\min}(A) = -\infty$;若 $s = t$,则 $f_{\max}(A) = +\infty$ 。

若 A' 属于第 $s+1$ 个划分,则 $f_{\max}(A) = f_{\min}(A')$,且满足:

$$\int_{f_{\min}(A)}^{f_{\max}(A)} f(x) dx = w_i$$

令 $Y = (X - E) / D$,则 Y 服从标准正态分布:

$$\int_{\frac{f_{\min}(A) - E(N, attr)}{D(N, attr)}}^{\frac{f_{\max}(A) - E(N, attr)}{D(N, attr)}} f'(x) dx = w_i$$

设 $\Phi(x)$ 为标准正态分布的累积分布函数,即:

$$\Phi(x) = \int_{-\infty}^x f'(t) dt$$

且 $\Phi(x)$ 在定义域 R 上单调递增。

则有:

$$\Phi\left(\frac{f_{\min}(A) - E}{D}\right) = \int_{-\infty}^{\frac{f_{\min}(A) - E}{D}} f'(t) dt = \frac{s - 1}{t}$$
$$\Phi\left(\frac{f_{\max}(A) - E}{D}\right) = \int_{-\infty}^{\frac{f_{\max}(A) - E}{D}} f'(t) dt = \frac{s}{t}$$

若 $\Phi(x)$ 的反函数为 $\Phi^{-1}(x)$,则有:

$$f_{\min}(A) = \Phi^{-1}((s - 1)/t) \cdot D + E$$
$$f_{\max}(A) = \Phi^{-1}(s/t) \cdot D + E$$

一般地,函数 $\Phi(x)$ 的定义域为 $[0, +\infty)$,值域为 $[0.5, 1)$,且 $\Phi(-x) = 1 - \Phi(x)$ 。有 $\Phi^{-1}(x)$ 的定义域为 $[0.5, 1)$,值域为 $[0, +\infty)$ 。由 $\Phi(-x) = 1 - \Phi(x)$,有 $-\Phi^{-1}(x) = \Phi^{-1}(1 - x)$ 。

所以上式可化为:

$$f_{\min}(A) = \begin{cases} \varphi^{-1}\left(\frac{s-1}{t}\right) \cdot D + E, & s \geq \frac{1}{2}t + 1 \\ \varphi^{-1}\left(\frac{t-s+1}{t}\right) \cdot D + E, & s \neq 1 \text{ and } s < \frac{1}{2}t + 1 \\ -\infty, & s = 1 \end{cases}$$
$$f_{\max}(A) = \begin{cases} \varphi^{-1}\left(\frac{s}{t}\right) \cdot D + E, & s \neq t \text{ and } s \geq \frac{1}{2}t \\ \varphi^{-1}\left(\frac{t-s+1}{t}\right) \cdot D + E, & s < \frac{1}{2}t \\ +\infty, & s = t \end{cases}$$

根据上式可以计算 $f_{\min}(A)$ 和 $f_{\max}(A)$ 。为了避免对“无穷”的计算操作,利用论域 $U = [U_{\min}, U_{\max}]$ 的上下限对正负无穷进行取代,即:

$$f_{\min}(A) =$$

$$\begin{cases} \varphi^{-1}(\frac{s-1}{t}) \cdot D + E, s \geq \frac{1}{2}t + 1 \\ \varphi^{-1}(\frac{t-s+1}{t}) \cdot D + E, s \neq 1 \text{ and } s < \frac{1}{2}t + 1 \\ U_{\min}, s = 1 \end{cases}$$
$$f_{\max}(A) = \begin{cases} \varphi^{-1}(\frac{s}{t}) \cdot D + E, s \neq t \text{ and } s \geq \frac{1}{2}t \\ \varphi^{-1}(\frac{t-s+1}{t}) \cdot D + E, s < \frac{1}{2}t \\ U_{\max}, s = t \end{cases}$$

