

笛卡尔积运算对数据库数据质量的传递影响<sup>\*</sup>)

陈卫东 张维明

(国防科技大学信息系统与管理学院 长沙 410073)

**摘要** 本文在属性粒度给出了数据质量评价模型,定义了正确性评价指标。在分析属性量化前后错误率不同和假设错误出现随机性的基础上,给出了两种错误率之间的定量关系。还研究了笛卡尔积运算的质量传播关系,并分别用属性量化前后错误率进行了量化表示。

**关键词** 数据库,数据质量,质量传递,笛卡尔积,正确性

## Data Quality Propagation of Cartesian Product for Relational Database

CHEN Wei-dong ZHANG Wei-ming

(Information System and Management Dept. of NUDT, Changsha 410073, China)

**Abstract** The paper presents a data quality model at attribute level, and defines the accuracy metric. After analyzing the difference of error rate before and after quantization, and with the assumption of random errors occurrence, the quantities relationship between them is presented. The paper also discusses the quality propagation of Cartesian Product operation, and represents the propagation formula with error rate before and after quantization separately.

**Keywords** Database, Data quality, Quality propagation, Cartesian product, Accuracy

## 1 引言

在现代社会里,数据是企业信息化的基础和重要资源。然而,随着应用系统数据量的急剧增加、新应用的不断出现以及应用之间的相互整合,数据质量问题变得日益突出。质量低劣的数据已成为影响企业进行正确决策的一个重要因素。在与数据质量有关的众多理论、技术和方法中,本文探讨数据库数据质量在关系代数运算下的质量传播问题。

数据库数据质量的研究可以在不同粒度下进行。元组粒度的数据质量研究,将元组作为一个整体看待,以元组是否完整和正确来反映关系的质量状况。其中 Kon<sup>[1]</sup>定义了元组不正确(Inaccuracy)、误属(Mismembership)和不完整(Incompleteness)指标,概要描述了查询操作对元组质量的影响。Reddy<sup>[2,3]</sup>将元组分为完全正确、部分正确和误属三种类型,针对正确性评价指标,研究了选择、投影、笛卡尔积运算质量传递问题。Parssian<sup>[4,5]</sup>在 Kon 和 Reddy 研究的基础上提出了数据质量评价模型,定量研究了选择、投影、笛卡尔积等运算对正确性(Accuracy)和不完整性(Incompleteness)指标的质量传递问题。在属性粒度的研究中,Metro<sup>[6,7]</sup>用同构划分的方法将关系分解为若干质量同构的块,SQL 查询得到新关系的质量以同构块的质量来测度,该方法是针对静态数据的评价,同构划分的有效性难以确定。Scannapieco<sup>[8]</sup>在元组和属性粒度定义了完整性指标,并对并、交等运算的传播问题进行了概要研究。

本文通过扩展 Parssian 评价模型,在属性粒度方面研究正确性指标对笛卡尔积运算下质量传递的影响。

## 2 数据质量模型

## 2.1 模型

Parssian 研究数据质量在元组粒度的传播问题<sup>[5]</sup>,以 Kon<sup>[1]</sup>提出的基本模型为基础,数据质量模型如图 1 所示<sup>[5]</sup>。其中概念关系代表实体的真实情况,是实体抽象完美的反映。实例关系是数据在数据库中的实际存储,存在错误和不完整。由于获取时的偏差,使得实例关系中的元组与概念关系中对应的元组出现差异,即表现为质量问题。对于不同类型的元组,模型将它们分为三类:正确实例( $S_A$ )是概念关系元组的准确反映;错误实例( $S_I$ )是错误地获取了原本正确的元组;无对应的实例( $S_M$ )则是实例关系中包含了本不属于概念关系中的元组;实例关系中缺失了本应在概念关系中存在但未获取的元组为  $S_C$ 。

