控制与决策
Control and Decision

Tarski代数视角下函数依赖与属性蕴含的关系

安秋生， 孔祥玉， 曹泽豪

引用本文:
安秋生, 孔祥玉, 曹泽豪. Tarski代数视角下函数依赖与属性蕴含的关系[J]. 控制与决策, 2020, 35(7): 1659–1666.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1183

您可能感兴趣的其他文章

Articles you may be interested in

强化属性依赖关系的K阶贝叶斯分类模型

K–dependence Bayesian classifiers for strengthening attribute dependencies

控制与决策. 2019, 34(6): 1234–1240 https://doi.org/10.13195/j.kzyjc.2017.1595

基于非线性复杂系统观的直觉模糊变权多属性决策方法

Intuitionistic fuzzy multi–attribute decision making method with variable weights based on the perspective of nonlinear complex systems

控制与决策. 2017, 32(11): 2013–2020 https://doi.org/10.13195/j.kzyjc.2016.1301

基于改进的TODIM方法的区间灰数多属性决策模型

Multiple attribute decision–making model with interval grey number based on improved TODIM method

控制与决策. 2016(2): 261–266 https://doi.org/10.13195/j.kzyjc.2014.1419

基于证据推理和第3代前景理论的不确定性多属性决策方法

Method for uncertain multi–attribute decision making based on evidential reasoning and third–generation prospect theory

控制与决策. 2016, 31(1): 105–113 https://doi.org/10.13195/j.kzyjc.2014.1472

基于高斯核函数的朴素贝叶斯分类器依赖扩展

Dependency extension of naive Bayesian classifiers based on Gaussian kernel function

控制与决策. 2015, 30(12): 2280–2284 https://doi.org/10.13195/j.kzyjc.2014.1650

文章编号: 1001-0920(2020)07-1659-08 DOI: 10.13195/j.kzyjc.2018.1183

Tarski代数视角下函数依赖与属性蕴含的关系

安秋生[1†], 孔祥玉[2], 曹泽豪[2]

(1. 山西师范大学 数计学院，山西 临汾 041000；2. 火箭军工程大学 导弹工程学院，西安 710025)

摘 要: 研究经典函数依赖与属性蕴含之间的关系. 首先介绍Tarski量词消除理论, 分别研究Tarski量词消除理论下的函数依赖表示方式和属性蕴含表示方式, 得出无量词Tarski代数下函数依赖与属性蕴含表示的统一数学模型; 然后, 进一步从形式概念分析的角度进行深入研究, 得出不同转换背景下函数依赖与属性蕴含两者成立的关系, 并从函数的观点分析两者的根本语义差别; 最后, 基于Armstrong公理的满足性讨论函数依赖与属性蕴含所满足的Armstrong公理, 基于Duquenne-Guigues基的满足性得出最小依赖集与Duquenne-Guigues基的关系, 并对函数依赖与属性蕴含之间的关系进行全面总结.

关键词: Tarski代数；函数依赖；属性蕴含；转换背景；Armstrong公理；Duquenne-Guigues基

中图分类号: TP301 **文献标志码:** A

The study of relationship between functional dependency and attribute implication based on Tarski algebra view

AN Qiu-sheng[1†], *KONG Xiang-yu*[2], *CAO Ze-hao*[2]

(1. School of Mathematics and Computer Science，Shanxi Normal University，Linfen 041000，China；2. College of Missile Engineering，Rocket Force University of Engineering，Xi'an 710025, China)

Abstract: The relationship between classical functional dependency and attribute implication is investigated emphatically. Firstly, the quantifier elimination theory of Alfred Tarski is introduced. Then, the representation forms of functional dependency and the attribute implication under the quantifier elimination theory of Alfred Tarski are researched respectively, and the unified mathematics model of functional dependency and attribute implication under the Alfred Tarski algebra without quantifier is established. Then, the relationship between existences of functional dependency and attribute implication is discussed under different conversion contexts from the view of formal concept analysis, and the fundamental semantics difference between them is studied from the view of function. Finally, the problem on whether functional dependency and attribute implication can satisfy the Armsrong's axiom are researched from their satisfaction degrees according to Armsrong's axiom, the relationship between minimum dependency set and Duquenne-Guigues base is given from their satisfaction degrees according to Duquenne-Guigues base, and the relationship between classical functional dependency and attribute implication is summarized in all directions.

Keywords: Tarski algebra；functional dependency；attribute implication；convert context；Armstrong axiom；Duquenne-Guigues base

0 引 言

函数依赖是关系数据库理论的重要概念, 属性蕴含是形式概念分析(FCA)中的概念, 它们分属不同的学科却又密切相关. 对于它们之间的关系, 近年来已有多位国内外学者进行了研究, 文献[1]认为函数依赖与属性蕴含具有等价性, 文献[2]认为存在于概念格中的属性蕴含集与存在于对应数据库表中的函数依赖集语义上相同, 文献[3]对于函数依赖与属性蕴含关系的研究结论与文献[1]相同, 但是文献[4]却持不同观点, 认为它们并不等价, 那么两者之间的关系究竟如何? 本文将对此进行更深入的研究[5].

经典关系数据库以关系代数为基础, 而关系代数的基础则是传统的集合论和量化逻辑, 量化逻辑以量词与变量为基础, 但是增加了推理的复杂性. 文献[6]指出: "涉及到关系中的依赖性、关系与操作的泛化理论仍然缺乏", 著名学者Tarski[7]提出的无变量集合理论的形式化方法(或称Tarski代数)则为文献[6]所提出的问题提供了可能的解决方案. 本文尝试以

收稿日期: 2018-09-03；**修回日期:** 2019-01-28.