对区间 $f_{\text{area}}(A)$ 进行求平均操作,即均值:

$$\text{ave} = \frac{1}{f_{\max}(A) - f_{\min}(A)} \int_{f_{\min}(A)}^{f_{\max}(A)} f(x) dx = \frac{1}{t[f_{\max}(A) - f_{\min}(A)]}$$

定义:

若 $s = (t + 1)/2$, 则 $E(N, \text{attr})$ 为模糊集 A 的隶属函数的极大值点 $\max(A)$; 否则, 均值 ave 与正态分

$$A_1(x) = \begin{cases} 1, U_{\min} \leq x < \max_1 \\ \cos[\frac{\pi/3}{\max_2 - \max_1}(x - \max_1)], \max_1 < x \leq \max_2 \\ \frac{1}{2} - \sin[\frac{\pi/6}{\max_3 - \max_2}(x - \max_2)], \max_2 < x \leq \max_3 \\ 0, \max_3 < x \end{cases}$$
$$A_i(x) = \begin{cases} 0, U_{\min} \leq x < \max_{i-2} \\ 1 - \cos[\frac{\pi/3}{\max_{i-1} - \max_{i-2}}(x - \max_{i-2})], \max_{i-2} \leq x < \max_{i-1} \\ \sin(\frac{x - \max_{i-1}}{\max_i - \max_{i-1}} * \pi/3 + \pi/6), \max_{i-1} \leq x < \max_i \\ 1, \max_i < x \leq U_{\max} \end{cases}$$

若 $2 \leq i \leq t-1$, 则

$$A_i(x) = \begin{cases} 0, U_{\min} \leq x < \max_{i-2} \\ 1 - \cos[\frac{\pi/3}{\max_{i-1} - \max_{i-2}}(x - \max_{i-2})], \max_{i-2} \leq x < \max_{i-1} \\ \sin(\frac{x - \max_{i-1}}{\max_i - \max_{i-1}} * \pi/3 + \pi/6), \max_{i-1} \leq x < \max_i \\ \cos[\frac{\pi/3}{\max_{i+1} - \max_i}(x - \max_i)], \max_i < x \leq \max_{i+1} \\ \frac{1}{2} - \sin[\frac{\pi/6}{\max_{i+2} - \max_{i+1}}(x - \max_{i+1})], \max_{i+1} < x \leq \max_{i+2} \\ 0, \max_{i+2} < x \leq U_{\max} \end{cases}$$

至此完成了基于正态分布的模糊集对应隶属函数的自动生成。设置阈值 ε , 令 $A_i(x) > \varepsilon$, 得到的不等式的解即为 $A_i(x)$ 所对应模糊集的区间匹配。

3 实例分析

第二节叙述的方法可以用在很大的数据量上, 利

布曲线的交点的横坐标为模糊集 A 的隶属函数的极大值点 $\max(A)$ 。

若 $f(x)$ 在 $x \geq E(N, \text{attr})$ 区间的反函数为 $f^{-1}(x)$, 则有:

$$\max(A) = \begin{cases} 2E - f^{-1}(\text{ave}), s \leq \lfloor \frac{t}{2} \rfloor \\ E(N, \text{attr}), s = \frac{t+1}{2} \\ f^{-1}(\text{ave}), s > \lfloor \frac{t+3}{2} \rfloor \end{cases}$$

其中, $f^{-1}(x) = E + \sqrt{-2D^2 \ln(\sqrt{2\pi} D \cdot x)}$

求得模糊集 A 的隶属函数的极大值点 $\max(A)$ 。若论集 U 共有 t 个划分, 其每个划分的模糊集 A_1, A_2, \dots, A_t 对应的极大值点分别为 $\max_1, \max_2, \dots, \max_t$ 。不妨令 $\max_0 = U_{\min}$, $\max_{t+1} = U_{\max}$, 则用正、余弦曲线描述每个模糊集的隶属函数:

用正态分布函数直接计算出每个模糊集对应的区间匹配。为了方便起见, 选择一个班的成绩作为研究对象, 分别计算出成绩偏低、中等以及优秀三个模糊集的隶属函数, 设置合适的 ε 值, 得到区间匹配。