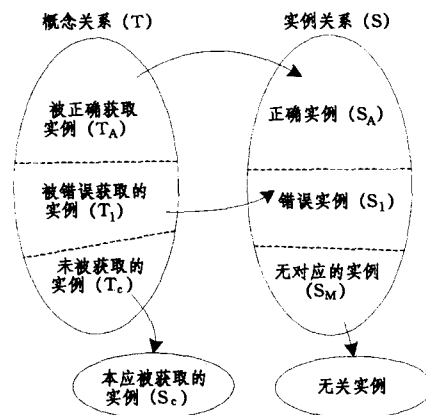


图 1 Parssian 模型

在图 1 的模型中,正确性定义为正确实例元组( $S_A$  部分)占总元组的比率。这意味着任何属性出现错误的元组即认为是不正确的,属于错误实例。显然,它是严格条件下的定义。

<sup>\*</sup> 国家自然科学基金项目(编号:60504036)资助。陈卫东 博士研究生,研究领域为信息系统与智能决策;张维明 教授,博士生导师,研究领域为信息系统与智能决策。

然而,在很多情况下,元组中一个数据项出错与多数数据项出现错误,其质量对使用者而言是不同的,因此需要在属性粒度评价数据质量。

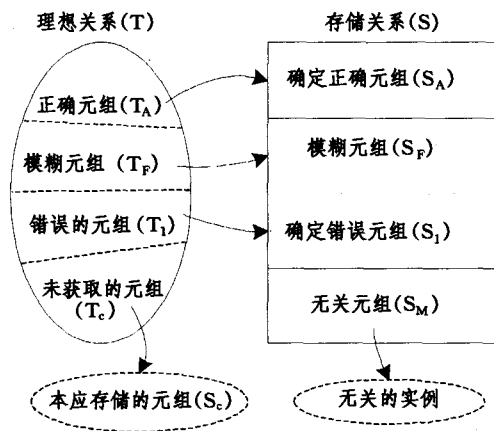


图2是笔者定义的在属性粒度的数据质量模型。其中理想关系是实体完整无误的反映,存储关系是理想关系的实际映射,存在错误和缺失。正确性指标定义在属性粒度,即数据项是否正确。这样,元组正确与否已不再是非此即彼的二值问题,经量化后处于0~1区间。于是可以将元组分为四种类型:①确定正确部分( $S_A$ )由没有任何错误的元组组成;②部分属性正确的元组介于正确和不正确之间形成模糊元组( $S_F$ )部分;③确定错误元组( $S_I$ )由完全错误的元组组成;④无关元组( $S_M$ )和本应存储元组( $S_C$ )。其中确定正确部分、无关元组部分和本应存储元组部分与 Parssian 模型相同。

记正确性指标为  $\alpha$ ,在属性粒度定义关系  $S$  的正确性测度如下。

**定义1(数据项正确性)** 如果  $i$  行  $j$  列数据项正确则为1,否则为0。

**定义2(关系正确性,  $\alpha_S$ )** 关系  $S$  中正确数据项个数与总数据项之比。

**定义3(误属性,  $\gamma_S$ )** 误属元组占总元组的比率,即  $\gamma_S = |S_M|/|S|$ 。

关系可以分为键属性和非键属性两个部分,考虑到键属性是元组的唯一标识,分别定义键属性和非键属性正确性。

表1 数据库存储模式数据(S)

| #  | 姓名* | 籍贯 | 地址                 | 工作时间    | 职务 | 军衔 |
|----|-----|----|--------------------|---------|----|----|
| 1  | 舒望清 | 江苏 | 扬州市岳口陈场林场          | 1969.08 | 副师 | 大校 |
| 2  | 王运生 | 上海 | 宝山区辛庄村             | 1973.10 | 正团 | 中校 |
| 3  | 张德运 | 北京 | 怀柔县车埠镇76号          | 1975.07 | 副团 | 中校 |
| 4  | 李可  | 浙江 | 番禺市成全镇农村村          | 1980.09 | 正营 | 少校 |
| 5  | 王平  | 四川 | 隆昌市隆化镇南卫村          | 1994.07 | 正营 | 少校 |
| 6  | 周鲁  | 湖南 | 益阳市太平街3号           | 1992.10 | 副营 | 上尉 |
| 7  | 赵兴东 | 江西 | 九江市邓苍村十一组          | 2002.07 | 正连 | 上尉 |
| 8  | 张剑  | 山西 | 太原市中洲农场<br>红旗大道25号 | 2003.07 | 副连 | 中尉 |
| 9  | 王伟亮 | 湖北 | 信阳市花桥镇             | 2005.09 | 副连 | 中尉 |
| 10 | 张萍  | 湖北 | 麻城市                | 2003.07 | 副连 | 少尉 |