基金项目: 国家自然科学基金面上项目(61074072, 61374120).

责任编委: 阳春华.

†通讯作者. E-mail: aaqqss@sina.com.

Tarski代数方法为基础,研究此视角下的函数依赖与属性蕴含,并结合其他方法解决他们的关系问题.

1 Tarski量词消除理论

Tarski[7]于1948年提出了一种关于实数封闭域(real closed fields)基础理论的量词消除方法,该方法同时也可以作为一种决策方法.

一般的自由量词公式是由多项式方程$(f(x)=0)$与不等式$(f(x)\leqslant 0)$使用布尔操作算子$\bigwedge$、$\bigvee$、$\Rightarrow$、$(\,)$等构成的表达式,也允许严格的不等式$f(x)>0$及$f(x)\neq 0$. 具有变量$x=(x_1,x_2,\ldots,x_n)$的公式(前束式)是下列类型的表达式:

$$(Q_1x_1)\ldots(Q_sx_s)F(f_1(x),\ldots,f_r(x)).$$

其中:Q_i为量词$\forall$或$\exists$,如果有量词对应于变量x_i,则称x_i被量化,否则称x_i是自由的,$F(f_1(x),\ldots,f_n(x))$为自由量词公式.

Tarski已经证明,每一个包含量词的公式均等价于一个无量词的公式.

引理1[8] 实数域上的每个一序公式存在一个等价的量词自由公式,并且有明确的算法计算该量词自由公式.

Tarski终生努力的目标之一就是"开发从逻辑表示中去除'量词'的方法",这种努力从根本上最终形成了他的"没有变量的集合理论的形式化方法"——Tarski量词消除理论.

2 Tarski量词消除理论视角下的函数依赖与属性蕴含

2.1 Tarski量词消除理论下的函数依赖

在经典关系数据库中,设R为属性的有限集,对于任意$A\in R$,其所有可能值的集合称为值域,记为$\mathrm{Dom}(A)$,$\forall A\in R,t(A)\in\mathrm{Dom}(A)$,$R$为关系模式,$r$为$R$下的关系,关系数据库是多个$r$的集合.

定义1[6] 给定任意n元关系T的关系模式S的子集x、$y\subseteq S$,该关系满足函数依赖$x\to y$当且仅当对于所有元组对$t,t'\in T$,当其属性值在x上相等时,它们在y上也相等,即

$$\forall t,t'\in T:(t[x]=t'[x]\Rightarrow t[y]=t'[y]).\qquad(1)$$

在式(1)中具有量词$\forall$及变量t、t',因此式(1)是被量化的,下面将利用Tarski量词消除方法消去其中的量词及变量(该推导过程部分参考文献[6],但结果不同).

首先,考虑到在关系数据库中属性值$t[x]$是由映射得到的,可以将$t[x]$变换为$x[t]$,即与属性集x关联的投影函数,则式(1)变换为

$$\forall t,t'\in T:(x(t)=x'(t)\Rightarrow y(t)=y(t')).\qquad(2)$$

为了消去其中的量词,给定函数$f:A\to B$,设二元关系$R\subseteq A\times A$,目的是为了检查集合A的两个值在f下是否有相同的映像,即$a'Ra\equiv f(a')=f(a)$是否成立. 设f^0表示f逆,即$a(f^0)b$当且仅当$b=f(a)$. 函数f在a取值为b,且$(\cdot)$表示二元关系函数复合的扩展. 一般地,有

$$b(R\cdot S)c\equiv\exists a:bRa\bigwedge aSc.\qquad(3)$$

那么有

$$\begin{aligned}&a'Ra\equiv f(a')=f(a)\equiv\\&\exists b:b=f(a')=f(a)\equiv\\&\exists b:b=f(a')\bigwedge b=f(a)\equiv\\&\exists b:a'(f^0)b\bigwedge a(f^0)b\equiv\\&\exists b:a'(f^0)b\bigwedge b(f)a\equiv\\&a'(f^0\cdot f)a.\end{aligned}$$

即$a'Ra\equiv f(a')=f(a)$可以用$a'(f^0\cdot f)a$表示. 使用逆运算与复合运算,式(2)的右边可以改写为

$$\begin{aligned}&(x(t)=x'(t)\Rightarrow y(t)=y(t'))\equiv\\&t(x^0\cdot x)t'\Rightarrow t(y^0\cdot y)t'.\end{aligned}$$

蕴含可以用关系包含表示为

$$R\subseteq S\equiv\forall b,a:bRa\Rightarrow bSa.$$

直接表示成包含$x^0\cdot x\subseteq y^0\cdot y$将导致明显的错误,因为量词的双重辖域$(t\in T\bigwedge t'\in T)$对此并不适用,为此将式(2)变换为

$$\forall t,t':(t\in T\bigwedge t'\in T\bigwedge t(x^0\cdot x)t')\Rightarrow t(y^0\cdot y)t'.\qquad(4)$$

对于元组集合T,定义二元关系$[T]$,有$b[T]a\equiv b=a\bigwedge a\in T$(实际上$b[T]a\equiv a[T]b$). 那么$t\in T$可以表示为$\exists u:u[T]t\equiv\exists u:t[T]u,t'\in T$情况相同,因此有

$$\begin{aligned}&\forall t,t':(t\in T\bigwedge t'\in T\bigwedge t(x^0\cdot x)t')\equiv\\&\exists u,u':t[T]u\bigwedge t'[T]u'\bigwedge t(x^0\cdot x)t'\equiv\\&\exists u,u':t[T]u\bigwedge u(x^0\cdot x)u'\bigwedge u'[T]^0t'.\end{aligned}$$