3.1 区间匹配实现一般流程

图 2 描述了模糊集区间匹配的一般实现过程。

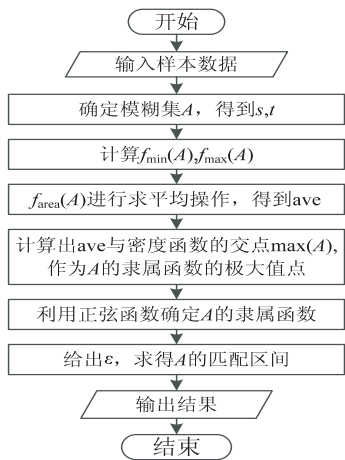


图2 模糊集区间匹配的实现过程

3.2 区间匹配计算过程

(1)确定模糊集以及 s, t 。

由表1 计算出学生成绩的平均值 $E \approx 77.6$, 标准差 $D \approx 15.3$, 方差 $D^2 \approx 234$ 。将学生成绩 $[0, 100]$ 划分为“低”、“中等”、“优异”三个模糊区间, 由此可知 $t = 3, s$ 取 $1, 2, 3$ 。

表1 学生成绩

| 编号 | 成绩 | 编号 | 成绩 | 编号 | 成绩 |
|----|----|----|-----|-----|------|
| 1 | 88 | 12 | 100 | 23 | 78 |
| 2 | 87 | 13 | 44 | 24 | 83 |
| 3 | 65 | 14 | 35 | 25 | 78 |
| 4 | 67 | 15 | 87 | 26 | 98 |
| 5 | 89 | 16 | 67 | 27 | 92 |
| 6 | 82 | 17 | 68 | 28 | 67 |
| 7 | 56 | 18 | 69 | 29 | 65 |
| 8 | 77 | 19 | 86 | 30 | 75 |
| 9 | 87 | 20 | 67 | 标准差 | 15.3 |
| 10 | 90 | 21 | 89 | 平均数 | 77.6 |
| 11 | 98 | 22 | 94 | | |

(2)计算 $f_{\min}(A)$ 、 $f_{\max}(A)$ 。

当 $s = 1$ 时, $f_{\min}(A) = -\infty$ 。因为 $\Phi(x) = \int_{-\infty}^x f(t) dt$ 是标准正态分布函数的累积分布函数, 查正态分布表可知 $\varphi^{-1}(\frac{2}{3}) = 0.43$, 则

$$f_{\max}(A) = -\varphi^{-1}(\frac{t-s}{t}) * D + E = -\varphi^{-1}(\frac{2}{3}) * 15.3 + 77.6 \approx 71$$

当 $s = 2$ 时, 因为若 A' 属于第 $s + 1$ 个划分, 则 $f_{\max}(A) = f_{\min}(A')$, 所以

$$f_{\min}(A) = 71$$
$$f_{\max}(A) = -\varphi^{-1}(\frac{s}{t}) * D + E = \varphi^{-1}(\frac{2}{3}) * 15.3 + 77.6 \approx 84.2$$

当 $s = 3$ 时, $f_{\min}(A) = 84.2, f_{\max}(A) = +\infty$ 。

根据上式可以计算出 $f_{\min}(A)$ 和 $f_{\max}(A)$ 。

为了避免对“无穷”的计算操作, 利用论域 $U = [U_{\min}, U_{\max}]$ 的上下限对正负无穷进行取代, 即

当 $s = 1$ 时, $f_{\min}(A) = 0, f_{\max}(A) = 71$;

当 $s = 2$ 时 $f_{\min}(A) = 71, f_{\max}(A) = 84.2$;

当 $s = 3$ 时, $f_{\min}(A) = 84.2, f_{\max}(A) = 100$ 。

(3)根据 $f_{\min}(A)$ 、 $f_{\max}(A)$ 计算 ave 。

$$s = 1 \text{ 时, } \text{ave} = \frac{1}{f_{\max}(A) - f_{\min}(A)} \int_{f_{\min}(A)}^{f_{\max}(A)} f(x) dx = \frac{1}{t[f_{\max}(A) - f_{\min}(A)]} = \frac{1}{3} * \frac{1}{71} = \frac{1}{213}。$$