**定义4 键属性正确性( $\alpha_{K_S}$ )** 为键属性所有正确数据项与总数据项的数量比。非键属性正确性为非键属性所有正确项与总数据项的数量比,记为  $\alpha_{Q_S}$ 。

显然  $\alpha_S = \alpha_{K_S} + \alpha_{Q_S}$ ,可以证明  $\alpha_{K_S} = k(1-\gamma_S)/m$ ,其中  $k$  为键属性个数,  $m$  为属性总数。如表1数据(姓名是键,斜体下划线为错误数据项),1#,2#,3#为完全正确元组( $S_A$ ),4#,5#,6#,7#为部分正确的元组( $S_F$ ),8#是确定错误元组( $S_I$ )(键属性是正确的,否则是误属的),9#,10#元组的键属性错误,不应属于关系  $S$ ,因而是误属元组( $S_M$ )。

根据定义对表1进行量化,得到表2。

表2 示例表量化结果表

| #  | 姓名* | 籍贯 | 地址 | 工作时间 | 职务 | 军衔 |
|----|-----|----|----|------|----|----|
| 1  | 1   | 1  | 1  | 1    | 1  | 1  |
| 2  | 1   | 1  | 1  | 1    | 1  | 1  |
| 3  | 1   | 1  | 1  | 1    | 1  | 1  |
| 4  | 1   | 0  | 1  | 1    | 0  | 1  |
| 5  | 1   | 1  | 1  | 0    | 1  | 1  |
| 6  | 1   | 1  | 0  | 1    | 1  | 0  |
| 7  | 1   | 1  | 1  | 0    | 1  | 1  |
| 8  | 1   | 0  | 0  | 0    | 0  | 0  |
| 9  | 0   | 0  | 0  | 0    | 0  | 0  |
| 10 | 0   | 0  | 0  | 0    | 0  | 0  |

表2中9#、10#是误属元组,即便有部分属性是正确的,但它们对关系  $S$  而言仍然是错误的,因此元组量化全为0。

### 2.2 模型比较

Parssian 模型是元组粒度的质量评价模型,非键属性数据项中只要错误即认为是错误,于是可以将模型简化为一个列向量形式的量化表示(图3)。根据其定义可知,正确性是量化向量中为1的个数与总数之比,也可以理解为度量  $S_A$  元组键属性  $K_S$  在关系代数运算质量上的变化情况(图中以方格显示部分)。显然,模型忽略了  $S_I'$  部分中原本正确的数据项。

根据本文的模型,正确性定义为关系所有为1的个数与总个数之比。量化分布如图3所示,从图中可以看出本文模型定义的正确性指标覆盖了  $S_A, S_F$  和  $S_I$  三部分。因此,较 Parssian 模型细致,结果更准确。

