文章编号: 1001- 2486(2009) 05- 0058- 06

内置谓词函数依赖及其推理规则*

胡艳丽, 张维明, 肖 卫东, 汤大权, 唐九阳
(国防科技大学 C^4ISR 技术国防科技重点实验室, 湖南 长沙 410073)

摘 要: 研究内置谓词函数依赖及其推理规则。首先提出内置谓词函数依赖, 定义了内置谓词函数依赖的语法和语义; 其次提出属性- 约束集闭包概念, 提出计算属性- 约束集闭包的算法, 判断内置谓词函数依赖逻辑蕴涵; 然后提出内置谓词函数依赖的推理规则集 $\mathcal{A}$, 证明推理规则集 $\mathcal{A}$ 是可靠且完备的, 用于内置谓词函数依赖蕴涵分析的形式化证明; 最后讨论了内置谓词函数依赖的应用。

关键词: 函数依赖; 内置谓词; 推理规则; 逻辑蕴涵; 可靠性; 完备性
中图分类号: TP311. 131 文献标识码: A

Functional Dependencies with Built-in Predicates and Its Axiomatization

HU Yan-li, ZHANG Wei-ming, XIAO Wei-dong, TANG Da-quan, TANG Jie-yang
(Key Laboratory of C^4ISR Technology, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: The increasing demand for data quality technology has motivated revisions of classical dependencies to capture more inconsistencies in real-life data. A class of integrity constraints, referred to as functional dependencies with built-in predicates (PFDs), is proposed for relational databases and their axiomatization is investigated. In contrast to traditional functional dependencies (FDs) developed mainly for schema design, PFDs generalize the notions of FDs to apply to subsets of relations specified by constraints in the context of interpreted data, and aim at capturing the consistency of data by enforcing bindings of ranges of semantically related values. For the implication analysis of PFDs, which is to decide whether or not a set of PFDs entails another PFD, we provide an inference system analogous to Armstrong's axioms for FDs, and prove the soundness and completeness of the inference system. This work is a step towards a practical constraint-based method for improving data quality since inconsistencies and errors in databases often emerge as violations of integrity constraints.

Key words: functional dependencies; built-in predicates; inference rule; logical implication; soundness; completeness

信息技术的普及和发展使得数据成为信息时代最重要的战略资源之一, 为科学研究以及辅助政治、经济、商业、军事等领域的决策提供重要依据。可靠、准确的数据是进行正确决策的基础。数据依赖(data dependency)指定数据满足的语义关系, 是实现数据语义完整性的重要方法。函数依赖是最重要的数据依赖, 然而函数依赖只涉及关系模式层次的信息, 无法表达具有数据取值关联的语义约束。因此本文提出内置谓词函数依赖, 通过内置谓词约束增强函数依赖的表达能力, 并提出可靠且完备的推理规则集进行内置谓词函数依赖蕴涵分析。

1 相关工作

函数依赖(Functional Dependencies, FDs)[1]能够表达属性取值是否相等, 无法表达具有数据取值关联的语义约束。而现实生活中普遍存在具有数据取值关联的语义约束, 如销量超过 50 万份、100 万份和 1000 万份的唱片分别为金唱片、白金唱片和钻石唱片; 再如国际航班经济舱单件托运行李重量大于 25kg 时, 要施以 50 磅的罚款等。因此研究人员纷纷探索增强函数依赖表达能力的方法。

* 收稿日期: 2009- 03- 23
基金项目: 国家自然科学基金资助项目(60902094, 60903225, 70701038, 60903206)
作者简介: 胡艳丽(1979—), 女, 博士生。

条件函数依赖(Conditional Functional Dependency)[2] 通过等值约束增强函数依赖的表达能力, 但无法表达复杂的数据取值关联。受限函数依赖(constrained functional dependencies)[3] 可以定义函数依赖成立的条件, 但无法表达依赖结论属性约束。扩展条件函数依赖(extended conditional functional dependency, eCFD) 对条件函数依赖进行扩展, 表示属性取值约束为有限集合的依赖关系[4]。