由于$\bigwedge$是可交换的,$u=t,u'=t',[T]^0=[T]$,由公式$b(R\cdot S)c\equiv\exists a:bRa\bigwedge aSc$从右向左(将$\exists a:bRa\bigwedge aSc$变为$b(R\cdot S)c$)组合两次,有

$$\begin{aligned}&(t[T]u\bigwedge u(x^0\cdot x)u')\bigwedge u'[T]^0t'\equiv\\&t([T]\cdot x^0\cdot x)u'\bigwedge u'[T]^0t'\equiv\\&t([T]\cdot x^0\cdot x\cdot[T]^0)t'.\end{aligned}$$

式(4)可以变为

$$\forall t,t':(t\in T\bigwedge t'\in T\bigwedge t(x^0\cdot x)t'\Rightarrow$$
$$t(y^0\cdot y)t'\equiv t([T]\cdot x^0\cdot x\cdot[T]^0)t'\Rightarrow$$
$$t(y^0\cdot y)t'\equiv[T]\cdot x^0\cdot x\cdot[T]^0\subseteq y^0\cdot y. \tag{5}$$

注意到,在式(1)中出现的量词在式(5)中已经消除,进一步,可将式(5)变换为

$$R\cdot f^0\cdot f\cdot R^0\subseteq g^0\cdot g. \tag{6}$$

式(6)中R可以看作是一种数据集,同时可以是不确定计算的“规格”,或者是有限状态自动机的变迁关系,f(相对于g)可以看作是R的输出产生的一个函数,$\forall b',b:b'Ra'$意味着$b'=a'$且$a'\in R$,bRa意味着$b=a$且$a\in R$,而$\forall a',a:f(a')=f(a)$实际上是$a'(f^0\cdot f)a$. 考虑到在偏序恒等(partial identities)关系下$R=R^0$,从量化逻辑的观点将式(6)与$R\subseteq S\equiv\forall b,a:bRa\Rightarrow bSa$相结合,则式(6)可以扩展如下:

$$(\forall a',a,b',b,bRa:f(a')=f(a)\bigwedge b'Ra')\Rightarrow$$
$$(g(b')=g(b)). \tag{7}$$

注意到,本文得到式(7)与文献[6]结果$\forall a',a:f(a')=f(a)\Rightarrow(\forall b',b:b'Ra'\bigwedge bRa)\Rightarrow g(b')=g(b)$不同. 可以将式(7)表示成$R:f\rightarrow g$或者$f\overset{R}{\rightarrow}g$，也可以说$R$满足函数依赖FD: $f\rightarrow g$,这里R为数据集合.

2.2 Tarski量词消除理论下的属性蕴含

形式概念分析(FCA)作为一种数据分析工具,从输入数据中可以产生两种重要的输出:一是概念格,二是属性蕴含. 在FCA的研究中,概念的发现与推理、依赖性的发现与推理以及依赖性的可视化研究是重要的3个方面,其中形式背景的依赖性研究即为属性蕴含,属性蕴含描述形式背景中有效数据的属性依赖,属性蕴含在数据库研究中也称为值依赖.

定义2 (属性蕴含) 设$K=(U,M,I)$为形式背景,其中U为对象的有限集,M为属性的有限集,$I\subseteq U\times M$为对象与属性间的二元关系. M上的属性蕴含为$X\Rightarrow Y$,其中$X\subseteq M$且$Y\subseteq M$. 所有在K中成立的属性蕴含集合记为IMP(K),称为K的蕴含理论[9].

对于在形式背景中成立的属性蕴含$X\Rightarrow Y$,一般有$X'\subseteq Y'$、$\{X\}'\subseteq\{X,Y\}''$、$Y\subseteq X''$三种解释(或判定方法),为研究方便,采用$X'\subseteq Y'$,将其形式化表示为

$$X'\subseteq Y'\equiv$$
$$\{u|\forall u\in U,x\in X,(u,x)\in I\}\subseteq$$
$$\{u|\forall u\in U,y\in Y,(u,y)\in I\}. \tag{8}$$

其中

$$X'=\{u|\forall u\in U,x\in X,(u,x)\in I\},$$
$$Y'=\{u|\forall u\in U,y\in Y,(u,y)\in I\}.$$

对每个对象$u\in U$建立映射$u:M\rightarrow\{0,1\}$,若$(u,x)\in I$,则$u(x)=1$,表示$u(x)$存在,否则$u(x)=0$,表示$u(x)$不存在.

对于u与u'而言,若$u(x)$存在,则$u(x)=1$,同样,若$u'(x)$存在,则$u'(x)=1$,因此$u(x)=u'(x)$表示$u(x)$与$u'(x)$均存在(其值为1). 同理,$u(y)=u'(y)$表示$u(y)$与$u'(y)$均存在(其值为1). 这样式(8)可以解释为对于所有$u,u'\in U$,若$u(x)=u'(x)$成立(或存在),则$u(y)=u'(y)$一定成立,因此,式(8)可以变换为

$$\forall u,u'\in U:(u(x)=u'(x)=1\Rightarrow u(y)=u'(y)=1)\equiv$$
$$\forall u,u'\in U:(u(x)=u'(x)\Rightarrow u(y)=u'(y)). \tag{9}$$