$s = 2$ 时, $\text{ave} = (1/3) * (1/13.2) = 1/39.6$ 。

$s = 3$ 时, $\text{ave} = (1/3) * (1/15.8) = 1/47.40$ 。

(4) 计算模糊集的隶属函数的极大值点 $\max(A)$ 。

$s = 1$ 时, $f^{-1}(x) = E + \sqrt{-2D^2 \ln(\sqrt{2\pi D} * \text{ave})} \approx 105.93, \max(A) = 2E - f^{-1}(x) \approx 49.3$ 。

$s = 2$ 时, $\max(A) = E(N, \text{attr}) = 77.6$ 。

$s = 3$ 时, $\max(A) = f^{-1}(x) = E +$

$$\sqrt{-2D^2 \ln \sqrt{2\pi D} * \text{ave}} \approx 87.6。$$

(5)确定 A 的隶属函数。

模糊集 $A_1、A_2、A_3$ 对应的极大值点分别为 $\max_1 = 49.3, \max_2 = 77.6, \max_3 = 87.6, \max_0 = U_{\min} = 0, \max_4 = U_{\max} = 100$ 。

$$A_1(x) = \begin{cases} 1, 0 \leq x < 49.3 \\ \cos[\frac{\pi/3}{28.3}(x - 49.3)], 49.3 < x \leq 77.6 \\ \frac{1}{2} - \sin[\frac{\pi/6}{10}(x - 77.6)], 77.6 < x \leq 87.6 \\ 0, 87.6 < x \end{cases}$$

$$A_{i=2}(x) = \begin{cases} 0, U_{\min} \leq x < \max_{i-2} \\ 1 - \cos[\frac{\pi/3}{49.3}(x - 0)], 0 \leq x < 49.3 \\ \sin(\frac{x - 49.3}{28.3} * \pi/3 + \pi/6), 49.3 \leq x < 77.6 \\ \cos[\frac{\pi/3}{10}(x - 77.6)], 77.6 < x \leq 87.6 \\ \frac{1}{2} - \sin[\frac{\pi/6}{12.4}(x - 87.6)], 87.6 < x \leq 100 \\ 0, \max_{i+2} < x \leq U_{\max} \end{cases}$$

$$A_i(x) = \begin{cases} 0, U_{\min} \leq x < 49.3 \\ 1 - \cos[\frac{\pi/3}{28.3}(x - 49.3)], 49.3 \leq x < 77.6 \\ \sin(\frac{x - 77.6}{10} * \pi/3 + \pi/6), 77.6 \leq x < 87.6 \\ 1, 87.6 < x \leq U_{\max} \end{cases}$$

3.3 在不同的 ε 下计算区间匹配

要查找成绩优秀的学生,取 $\varepsilon = 0.5$,使 $A_3(x) > 0.5$,得 x 的区间为 $[77.6, 100]$,查表将成绩按升序排序可知对应的结果如表 2 所示。

表 2 $\varepsilon = 0.5$ 时所得结果

| 编号 | 成绩 | 编号 | 成绩 |
|----|----|----|-----|
| 23 | 78 | 5 | 89 |
| 25 | 78 | 21 | 89 |
| 6 | 82 | 10 | 90 |
| 24 | 83 | 27 | 92 |
| 19 | 86 | 22 | 94 |
| 2 | 87 | 11 | 98 |
| 9 | 87 | 26 | 98 |
| 15 | 87 | 12 | 100 |
| 1 | 88 | | |

取 $\varepsilon = 0.87$,使 $A_3(x) > 0.87$,得 x 区间为 $[82.6, 100]$,查表将成绩按升序排序可知对应结果如表 3 所示。

表 3 $\varepsilon = 0.87$ 时所得结果

| 编号 | 成绩 | 编号 | 成绩 |
|----|----|----|-----|
| 24 | 83 | 21 | 89 |
| 19 | 86 | 10 | 90 |
| 2 | 87 | 27 | 92 |
| 9 | 87 | 22 | 94 |
| 15 | 87 | 11 | 98 |
| 1 | 88 | 26 | 98 |
| 5 | 89 | 12 | 100 |