数据依赖的公理化是数据库理论研究的重要课题之一。数据依赖推理规则集由一组推理规则构成, 基于这些规则可以从一组数据依赖推理生成新的依赖, 是判断依赖逻辑蕴涵(logical implication) 关系的基础。条件函数依赖和受限函数依赖具有可靠且完备的推理规则, 扩展条件函数依赖是否可公理化则缺乏研究。

综上所述, 建立数据取值关联是增强函数依赖表达能力的主要方法。本文提出内置谓词函数依赖, 在上述依赖基础上进一步增强函数依赖的表达能力, 并深入研究内置谓词函数依赖的推理规则。

2 内置谓词函数依赖

设模式 $R(U)$, $X, Y, Z \subseteq U$ 表示属性集合, $A, B, C \in U$ 表示单个属性。令 $\boldsymbol{V}$ 是表示属性取值的无穷变量集合, 变量的值域 $\boldsymbol{D}$ 为无穷常量集合。

定义 1(内置谓词约束) 设变量 $v \in \boldsymbol{V}$ 表示元组在属性 $A \in U$ 上的取值, 其内置谓词约束 $\varphi(v)$ 可归纳定义如下:

(1) $v\theta c$ 是内置谓词约束;

(2) 若 $\varphi_1(v)$ 和 $\varphi_2(v)$ 是内置谓词约束, 则 $\varphi_1(v) \wedge \varphi_2(v)$ 是内置谓词约束;

(3) 若 $\varphi_1(v)$ 和 $\varphi_2(v)$ 是内置谓词约束, 则 $\varphi_1(v) \vee \varphi_2(v)$ 是内置谓词约束;

(4) 若 $v\theta c$ 是内置谓词约束, 则 $(v\theta c)$ 是内置谓词约束;

因此, 属性 A 取值的内置谓词约束记作 $\varphi(v) ::= v\theta c \mid (v\theta c) \mid \varphi_1(v) \wedge \varphi_2(v) \mid \varphi_1(v) \vee \varphi_2(v)$, 其中, $\theta \in \{=, \leqslant, <, \neq, >, \geqslant\}$, $c \in \boldsymbol{D}$, $\wedge$ 和 $\vee$ 分别是合取和析取联结词, "::=" 表示归纳定义。

设一组变量 $v_1, \cdots, v_n$ 分别表示元组在属性集 $X = A_1, \cdots, A_n$ 上的取值, 其内置谓词约束 $\varphi(v_1, \cdots, v_n)$ 是由相应变量的内置谓词约束构成的 n 元组, 记作 $\varphi(v_1, \cdots, v_n) = (\varphi(v_1), \cdots, \varphi(v_n))$, 其中, $\varphi(v_i)$ 是属性 A_i 的内置谓词约束。

定义 2(赋值) 设属性 A 的内置谓词约束 $\varphi(v)$, 将变量 v 用值域中的常量 c 代替称为对 $\varphi(v)$ 的赋值, 记作 $\varphi(c/v)$。一个赋值适合内置谓词约束 $\varphi(v)$, 当且仅当在这个赋值下 $\varphi(v)$ 为真。值域中所有使 $\varphi(v)$ 为真的常量构成 $\varphi(v)$ 的保真集合, 记作 $C(\varphi(v)) = \{c \mid c \in D$, 且 $\varphi(c/v) = \text{true}\}$。

设内置谓词约束 $\varphi_i(v_j)$ 和 $\varphi_k(v_l)$, 若 $C(\varphi_k(v_l)) \subseteq C(\varphi_i(v_j))$, 则称 $\varphi_k(v_l)$ 严于 $\varphi_i(v_j)$, 记作 $\varphi_k(v_l) \preccurlyeq \varphi_i(v_j)$。若 $C(\varphi_k(v_l)) = C(\varphi_i(v_j))$, 则称 $\varphi_i(v_j)$ 和 $\varphi_k(v_l)$ 等价, 记作 $\varphi_k(v_l) \equiv \varphi_i(v_j)$。