其形式与函数依赖表示相同,可以采用与函数依赖同样的方法将式(9)变换为

$$[U]\cdot x^0\cdot x\cdot[U]^0\subseteq y^0\cdot y. \tag{10}$$

在式(9)中出现的量词在(10)中已没有出现,进一步可将式(10)变换为

$$R\cdot f^0\cdot f\cdot R^0\subseteq g^0\cdot g. \tag{11}$$

式(11)同函数依赖变为

$$(\forall u',u,t',t,tRu:f(u')=f(u)\bigwedge t'Ru')\Rightarrow$$
$$(g(t')=g(t)), \tag{12}$$
$$(\forall u',u,t',t,tRu:f(u')=f(u)=1\bigwedge t'Ru')\Rightarrow$$
$$(g(t')=g(t)=1). \tag{13}$$

注意到,式(12)和(13)中的$f(u')=f(u)$不同于函数依赖,函数依赖表示两个属性值相等,这里则代表$f(u')$、$f(u)$两个属性值均存在(用1表示),$g(t')=g(t)$亦然.

2.3 Tarski代数为基础的函数依赖与属性蕴含关系

前文给出了Tarski理论下函数依赖与属性蕴含无量词表达式统一的数学模型,本节将以Tarski理论为基础,从不同角度对函数依赖与属性蕴含的语义逐步展开研究.

2.3.1 Tarski代数下的语义分析

函数依赖在Tarski量词消除理论下的代数表示为

$$R\cdot f^0\cdot f\cdot R^0\subseteq g^0\cdot g.$$

其中: R为数据库关系表属性值之间的等价关系(或不可分辨关系), f、g为属性到元组之间的映射函

数. 对于包含关系,有

$$R \subseteq S \equiv \forall b, a : bRa \Rightarrow bSa. \tag{14}$$

合并式(6)和(14)可以得到

$$\forall b, a : b(R \cdot f^0 \cdot f \cdot R^0)a \Rightarrow b(g^0 \cdot g)a. \tag{15}$$

由式(15),函数依赖在Tarski量词消除理论下的语义可以解释为:定义元组变量之间的等价关系R(如bRa、$b'Ra'$),由复合函数$f^0 \cdot f$完成属性值之间的关系定义(如$f(a') = f(a)$),定义等价关系R^0,由此推导出$g^0 \cdot g$成立(如$g(b) = g(a)$). 本文推导的函数依赖量化逻辑表示公式(7)充分体现了上述语义.

属性蕴含在Tarski量词消除理论下的代数表示为

$$R \cdot f^0 \cdot f \cdot R^0 \subseteq g^0 \cdot g.$$

其中:R为形式背景中存在的属性值(只有部分属性值存在)之间的等价关系(或不可分辨关系),f、g为由属性域到对象之间的映射函数,对于包含关系,有

$$R \subseteq S \equiv \forall b, a : bRa \Rightarrow bSa.$$

合并式(11)和(14)可以得到

$$\forall b, a : b(R \cdot f^0 \cdot f \cdot R^0)a \Rightarrow b(g^0 \cdot g)a. \tag{16}$$

由式(16),属性蕴含在Tarski量词消除理论下的语义可以解释为:定义元组变量(如a'、a)之间的等价关系R,由复合函数$f^0 \cdot f$完成属性值之间的关系定义(如$f(a') = f(a)$,表示两个属性值均存在),定义等价关系R^0,由此推导出$g^0 \cdot g$成立(如$g(b) = g(a)$,同样指两个属性值均存在). 本文推导的属性蕴含的量化逻辑表示

$$(\forall u', u, t', t, tRu : f(u') = f(u) = 1 \bigwedge t'Ru') \Rightarrow$$
$$(g(t') = g(t) = 1)$$

充分体现了上述语义.

可见,函数依赖与属性蕴含在Tarski量词消除理论下虽然具有统一的数学模型,但是语义截然不同. 下面进一步从FCA及函数的角度对其语义差别进行深入研究.

2.3.2 不同转换背景下"函数依赖成立"与"属性蕴含成立"的关系

虽然函数依赖是数据库领域概念,属性蕴含是FCA领域概念,但实际上一个多值背景即为数据库中的一个关系,因此本节研究不同转换背景下"函数依赖成立"与"属性蕴含成立"两者之间的关系.

按照定义1,在数据库关系中,判定属性集X与Y间函数依赖关系时需要考虑属性值间的关系,设t、s为关系模式下关系r的任意两个元组,它们在属性子集X、Y上的取值分别为$t[X]$、$s[X]$、$t[Y]$、$s[Y]$. 从数学意义上讲,对于$t[X]$、$s[X]$是否相等以及$t[Y]$、$s[Y]$是否相等有4种组合,分别为

$$\text{a)} \begin{cases} t[X] = s[X], \\ t[Y] = s[Y]; \end{cases}$$

$$\text{b)} \begin{cases} t[X] = s[X], \\ t[Y] \neq s[Y]; \end{cases}$$

$$\text{c)} \begin{cases} t[X] \neq s[X], \\ t[Y] = s[Y]; \end{cases}$$

$$\text{d)} \begin{cases} t[X] \neq s[X], \\ t[Y] \neq s[Y]. \end{cases}$$

函数依赖要排除的实际是组合b),而组合c)和d)考虑的是$t[X]$与$s[X]$不等的情况,函数依赖对于$t[Y]$、$s[Y]$是否相等不作要求,所以只剩组合a)真正意义上满足函数依赖的定义,为此可以将组合a)视为函数依赖成立的条件,即

$$\begin{aligned} &\forall t, s \in r, \\ &\text{if } t[X] = s[X] \text{ and } t[Y] = s[Y], \\ &\text{then } X \to Y, \\ &\text{holds.} \end{aligned}$$

下面所研究的函数依赖即为上述情况.