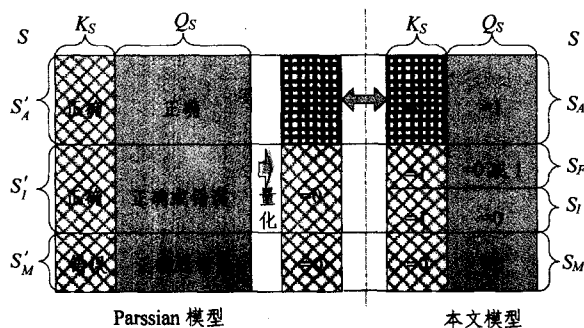


图3 模型量化分布比较

### 3 属性错误率

对比表1和表2可以看出,由于量化过程受到  $S_M$  部分的影响,使得  $S_M$  中原本正确的数据项量化为0,导致属性错误率在量化前后存在差异。

**定义 5** 属性量化前错误率  $E_j$  为属性  $A_j$  在量化之前错误数据项个数与总个数之比。表 1 量化前属性错误率分别为: 0.2, 0.3, 0.2, 0.4, 0.2, 0.2。

**定义 6** 属性量化后错误率  $\hat{E}_j$  为属性  $A_j$  在量化之后错误数据项个数与总个数之比。表 2 量化后错误率为: 0.2, 0.4, 0.4, 0.5, 0.4, 0.4。

量化前属性错误率反映  $A_j$  固有的质量分布, 量化后错误率反映质量的评价。

假设: 关系  $S$  中属性错误项的出现是随机的。

该假设意味着量化前错误率在  $S_M$  中也为  $E_j$ 。可以证明关系  $S$  的非键属性  $A_j$  的量化前后错误率有  $\hat{E}_j \geq E_j$ , 且有

$$\hat{E}_j = \gamma_s + (1 - \gamma_s) E_j \quad (1)$$

由式(1)计算得到量化后错误率分别为: 0.44, 0.36, 0.52, 0.36, 0.36。式(1)表明误属元组数量的增加, 使得关系的正确性进一步下降, 质量更糟。

### 4 笛卡尔积运算质量传递

设关系  $S_1$  有  $m_1$  个属性,  $k_1$  个键属性。  $S_2$  有  $m_2$  个属性,  $k_2$  个键属性。  $S_1$  和  $S_2$  的笛卡尔积是一个有  $m_1 + m_2$  个属性的元组集合, 每个元组  $m_1$  个分量来自  $S_1$  的一个元组, 有  $m_2$  个分量来自  $S_2$  的一个元组, 记为  $S_1 \times S_2$ 。乘积结果集合  $R$  为  $S_1$  和  $S_2$  元组的任意组合, 共有  $|S_1| \times |S_2|$  个元组,  $k_1 + k_2$  个键属性。

#### 4.1 笛卡尔积运算分析

关系  $S_1$  和  $S_2$  的笛卡尔积结果  $R$  中包含了  $S_1, S_2$  中所有元组的任意组合, 这样  $S_1, S_2$  不同归属类型的元组互相交叉合并形成新的关系  $R$  后, 将使  $R$  元组的正确性产生改变。

记  $S_1$  的四个组成部分分别为  $S_{1A}, S_{1F}, S_{1I}, S_{1M}$ , 键属性为  $S_{1K}$ ;  $S_2$  分别为  $S_{2A}, S_{2F}, S_{2I}, S_{2M}$ , 键属性为  $S_{2K}$ 。  $S_1 \times S_2$  乘积前形态如图 4 所示。

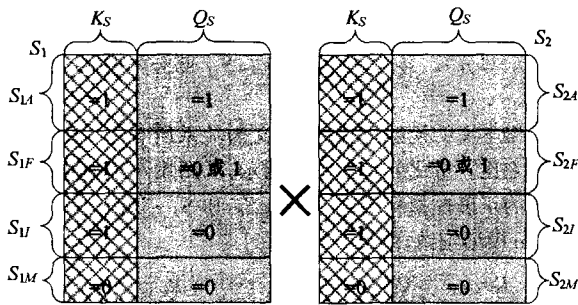


图 4 笛卡尔积运算前形态

$S_1 \times S_2$  的乘积可以分解为

$$\begin{aligned} S_1 \times S_2 = & [(S_{1A} \times S_{2A}) \cup (S_{1F} \times S_{2A}) \cup (S_{1I} \times S_{2A}) \cup (S_{1M} \\ & \times S_{2A})] \\ & \cup [(S_{1A} \times S_{2F}) \cup (S_{1F} \times S_{2F}) \cup (S_{1I} \times S_{2F}) \cup \\ & (S_{1M} \times S_{2F})] \\ & \cup [(S_{1A} \times S_{2I}) \cup (S_{1F} \times S_{2I}) \cup (S_{1I} \times S_{2I}) \cup (S_{1M} \\ & \times S_{2I})] \\ & \cup [(S_{1A} \times S_{2M}) \cup (S_{1F} \times S_{2M}) \cup (S_{1I} \times S_{2M}) \cup \\ & (S_{1M} \times S_{2M})] \end{aligned}$$