设属性集 X 的内置谓词约束分别为 $\varphi_k(v_1, \cdots, v_n)$ 和 $\varphi_i(v_1, \cdots, v_n)$。若对于每个变量 $v_j \in \{v_1, \cdots, v_n\}$, 其内置谓词约束都满足 $\varphi_k(v_j) \preccurlyeq \varphi_i(v_j)$, 则称 $\varphi_k(v_1, \cdots, v_n)$ 严于 $\varphi_i(v_1, \cdots, v_n)$, 记作 $\varphi_k(v_1, \cdots, v_n) \preccurlyeq \varphi_i(v_1, \cdots, v_n)$。类似地, 可定义 $\varphi_k(v_1, \cdots, v_n)$ 与 $\varphi_i(v_1, \cdots, v_n)$ 等价, 记作 $\varphi_k(v_1, \cdots, v_n) \equiv \varphi_i(v_1, \cdots, v_n)$。

为表达方便, 将属性 A 的内置谓词约束 $\varphi(v)$ 记作为 $\varphi(A)$。类似地, 属性集 X 的内置谓词约束 $\varphi(v_1, \cdots, v_n)$ 简写为 $\varphi(X)$。

定义 3(属性- 约束集) 设属性集 X 及其内置谓词约束 $\varphi(X)$, 属性- 约束集 $S_{(X, \varphi(X))}$ 是由 X 中的属性及其内置谓词约束共同构成的集合, 记作 $S_{(X, \varphi(X))} = \{(A, \varphi(A)) \mid A \in X, \varphi(A)$ 是 $\varphi(X)$ 在属性 A 上的内置谓词约束$\}$。

设属性- 约束集 $S_{(X, \varphi(X))}$ 和 $S_{(Y, \varphi(Y))}$, 若属性- 约束 $(A, \varphi(A)) \in S_{(X, \varphi(X))}$ 都满足 $(A, \varphi(A)) \in S_{(Y, \varphi(Y))}$, 则称 $S_{(X, \varphi(X))}$ 包含于 $S_{(Y, \varphi(Y))}$, 记作 $S_{(X, \varphi(X))} \subseteq S_{(Y, \varphi(Y))}$。

定义 4(内置谓词函数依赖) 设模式 $R(U)$, R 上成立的内置谓词函数依赖形式为 $\Phi = (R: X \rightarrow Y \mid Z, \varphi)$, 其中:

(1) $XYZ \subseteq U$ 且 $Y \cap Z = \cong$, 其中 X 称为 的左部属性集, Y 和 Z 分别称为 的依赖属性集和影响

属性集;

(2)φ是表示元组在属性集XYZ上取值的内置谓词约束,对于属性$A\in XYZ$,$\varphi(A)$是指定元组在属性A上取值的内置谓词约束。

如Φ所含属性对应内置谓词约束的保真集合均不为空,那么Φ称为良构的内置谓词函数依赖。因为内置谓词函数依赖包含有限个属性,且能在多项式时间内判断每个约束的保真集合是否为空,因此以下研究均针对良构的内置谓词函数依赖。

内置谓词函数依赖在取值匹配内置谓词约束的数据集上成立。文献[2]采用符号"$\asymp$"表示数据项与等值约束匹配,类似地,本文采用符号"$\asymp_p$"表示数据项与内置谓词约束匹配。

定义5(匹配) 设模式$R(U)$的实例I,属性A的内置谓词约束$\varphi(A)$,如果属性取值$t[A]$使$\varphi(A)$为真,那么称$t[A]$与$\varphi(A)$匹配,记作$t[A]\asymp_p\varphi(A)$,其中元组$t\in I$。对于约束集$\varphi(X)$,若任意属性$A\in X$均满足$t[A]$与$\varphi(A)$匹配,则称$t[X]$与$\varphi(X)$匹配,记作$t[X]\asymp_p\varphi(X)$。

定义6(内置谓词函数依赖语义) 给定模式R的实例I和R上成立的内置谓词函数依赖$\Phi=(R:X\rightarrow Y|Z,\varphi)$,对于$I$中在属性集$X$上取值满足内置谓词约束$\varphi(X)$的元组集$I_{\varphi(X)}=\{t\in I|t[X]\asymp_p\varphi(X)\}$,有

(1)对于任意元组$t_1,t_2\in I_{\varphi(X)}$,若$t_1[X]=t_2[X]$,则$t_1[Y]=t_2[Y]$;