定义3[5] 多值背景(G, M, W, I)由集合G、M、W及其间的三元关系$I(I \subseteq G \times M \times W)$组成,且下式成立:

$$(g, m, w) \in I,\ (g, m, v) \in I \Rightarrow w = v.$$

其中:G的元素称为对象;M的元素称为(多值)属性;W的元素称为属性值;$(g, m, v) \in I$,对于对象g,读作"对象g的m属性具有w",如果w有n个元素,则(G, M, W, I)称为n值背景或多值背景.

形式背景中的属性蕴含$A \Rightarrow B$从数学意义上讲是一种"蕴含式"规则,而"蕴含"的数学意义为$A \Rightarrow B \equiv \neg A \bigvee B$,蕴含式的真值如表1所示.

表1 蕴含真值表

A	B	$A \Rightarrow B$
T	T	T
T	F	F
F	T	T
F	F	T

由表1可以得出属性蕴含"$A \Rightarrow B$"同样包含4种组合:

a) if $A = \mathrm{T}$ and $B = \mathrm{T}$, then $(A \Rightarrow B) \equiv \mathrm{T}$;

b) if $A = \mathrm{T}$ and $B = \mathrm{F}$, then $(A \Rightarrow B) \equiv \mathrm{F}$;

c) if $A = \mathrm{F}$ and $B = \mathrm{T}$, then $(A \Rightarrow B) \equiv \mathrm{T}$;

d) if $A = \mathrm{F}$ and $B = \mathrm{F}$, then $(A \Rightarrow B) \equiv \mathrm{T}$.

在形式背景中,用$x[A] = 1$表示对象x拥有属性A,按照属性蕴含的定义,组合a)是真正满足属性蕴含的定义,组合b)不满足(属性的有效性为假),组合c)和d)考虑的前提是x不拥有属性A,即A取值为假的情况,没有实际意义,一般不予考虑. 因此组合a)是真正意义上的属性蕴含,下面所研究的属性蕴含(也称强属性蕴含)即为上述情况.

定义4 (强属性蕴含)　在形式背景$K = (U, M, I)$中,$\forall x \in U, A、B \subseteq M$,当$x[A]$存在时,$x[B]$一定存在,称此时的属性蕴含$A \Rightarrow B$为强属性蕴含.

定义5 (弱属性蕴含)　在形式背景$K = (U, M, I)$中,$\forall x \in U, A、B \subseteq M$,当$x[A]$不存在时,无论$x[B]$是否存在,称此时的属性蕴含$A \Rightarrow B$为弱属性蕴含.

本文研究转换为"函数依赖成立"与形式背景中"强属性蕴含成立"的关系研究. 首先进行概念标尺(将关系转换为形式背景)的选择,选择的原则是能够将关系转换为形式背景(即多值背景转换为单值背景)且具有通用性.

定义6 [1]　一个用于多值背景的属性m的标尺是一个(单值)背景$S_m = (G_m, M_m, I_m)$,满足$m(G) \subseteq G_m$标尺中的对象是属性m的各个值,标尺中的属性是单值背景中的属性.

为方便起见,本文采用额定标尺.

定义7 [1]　额定标尺$N_n := (n, n, =)$用于取值互相排斥的属性换算,n表示对象,"$=$"表示同一对象在不同的属性上取值是否相等(或存在).

例1 [4]　给出太阳系行星关系示例如表2所示.

表2　太阳系行星关系示例

	planet size			distance from sun		Moon		Period	
	ss	sm	sl	dn	df	my	mn	ps	pl
Mercury	×			×			×	×	
Venus	×			×			×	×	
Earth	×			×		×		×	
Mars	×			×		×		×	
Juplter			×		×	×		×	
Saturn			×		×	×		×	
Uranus		×			×	×			×
Nepun		×			×	×			×
Pluto	×				×	×			×

命题1 [10]　关系模式R中任意关系r上的函数依赖$X \to Y$成立,当且仅当形式背景(O, M, R)上的属性蕴含$X \Rightarrow Y$成立,这里$O = \{(t_1, t_2) | t_1, t_2 \in r, t_1 \neq t_2\}$且$\forall A \in M, (t_1, t_2) R_A \Leftrightarrow t_1[A] = t_2[A]$.

该结论是否普遍成立? 本文将在上述额定标尺下展开研究. 表2中显然属性蕴含distance-near⇒size-small成立,但是由函数依赖的定义,函数依赖distance-near→size-small显然不成立,根据语义,"行星离太阳的距离"并不能函数决定"行星的大小",因此给出如下命题.

命题2　对于任意关系r,(单值)背景为$S_m = (G_m, M_m, I_m)$,其对应的转换背景为额定标尺$N_n := (n, n, =) \leqslant$,当强属性蕴含$X \Rightarrow Y$在背景$S_m$中成立时,函数依赖$X \to Y$在$r$中未必成立.

由命题2可见,在额定标尺背景下,由属性蕴含推导函数依赖较为困难,因为事实上虽然属性蕴含成立但函数依赖不一定成立. 当函数依赖在此背景下成立时,相应的属性蕴含是否成立? 仍以表2为例,由行星的大小与其距太阳的距离可以决定其运行周期,函数依赖(planet size, distance fromsun)→period成立,但属性蕴含(planet size, distance from sun)⇒Period不成立,因此给出如下命题.

命题3　对于任意关系r,(单值)背景为$S_m = (G_m, M_m, I_m)$,其对应的转换背景为额定标尺$N_n := (n, n, =)$,当函数依赖$X \to Y$成立时,强属性蕴含$X \Rightarrow Y$在背景S_m中未必成立.