注意到  $S_{1M} \times S_{2A}, S_{1M} \times S_{2F}, S_{1M} \times S_{2I}, S_{1A} \times S_{2M}, S_{1F} \times S_{2M}, S_{1I} \times S_{2M}$  项乘积结果元组键属性中包含了  $S_{1K}$  或  $S_{2K}$ , 故元组必定属于  $R_M$ , 量化值全部为 0, 使得  $R$  的属性  $A_j$  量化后错误率  $\hat{E}_j^R$  发生变化。二者有以下数量关系。

**定理 1** 关系  $R = S_1 \times S_2$  的非键属性  $A_j^{S_1}$  (或  $A_j^{S_2}$ ), 在  $S_1$  (或  $S_2$ ) 中的量化后错误率记为  $\hat{E}_j^{S_1}$  (或  $\hat{E}_j^{S_2}$ ), 在  $R$  中的错误率记为  $\hat{E}_j^{R_{S_1}}$  (或  $\hat{E}_j^{R_{S_2}}$ ), 则

(1)  $\hat{E}_j^{S_1}$  和  $\hat{E}_j^{R_{S_1}}$  的关系为

$$1 - \hat{E}_j^{R_{S_1}} = (1 - \hat{E}_j^{S_1})(1 - \gamma_{S_2}) \quad (2)$$

(2)  $\hat{E}_j^{S_2}$  和  $\hat{E}_j^{R_{S_2}}$  的关系为

$$1 - \hat{E}_j^{R_{S_2}} = (1 - \hat{E}_j^{S_2})(1 - \gamma_{S_1}) \quad (3)$$

#### 4.2 质量传递

笛卡尔积的质量传递计算, 可以根据定理 1 得到的结果, 并根据定义分别计算键属性和非键属性正确性, 然后求和得到。质量传递关系如下。

**定理 2** 关系  $S_1$  和  $S_2$  笛卡尔积运算结果关系  $R$  正确性为:

(1) 以关系  $S$  的键属性 ( $\alpha_{K_S}$ ) 和非键属性 ( $\alpha_{Q_S}$ ) 正确性表示为

$$\alpha_R = \frac{m_1 m_2 (k_1 + k_2)}{k_1 k_2 (m_1 + m_2)} \alpha_{K_{S_1}} \alpha_{K_{S_2}} + \frac{m_2}{k_2} \alpha_{K_{S_2}} \alpha_{Q_{S_1}} + \frac{m_1}{k_1} \alpha_{K_{S_1}} \alpha_{Q_{S_2}} \quad (4)$$

(2) 以量化后错误率  $E$  表示为

$$\alpha_R = \frac{(1 - \gamma_{S_2})(m_1 - k_1 \gamma_{S_1}) - \sum_{j=k_1+1}^{m_1} \hat{E}_j^{S_1}}{m_1 + m_2} + \frac{(1 - \gamma_{S_1})(m_2 - k_2 \gamma_{S_2} - \sum_{j=k_2+1}^{m_2} \hat{E}_j^{S_2})}{m_1 + m_2} \quad (5)$$

(3) 以量化前错误率表示为

$$\alpha_R = (1 - \gamma_{S_2})(1 - \gamma_{S_1}) \left( 1 - \frac{\sum_{j=k_1+1}^{m_1} E_j^{S_1} + \sum_{j=k_2+1}^{m_2} E_j^{S_2}}{m_1 + m_2} \right) \quad (6)$$