(2)对于任意元组$t\in I_{\varphi(X)}$,$t[YZ]\asymp_p\varphi(YZ)$;

那么称实例I满足Φ,记作$I\models\Phi$。若实例I中不存在与$\varphi(X)$匹配的元组,则称I平凡满足Φ。给定R上成立的一组内置谓词函数依赖Σ,如果实例I满足Σ中的每一条内置谓词函数依赖,则称I满足Σ,记作$I\models\Sigma$。

根据内置谓词函数依赖语义,函数依赖是内置谓词函数依赖的特例,即函数依赖是所有约束均为'_'的内置谓词函数依赖。

3 内置谓词函数依赖的推理规则

内置谓词函数依赖蕴涵分析是指判断一个内置谓词函数依赖是否被一组内置谓词函数依赖逻辑蕴涵,可靠且完备的推理规则为判断内置谓词函数依赖逻辑蕴涵提供形式化证明。

3.1 内置谓词函数依赖逻辑蕴涵

定义7(逻辑蕴涵) 设Σ是关系模式R上成立的一组内置谓词函数依赖,Φ是一个内置谓词函数依赖,I为关系R的实例。若每个满足Σ的实例I也满足Φ,则称Σ逻辑蕴涵Φ,记作$\Sigma\models\Phi$。

由逻辑蕴涵定义可知,若$\Sigma\not\models\Phi$,则存在满足Σ但不满足Φ的实例,即存在实例$I\models\Sigma$,且$I\not\models\Phi$。

定义8(属性-约束集闭包) 设Σ是模式R上成立的一组内置谓词函数依赖,属性-约束集$S_{(X,\varphi(X))}$基于Σ的依赖属性-约束集闭包和影响属性-约束集闭包分别为$S^*_{(X,\varphi(X))_{fd}}=\{(A,\varphi(A))|\Sigma\models(R:X\rightarrow A|\varnothing,\varphi)\}$和$S^*_{(X,\varphi(X))_{\neg fd}}=\{(B,\varphi(B))|\Sigma\models(R:X\rightarrow\varnothing|B,\varphi)\}$。

引理1 设Σ是模式$R(U)$上成立的一组内置谓词函数依赖,$S_{(X,\varphi(X))}$是属性集$X\subseteq U$及其内置谓词约束$\varphi(X)$构成的属性-约束集,

(1)$\Sigma\models(R:X\rightarrow A|\varnothing,\varphi)$当且仅当$(A,\varphi(A))\in S^*_{(X,\varphi(X))_{fd}}$;

(2)$\Sigma\models(R:X\rightarrow\varnothing|B,\varphi)$当且仅当$(B,\varphi(B))\in S^*_{(X,\varphi(X))_{\neg fd}}$。

根据引理1,给定一组内置谓词函数依赖Σ和内置谓词函数依赖$\Phi=(R:X\rightarrow Y|Z,\varphi)$,如果知道如何计算$S_{(X,\varphi(X))}$基于$\Sigma$的属性-约束集闭包,就可以根据属性集$Y$和$Z$及其内置谓词约束$\varphi(Y)$和$\varphi(Z)$与属性-约束集闭包的关系判断$\Sigma$是否蕴涵$\Phi$。下面给出计算属性-约束集闭包的CLOSURE算法,如图1所示。

CLOSURE算法的输入是模式$R(U)$上成立的一组内置谓词函数依赖Σ和属性-约束集$S_{(X,\varphi(X))}$,输出是$S_{(X,\varphi(X))}$的属性-约束集闭包。CLOSURE算法首先根据$S_{(X,\varphi(X))}$对属性-约束集$fd_closure$和