由上可知, ε 越大,所得的结果区间越小,准确度越高。

4 结束语

以模糊集理论为基础,通过正态分布函数的密度函数的某些特性设定了一个总的模糊化自动匹配算法,可以利用样本中已有的信息对所有的模糊集都设置一个相对较为准确的隶属函数,可以通过该隶属函数对模糊集进行自动匹配,减少了用户自己设定隶属函数的主观性,大大提高了结果的准确度。该算法很大程度上增加了用户友好性,能够满足用户对模糊查询的需求,但是也存在一定的局限性。比如需要计算出所有数据的平均值和方差,数据量越大时,需要的内存越多,计算的效率会相应降低,这是今后需要改进的地方。但是当数据量越大时,数据的一般性能体现出来,所得结果的准确度也会相应提高。

参考文献:

[1] 庄英. 关系型数据库的模糊查询研究[D]. 南京:南京信息工程大学,2008.

[2] 樊新华. 基于关系数据库的模糊查询技术[J]. 计算机与数字工程,2009,37(10):149-152.

[3] 郑知卉. 关系数据库模糊聚合查询方法研究[D]. 沈阳:东北大学,2012.

[4] 陈逸菲. 基于模糊理论的关系数据库查询技术研究[D]. 南京:南京信息工程大学,2006.

[5] TAMANI N, LIÉTARD L, ROCACHER D. Bipolar SQLf: a flexible querying language for relational databases[C]//Proceedings of the 9th international conference on flexible query answering systems. Ghent, Belgium;Spring,2011:472-484.

[6] HOQUE A H M S, ALI M S, AKTARUZZAMAN M, et al. SK Mondol Performance comparison of fuzzy queries on fuzzy database and classical database[C]//5th international conference on electrical and computer engineering. Dhaka, Bangladesh:IEEE,2008:654-658.

[7] 王慧. 关系型数据库的模糊查询方法研究与实现[D]. 南京:南京信息工程大学,2009.

[8] 王慧,张颖超. 基于模糊逻辑带权重的模糊查询研究[J]. 计算机应用研究,2006,26(1):114-116.

[9] CASTELLTORT A, LAURENT A. Fuzzy queries over NoSQL graph databases: perspectives for extending the cypher language[C]//International conference on information processing and management of uncertainty in knowledge-based systems. [s. l.]:[s. n.],2014:384-395.

[10] BENALI-SOUGUI I, HIDRI M S, GRISSA-TOUZI A. No-FSQL: a graph-based fuzzy NoSQL querying model[J]. International Journal of Fuzzy Systems Applications,2016,5(2):54-63.

[11] CASTELLTORT A, LAURENT A. Fuzzy historical graph pattern matching a NoSQL graph database approach for fraud ring resolution[C]//IFIP international conference on artificial intelligence applications and innovations. [s. l.]:[s. n.],2015:151-167.

[12] KIANMEHR K, KOOCHAKZADEH N, ALHAJJ R, et al. Ask-Fuzzy: attractive visual fuzzy query builder[C]//IEEE international conference on data engineering. Washington, DC, USA:IEEE,2012:1241-1244.

[13] PIVERT O, SLAMA O, SMITS G, et al. SUGAR: a graph database fuzzy querying system[C]//IEEE tenth international conference on research challenges in information science. Grenoble, France:IEEE,2016:1-2.

[14] PIVERT O, SMITS G, THION V. Expression and efficient processing of fuzzy queries in a graph database context[C]//IEEE international conference on fuzzy systems. Istanbul, Turkey:IEEE,2015:1-8.

[15] 孟广武. 模糊数学的基本理论及其应用(Ⅲ)—截集,分解定理及扩展原理[J]. 聊城大学学报:自然科学版,2000,13(2):1-5.

[16] 刘普寅,吴孟达. 模糊理论及其应用[M]. 长沙:国防科技大学出版社,2001.