由命题3可见,在额定标尺背景下,从函数依赖集推导属性蕴含集也是困难的,因为事实上虽然函数依赖成立但是属性蕴含不一定成立.

那么众多学者认可的"命题1"应该如何理解?它在什么情况下成立? 首先分析文献[10]对于属性蕴含的定义及其结论.

定义8　背景(O, I, R)中的属性蕴含$I_1 \Rightarrow I_2$是I的子集序对,对于背景对象所有内涵I而言,$I_1 \not\subseteq I$或$I_2 \subseteq I$(为全文符号表示统一起见,用"⇒"替代原文的"→").

定义8中,"$I_1 \not\subseteq I$或$I_2 \subseteq I$"可以看作如下3种组合:

$$\text{a)} \begin{cases} I_1 \not\subseteq I, \\ I_2 \subseteq I; \end{cases} \quad \text{b)} \begin{cases} I_1 \not\subseteq I; \\ I_2 \not\subseteq I; \end{cases} \quad \text{c)} \begin{cases} I_1 \subseteq I, \\ I_2 \subseteq I. \end{cases}$$

组合a)和组合b)实际上是前文定义的"弱属性蕴含",组合c)是"强属性蕴含". 注意到,命题1中对象取值为$O = \{(t_1, t_2) | t_1, t_2 \in r, t_1 \neq t_2\}$,而任意两个元组间的关系定义为$\forall A \in M, (t_1, t_2) R_A \Leftrightarrow t_1[A] = t_2[A]$,显然采用的是如下转换背景.

定义9 [3](等价转换背景)　设$K = (B_2(G), M,$

$I), G = T$(相当于关系表), M为属性,对应于FCA中的概念. $B_2(G) = \{(t_i, t_j)|i \prec j$且$t_i, t_j \in T\}$为$G$中序对的集合, I为$(t_i, t_j)Im \Leftrightarrow t_i(m) = t_j(m), m \in M$.

定义9将原来数据库关系中"相等的属性值"对应的对象进行组合,形成二元组作为对象,组合后"相等的属性值"必然出现在转换后的形式背景中,因此原来数据库关系中成立的函数依赖在转换后的形式背景中属性蕴含也必然成立. 文献[11]正是采用了与文献[10]相同的转换背景进而得出等价结论.

断言1 对于每个形式背景$K = (G, M, I)$,存在关系r使得$X \Rightarrow Y$是形式背景K的属性蕴含,当且仅当$X \to Y$是关系r的函数依赖.

由以上研究可见,命题1和断言1成立的前提是其特殊的等价转换背景,并不是在所有转换背景下都成立. 从另一方面说,将经典的关系数据库转换为形式背景,为了保持原有的函数依赖,需要采用类似定义9中特殊的转换背景,这样在转换后的形式背景中与原函数依赖等价的属性蕴含才能保持.

2.3.3 从函数的观点看函数依赖和属性蕴含

文献[5]已经证明,从函数依赖和属性蕴含的函数表示来看它们是同构的,可以用统一的数学模型表示,即函数依赖和属性蕴含均可以用如下函数映射表示: 设有$f: W^X \to W^Y$,使得$f(m(g)|m \in X) = (n(g)|n \in Y)$对所有对象(或元组)$g \in G$成立,其中$W^X$、$W^Y$为属性$X$、$Y$的值域.

需要说明的是,函数依赖和属性蕴含虽然可以用数学的"函数"映射模型表示,但是数据库中的"函数依赖"和FCA中的"属性蕴含"与数学意义上的"函数"仍有差别,数学上的"函数"研究主要涉及函数的有界性、单调性、奇偶性、周期性、连续性和凹凸性等,而"函数依赖"和"属性蕴含"所涉及的"函数"则不同,说明如下.

为研究方便,假设数学意义上的函数为$F: x \to y$,对F而言,当自变量x固定时,对应的y值也固定;在数据库中,假设X、Y为属性, DOM(X)、DOM(Y)分别为其值域, f为随时间变化的函数, f: DOM$(X) \to$ DOM(Y). f与F不同,它允许随着时间变化其中数据库关系的属性值加以变化.

例2 关系模式. 学生(学号,姓名,年龄,年级,性别,总成绩)中,有函数依赖:学号$\to$年龄,年龄值每年变化.

为区别数学的函数,省略DOM,将f: DOM$(X) \to$ DOM(Y)记为$f: X \to Y$. 若有函数依赖$f: X \to Y$,则称Y函数依赖于X[12]. 对于属性蕴含而言,由于形式背景由数据库关系转化而来,设有属性蕴含$X \Rightarrow Y$,表示当对象在属性X上取值时(设值为1),一定在属性Y上取值(值为1),意味着当对象在X上的取值存在时在Y上的取值也是存在的. 从函数的观点看,属性蕴含与函数依赖不同,与数学上"函数"的取值有些类似.

3 Armstrong公理及Duquenne-Guigues基

3.1 对Armstrong公理的满足性

文献[5]已经证明函数依赖与属性蕴含均满足Armstrong公理及其附加的推理规则:

1) 自反律,若$Y \subseteq X \subseteq U$,则$X \to Y$为$F$所蕴含;

2) 增广律,若$X \to Y$为F所蕴含,且$Z \subseteq U$,则$XZ \to YZ$为F所蕴含;

3) 传递律,若$X \to Y, Y \to Z$为F所蕴含,则$X \to Z$为F所蕴含;

4) 合并规则,若$X \to Y, X \to Z$,则$X \to YZ$为F蕴含;

5) 伪传递规则,若$X \to Y, WY \to Z$,则$XW \to Z$为F所蕴含;

6) 分解规则,若$X \to Y, Z \subseteq Y$,则$X \to Z$为F所蕴含.