$nfd\_closure$ 进行初始化(第 1– 2行)。然后选择 Σ 中一组具有相同属性集的内置谓词函数依赖$\{\Phi_k\}$。若属性集 YA 包含于$fd_closure^{(X)}$,属性集 Y 的约束$fd_closure^{(\varphi)}(Y)$严于 $\varphi_k(Y)$,属性 A 的约束 $fd_closure^{(\varphi)}(A)$严于$\{\Phi_k\}$对属性 A 内置谓词约束的析取$\vee_{k=1}^{n}\varphi_k(A)$,则对于每个依赖属性 $B\in Z$,将$\{(B, \vee_{k_1=1}^{n}\varphi_{k_1}(B))\}$加入$fd_closure$(第 5– 7 行)。类似地,对于任意属性 $D\in Y$,若 D 包含于属性集$nfd_closure^{(X)}$,约束 $nfd_closure^{(\varphi)}(D)$严于 $\varphi_k(D)$,或者 $\varphi_k(D)$等价于‘_’,对于属性 A,若 A 包含于属性集$nfd_closure^{(X)}$,约束 $nfd_closure^{(\varphi)}(A)$严于$\vee_{k=1}^{n}\varphi_k(A)$,或者$\vee_{k=1}^{n}\varphi_k(A)$等价于‘_’,则对于每个影响属性 $C\in W$,将$\{(C, \vee_{k_1=1}^{n}\varphi_{k_1}(C))\}$加入 $nfd\_closure$(第 8– 10 行)。重复上述过程直至$fd_closure$和 $nfd\_closure$ 不再发生变化。

Input: A set Σ of FDPs and the set of $S_{(X,\varphi(X))}$
Output: the closure set of $S_{(X,\varphi(X))}$ under Σ

1 $fd\_closure := S_{(X,\varphi(X))}$;
2 $nfd\_closure := S_{(X,\varphi(X))}$;
3 repeat until no further change
4 Pick $\{(R: YA \rightarrow Z \mid W, \varphi_k) \mid k=1,2,\cdots,n\}$ from Σ with equivalent $\varphi_k(Y)$;
5 if $YA \subseteq fd\_closure^{(X)}$, $fd\_closure^{(\varphi)}(Y) \preceq \varphi_k(Y)$ and $fd\_closure^{(\varphi)}(A) \preceq (\vee_{k=1}^{n}\varphi_k(A))$ then
6 for each $B\in Z$
7 $fd\_closure := fd\_closure \cup \{(B, \vee_{k=1}^{n}\varphi_k(B))\}$;
8 if for any $D\in Y$, either $D\in nfd\_closure^{(X)}$, $nfd\_closure^{(\varphi)}(D) \preceq \varphi_k(D)$, or $\varphi_k(D)=$‘_’, and for A, either $A\in nfd\_closure^{(X)}$, $nfd\_closure^{(\varphi)}(A) \preceq (\vee_{k=1}^{n}\varphi_k(A))$, or $\vee_{k=1}^{n}\varphi_k(A)=$‘_’ then
9 for each $C\in W$
10 $nfd\_closure := nfd\_closure \cup \{(C, \vee_{k=1}^{n}\varphi_k(C))\}$;
11 return $fd\_closure$ and $nfd\_closure$;

图 1 CLOSURE 算法
Fig. 1 Algorithm CLOSURE

与函数依赖属性集闭包算法的证明类似,易证 CLOSURE 算法是计算属性– 约束集闭包的正确算法。

命题 1 设 Σ 是模式$R(U)$上成立的一组内置谓词函数依赖,$S_{(X,\varphi(X))}$是属性集 $X\subseteq U$ 及其内置谓词约束 $\varphi(X)$ 构成的属性– 约束集,CLOSURE 算法计算 $S_{(X,\varphi(X))}$ 的依赖属性– 约束集闭包 $S^{*}_{(X,\varphi(X))_{fd}}$ 和影响属性– 约束集闭包 $S^{*}_{(X,\varphi(X))_{-fd}}$。

3.2 推理规则

推理规则是数据依赖理论的重要研究课题之一,可靠且完备的推理规则通过推理规则(inference rules)支持逻辑蕴涵的形式化证明。图 2 给出内置谓词函数依赖的推理规则集 $\mathscr{A}$。