表 3 为示例关系

表 3 示例数据表

| 姓名 * | 调职年月    | 下次调职年份 |
|------|---------|--------|
| 王运生  | 2005.10 | 2008   |
| 张德运  | 2006.08 | 2009   |
| 李可   | 2000.01 | 2003   |
| 李伟   | 2003.07 | 2006   |

表 2( $S_1$ ) 与表 3( $S_2$ ) 做笛卡尔积运算。  $m_1 = 6, m_2 = 3, k_1 = k_2 = 1, \gamma_{S_1} = 0.2, \gamma_{S_2} = 0.25$ 。表 3 量化前错误率和量化后错误率如表 4 所示。

表 4 示例数据错误率

| 项              | 姓名  | 调职年月 | 下次调职年份 |
|----------------|-----|------|--------|
| 观察得到表 3 量化前错误率 | 0.1 | 0.1  | 0.1    |
| 观察得到表 3 量化后错误率 | 0.1 | 0.2  | 0.2    |
| 式计算量化后错误率      | -   | 0.32 | 0.32   |

按照定义直接计算  $S_1 \times S_2$  的正确性  $\alpha = 0.46$ , 以量化前错误率计算代入(5)式得到  $\alpha = 0.50$ , 以观察得到量化后错误率计算代入(5)式得到  $\alpha = 0.51$ , 以计算得到的量化后错误率计算代入(6)式得到  $\alpha = 0.50$ 。

**结束语** 本文结论与 Parssian 的结论  $\alpha'_R = \alpha'_{S_1} \alpha'_{S_2}$  不同。如果关系  $S_1$  和  $S_2$  所有属性均为关键字, 则本文传递公式简化为  $\alpha_R = \alpha_{S_1} \alpha_{S_2}$ , 即 Parssian 结论是本文结论的特例。

(下转第 216 页)

表 5 ID3 算法生成的规则

| tributes | Decision |
|----------|----------|
| a b c d  | e        |
| 2 2 0 0  | 0        |
| 2 2 0 1  | 0        |
| 0 2 1 0  | 1        |
| 0 2 0 0  | 1        |
| 1 1 1 2  | 1        |
| 0 1 1 1  | 1        |
| 0 1 0 0  | 2        |
| 0 0 1 2  | 2        |
| 1 0 1 2  | 2        |
| 0 0 0 1  | 2        |

针对决策系统  $\eta$ , 基于粗糙集的分类算法得到的规则可理解为, 当交易具有以下特征: (1) 有较大的进料加工进口付汇, 但没有或只有较少的进料加工出口及收汇; (2) 有较大的来料加工出口, 但没有或只有较少的来料加工国外直接进口 (来料大部分由国内通过深加工转厂转入), 则该交易为可疑交易。

表 5 给出了 ID3 算法在决策系统  $\eta$  上生成的规则。对比这两个算法得出的规则, 基于粗糙集的分类算法得到规则的前件基本为 ID3 算法得到规则前件的子集, 且决策结论也基本一致。因此, 从实验结果看, 该算法的属性约简功能具有较好的效果, 与粗糙集理论描述一致。

表 3 分辨矩阵

| $\emptyset$ | {b,c,d}     | {b,d}                | $\emptyset$ | $\emptyset$ | {a,b,c}     | {a,c}       | {a,b,c,d}   |
|-------------|-------------|----------------------|-------------|-------------|-------------|-------------|-------------|
| {b,c,d}     | $\emptyset$ | $\emptyset\{a,b,c\}$ | {a,d}       | {a,b,d}     | {a,b,d}     | {a,c,d}     |             |
| {b,d}       | $\emptyset$ | $\emptyset$          | {a,b}       | {a,b,c,d}   | {a,c,d}     | {a,b,c,d}   | {a,b,c,d}   |
| $\emptyset$ | {a,b,c}     | {a,b}                | $\emptyset$ | $\emptyset$ | {b,c,d}     | {a,c,d}     | {a,b,c,d}   |
| $\emptyset$ | {a,d}       | {a,b,c,d}            | $\emptyset$ | $\emptyset$ | {b}         | {a,b}       | {a,c,d}     |
| {a,b,c}     | {a,b,d}     | {a,b,d}              | {b,c,d}     | {b}         | $\emptyset$ | $\emptyset$ | $\emptyset$ |
| {a,c}       | {a,b,d}     | {a,b,c,d}            | {a,c,d}     | {a,b}       | $\emptyset$ | $\emptyset$ | $\emptyset$ |
| {a,b,c,d}   | {a,c,d}     | {a,b,c,d}            | {a,b,c,d}   | {a,c,d}     | $\emptyset$ | $\emptyset$ | $\emptyset$ |