3.2 对Duquenne-Guigues基的满足性

3.2.1 函数依赖的最小依赖集

定义10[13] 如果函数依赖集F满足如下条件,则称F为一个极小函数依赖集:

1) F中任一函数依赖的右部仅含有一个属性;

2) F中不存在$X \to A$,使得F与$F - \{X \to A\}$等价;

3) F中不存在这样的$X \to A$, X有真子集Z使得$F - \{X \to A\} \bigcup \{Z \to A\}$与$F$等价.

3.2.2 属性蕴含的Duquenne-Guigues基

定义11 (伪内涵)[14] 设$Y \subseteq$ DOM,若对于每个伪内涵$Y_1 \subseteq Y$,都有$f(g(Y_1)) \subseteq Y$,且$Y \neq f(g(Y))$,则Y是一个伪内涵.

伪内涵由以下原因得名: 若Y是一个内涵,则$Y = f(g(Y))$;另一方面$Y_1 \subseteq Y, f(g(Y_1)) \subseteq f(g(Y))$,所以$f(g(Y_1)) \subseteq Y$必然成立. "伪内涵"与"内涵"的区别在于$Y$与$f(g(Y)$相等与否.

定义12 (Duquenne-Guigues基)[14] 设$K = (G, M, I)$为一个形式背景,则属性蕴含集合$\{X \to (f(g(X)) - X)|X$是K的伪函数$\}$称为K的Duquenne-Guigues基,简称G-D基.

对于G-D基,学者们有以下结论.

定理1[15] 称G-D基为$K = (G, M, I)$完全的、

无冗余的依赖基.

3.2.3 最小依赖集与G-D基的关系

文献[3]指出,数据表T中函数依赖的最小依赖集与形式背景K(由T转换而来)中属性蕴含的G-D基是相同的,该结论是否具有通用性,分析如下.

由命题1～命题3可知,从数据库数据表向形式背景转换的背景不同,则其得出的属性蕴含也不同,从而函数依赖的最小依赖集与属性蕴含的G-D基也不相同.

定理2 从数据库数据表向形式背景转换过程中,当采用等价转换背景时,数据表函数依赖的最小依赖集与形式背景属性蕴含的G-D基是相同的.

证明 采用等价转换背景时,函数依赖与属性蕴含一一对应,它们均满足Armstrong公理,因此函数依赖的最小依赖集与形式背景属性蕴含的G-D基是相同的. □

定理3 从数据库数据表向形式背景转换过程中,当采用非等价转换背景时,由数据表函数依赖的最小依赖集与形式背景属性蕴含的*G-D*基未必相同.

定理3可以由命题2推论得出,此略.

基于本文研究结论,下面给出实例探讨函数依赖与属性蕴含之间关系的实际应用背景,同时分析经典函数依赖对于属性蕴含研究的意义.

例3 给出关系如表3所示,求出其极小函数依赖集,并将该关系表转换为形式背景,求解其最小蕴含集或G-D基.

表3 示例表

T	a	b	c
t_1	1	3	4
t_2	4	3	6
t_3	1	5	4
t_4	4	2	6

因为$A=\{a,b,c\}$为属性集,$T=\{t_1,t_2,t_3,t_4\}$为元组集,按照函数依赖定义,关系表中存在的函数依赖集为

$$F=\{a\to c, c\to a, ab\to c, bc\to a\}.$$

按照Armstrong公理系统的增广律,$ab\to c$可由$a\to c$生成,$bc\to a$可由$c\to a$生成,因此极小函数依赖集为$F'=\{a\to c, c\to a\}$.

关系表3实际是形式概念分析的一个多值背景,按照本文研究结论,利用定义9的等价转换背景,可以将表3转换为形式背景表4.

表4 背景表

K	a	b	c
(t_1,t_2)		×	
(t_1,t_3)	×		×
(t_1,t_4)			
(t_2,t_3)			
(t_2,t_4)	×		×
(t_3,t_4)			

表4中: $K=(B_2(G),M,I)$, $B_2(G)=\{(t_1,t_2),(t_1,t_3),(t_1,t_4),(t_2,t_3),(t_2,t_4),(t_3,t_4)\}$, $M=\{a,b,c\}$, $I\subseteq G\times M$表示G与M间二元关系,"×"表示对象拥有属性. 按照本文的研究结论"定理2",当采用等价转换背景时,关系表函数依赖的最小依赖集与形式背景属性蕴含的G-D基是相同的,由此得到最小蕴含集或G-D基为$AI=\{a\Rightarrow c, c\Rightarrow a\}$.

例3表明,在利用形式概念分析(FCA)理论进行各种各样的数据分析时,需要将经典关系数据库的关系表转换为相应的形式背景,而背景转换的方法有多种,当选择等价转换背景时,可以将由Armstrong公理得到的关系表最小依赖集快速变换为转换后的形式背景最小蕴含集或G-D基.