**结束语** 本文提出一个基于粗糙集的挖掘算法, 该算法用于生成决策模型, 该决策模型由一组规则组成, 首先实现属性约简, 形成分辨矩阵, 然后从中发现规则。这一算法属性约简具有较好有效性, 可减少未知样本参与分类的决策属性, 适合模型相对稳定、更新不频繁且建模过程可以在后台进行的应用。

参 考 文 献

[1] Pawlak Z. Rough Sets; Theoretical Aspects of Reasoning about Data. Dordrech: Kluwer Academic Publishers, 1991, 10-34  
 [2] Han B, Wu T J. Data Mining in Multisensor System Based on Rough Set Theory//Proceedings of the American Control Conference. Arlington, 2001; 4427-4431  
 [3] Rough 集理论与知识获取. 西安: 西安交通大学出版社, 2001;

156-176  
 [4] 张文修, 吴志伟. 粗糙集理论和研究综述. 模糊系统与数学, 2000, 14(4): 1-12  
 [5] 杨涛, 李龙澎. 一种基于粗糙集聚类的数据约简算法. 系统仿真学报, 2004, 16(10): 2195-2200  
 [6] 徐泉清, 朱玉文, 李亮. 一种结合粗糙集和 COBWEB 的聚类器. 计算机应用, 2005, 25(6): 1350-1352  
 [7] 干庆东, 马听, 戴华平. 基于粗糙集属性量度的数据库分解方法. 浙江大学学报(工学版), 2004, 38(9): 1196-1199  
 [8] 刘业政, 杨善林. 基于粗集理论的 Null 值估算方法研究. 计算机工程, 2001, 27(10): 41-42  
 [9] 苗谦谦, 王压. 粗糙集理论中概念与运算的信息表示. 软件学报, 1999, 10(2): 113-116  
 [10] 苗谦谦, 胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展, 1999, 36(6): 681-684

(上接第 212 页)

导致结论不同有两方面原因: 首先, 衡量的粒度不同, 定义不同, 使得结论不同; 另一方面, Parssian 乘积过程可以简化为量化向量的乘积, 而本模型笛卡尔乘积可以看成是键属性、非键属性分别乘积然后归并合成的结果。

参 考 文 献

[1] Kon H B, Madnick E, Siegel M D. Good Answers From Bad Data; A Data Management Strategy[C]. Massachusetts Institute of Technology (MIT), Sloan School of Management, 1995: 1-16  
 [2] Reddy M P, Wang R Y. A Data Quality Algebra for Estimating Query Result Quality[C]// CISMODO Conference. Bombay, 1996  
 [3] Yang W, Wang L, Richard Y, et al. Data Quality[M]. Kluwer,

2001  
 [4] Parssian A, Sumit S, Varghese J S. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product[J]. Management Science, 2004, 50(7): 967-982  
 [5] Parssian A, Sumit S, Varghese J S. Assessing Information Quality for the Composite Relational Operation Join[C]. 2002: 225-237  
 [6] Motro A, Igor R. Not All Answers Are Equally Good; Estimating the Quality of Database Answers[M]. Kluwer Academic Publishers, 1997: 1-21  
 [7] Motro A, Igor R. Estimating The Quality of Databases[C]// The 3rd International Conference on Flexible Query Answering Systems (FQAS). Cambridge, MA, 1998: 298-307  
 [8] Scannapieco M, Carlo B. Completeness in the Relational Model; a Comprehensive Framework[C]. 2004: 333-345