4 结 论

本文介绍了Tarski提出的量词消除理论,分别研究了Tarski量词消除理论下的函数依赖表示方式及Tarski量词消除理论下的属性蕴含表示方式,得出无量词Tarski代数下函数依赖与属性蕴含表示的统一数学模型;然后,进一步从形式概念分析的角度进行深入研究,得出了不同转换背景下函数依赖与属性蕴含两者成立的关系,并从函数的观点分析两者的根本语义差别;最后,基于Armstrong公理的满足性讨论了函数依赖与属性蕴含所满足的Armstrong公理,基于G-D基的满足性得出最小依赖集与G-D基的关系,并对函数依赖与属性蕴含之间的关系进行全面总结,得出如下结论:

1) 在Tarski代数方法意义下,函数依赖与数据蕴含表现出相同的数学结构,即在Tarski无量词表达方法中,函数依赖与数据蕴含具有统一的数学模型,但语义不同.

2) 在不同转换背景下,函数依赖成立与属性蕴含成立的关系表现出不同的情形: 在等价转换背景下,原来数据库关系中成立的函数依赖将表现为转换后形式背景的属性蕴含;在非等价转换背景下,原来数据库关系中成立的函数依赖与转换后形式背景下的

属性蕴含成立与否没有必然关系.

3) 从函数的观点看,函数依赖与数学上“函数”的语义差异是明显的,而属性蕴含则与数学上“函数”的取值有些类似.

4) 对函数依赖与属性蕴含所满足的Armstrong公理及G-D而言:当采用等价转换背景时,它们均满足Armstrong公理,因此函数依赖的最小依赖集与形式背景属性蕴含的G-D基是相同的;当采用非等价转换背景时,虽然它们均满足Armstrong公理,但是函数依赖的最小依赖集与形式背景属性蕴含的G-D基未必相同.

未来将进行的相关研究包括D-基(D-Basis)的属性蕴含度量、形式概念分析的匹配依赖及利用正、负属性进行不确定建模[16-18]等.

参考文献(References)

[1] Ganter B, Wille R. Formal concept analysis[M]. Berlin: Springer, 1999: 73-74.

[2] Baixeries J,Kaytoue M, Napoli A. Computing functional dependencies with pattern structures[C]. Proceedings of the Ninth International Conference on Concept Lattices and Their Applications. Spain: Laszlo Szathmary, 2012: 175-186.

[3] Baixeries J, Kaytoue M, Napoli A. Characterizing functional dependencies in formal concept analysis with pattern structures[J]. Annals of Mathematics & Artificial Intelligence, 2014, 72(1/2): 129-149.

[4] Carpineto C, Romano G, Adamo P D. Inferring dependencies from relations: A conceptual clustering approach[J]. Computational Intelligence, 2010, 15(4): 415-441.

[5] 安秋生, 孔祥玉. 函数依赖与属性蕴含的关系研究[J]. 小型微型计算机系统, 2017, 38(9): 2000-2005.
(An Q S, Kong X Y. Relationship study of functional dependency and attribute implication[J]. Journal of Chinese Computer Systems, 2017, 38(9): 2000-2005.)

[6] Oliveira J N. A relation-algebraic approach to the “Hoare logic” of functional dependencies[J]. Journal of Logical & Algebraic Methods in Programming, 2014, 83(2): 249-262.

[7] Tarski A. A decision method for elementary algebra and geometry[J]. Texts & Monographs in Symbolic Computation, 1951, 14(3): 1-63.

[8] Pablo A Parrilo. MIT 6.972 algebraic techniques and semidefinite optimization[R]. MIT: Pablo A Parrilo, 2006: 1-2.

[9] Radim Belohlavek. Introduction to formal concept analysis[EB/OL]. (2008-06-10)[2019-01-16]. http://www.doc88. com/p-3397343225220.html.

[10] Stephane Lopes, Jean-Marc Petit, Lotfi Lakhal. Functional and approximate dependency mining: Database and FCA points of view[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2002, 14(2/3): 93-114.

[11] Kuznetsov S O. Machine learning on the basis of formal concept analysis[J]. Automation & Remote Control, 2001, 62(10): 1543-1564.

[12] Bernstein P A. Synthesizing third normal form relations from functional dependencies[J]. Acm Transactions on Database Systems, 1976, 1(4): 277-298.

[13] 萨师煊, 王珊. 数据库系统概论[M]. 北京:高等教育出版社, 2017: 193-194.
(Sa S X, Wang S. Introduction to database system[M]. Beijing: Higher Education Press, 2017: 193-194.)

[14] 李昭原. 数据库技术新进展[M]. 北京: 清华大学出版社, 2007: 20-30.
(Li Z Y. New developments in database technology[M]. Beijing: Tsinghua University Press, 2007: 20-30.)

[15] Lin Y Z, Li H X. Correspondence between the orderings and the quasi-quotient groups on an ordered group[J]. Journal of Beijing Normal University, 2004, 40(4): 433-436.

[16] Kira Adaricheva, Nation J B, Gordon Okimoto, et al. LNAI 9113[M]. Berlin: Springer, 2015: 39-57.

[17] Jaume Baixeries, Victor Codocedo, Mehdi Kaytoue, et al. Characterizing approximate-matching dependencies in formal concept analysis with pattern structures[J]. Discrete Applied Mathematics, 2018, 249(20): 18-27.

[18] Eduard Bart, Jan Konecny. L-Concept lattices with positive and negative attributes: Modeling uncertainty and reduction of size[J]. Information Sciences, 2019, 472: 163-179.

作者简介

安秋生(1966−), 男, 教授, 博士, 从事数据库理论、粗糙集与概念格等研究, E-mail: aaqqss@sina.com;

孔祥玉(1967−), 男, 教授, 博士生导师, 从事神经网络与机器学习等研究, E-mail: xiangyukong01@163.com;

曹泽豪(1993−), 男, 硕士生, 从事系统特征提取、多元统计过程监控的研究, E-mail: 578021174@qq.com.

(责任编辑：郑晓蕾)