规则 1 为自反律,是对 Armstrong 公理系统自反律的推广,表示平凡的内置谓词函数依赖。规则 2 为增广律,是对 Armstrong 公理系统增广律的推广。规则 3 和规则 4 为传递律,是对 Armstrong 公理系统传递律的推广。规则 5 表明,对于属性集 XC 和 YZ,如果元组在属性集 XC 上的取值决定在属性集 Y 上的取值,并且在属性集 XC 上的取值匹配 $\varphi_i(XC)$和 $\varphi_j(XC)$时,在属性集 Y 上的取值分别匹配 $\varphi_i(Y)$和 $\varphi_j(Y)$,那么当元组在属性集 XC 上的取值匹配 $\varphi_i(XC)$或 $\varphi_j(XC)$时,在属性集 Y 上的取值匹配 $\varphi_i(Y)$或 $\varphi_j(Y)$;属性集 XC 影响属性集 Z 的情况类似。规则6 表明,对于属性集 XW 和 YZ,如果元组在属性集 X 和 W 上的取值分别决定在属性集 Y 上的取值,并且在属性集 X 和 W 上的取值分别匹配 $\varphi_i(X)$和 $\varphi_j(W)$时,在属性集 Y 上的取值分别匹配 $\varphi_i(Y)$和 $\varphi_j(Y)$,那么当元组在属性集 X 和 W 上的取值同时匹配 $\varphi_i(X)$和 $\varphi_j(W)$时,元组在属性集 Y 上的取值匹配 $\varphi_i(Y)$且 $\varphi_j(Y)$。属性集 X 和 W 影响属性集 Z 的情况类似。规则 7 表明,对于属性集 XA 和 Z,如果当元组在属性 A 上的取值为其值域中的任意常量

时,元组在属性集 Z 上的取值都满足内置谓词约束 $\varphi_i(Z)$,那么元组在属性 A 上的取值对元组在属性集 Z 上的取值无影响,属性集 Z 只受属性集 X 的影响。

规则 1	如果 $Y\subseteq X$,那么 $(R:X\rightarrow Y\mid \cong,\ \varphi_i)$ 和 $(R:X\rightarrow \cong\mid Y,\ \varphi_j)$ 在 R 上成立,其中对于任意属性 $A\in Y$,$\varphi_i(A_L)\preceq\varphi_i(A_R)$ 且 $\varphi_j(A_L)\preceq\varphi_j(A_R)$,$A_L$ 和 A_R 分别表示包含于依赖左部和右部的属性 A;
规则 2	如果 $(R:X\rightarrow Y\mid Z,\ \varphi_i)$ 在 R 上成立,那么 $(R:XW\rightarrow YW\mid Z,\ \varphi_j)$ 和 $(R:XW\rightarrow Y\mid ZW,\ \varphi_k)$ 在 R 上成立,其中 $\varphi_i(XYZ)\equiv\varphi_j(XYZ)\equiv\varphi_k(XYZ)$,$\varphi_j(W_L)\preceq\varphi_j(W_R)$ 且 $\varphi_k(W_L)\preceq\varphi_k(W_R)$,$W_L$ 和 W_R 分别表示位于依赖左部和右部的属性集 W;
规则 3	如果 $(R:X\rightarrow Y\mid Z,\ \varphi_i)$ 和 $(R:Y\rightarrow W\mid V,\ \varphi_j)$ 在 R 上成立,其中 $\varphi_i(Y)\preceq\varphi_j(Y)$,那么 $(R:X\rightarrow W\mid ZV,\ \varphi_k)$ 在 R 上成立,其中 $\varphi_i(XZ)\equiv\varphi_k(XZ)$,且 $\varphi_j(WV)\equiv\varphi_k(WV)$;
规则 4	如果 $(R:X\rightarrow Y\mid Z,\ \varphi_i)$ 和 $(R:Z\rightarrow \cong\mid V,\ \varphi_j)$ 在 R 上成立,其中 $\varphi_i(Z)\preceq\varphi_j(Z)$,那么 $(R:X\rightarrow Y\mid V,\ \varphi_k)$ 在 R 上成立,其中 $\varphi_i(XY)\equiv\varphi_k(XY)$,且 $\varphi_j(V)\equiv\varphi_k(V)$;
规则 5	如果 $(R:XC\rightarrow Y\mid Z,\ \varphi_i)$ 且 $(R:XC\rightarrow Y\mid Z,\ \varphi_j)$,其中 $\varphi_i(X)\equiv\varphi_j(X)$,那么 $(R:XC\rightarrow Y\mid Z,\ \varphi_k)$,其中 $\varphi_i(X)\equiv\varphi_k(X)$,$\varphi_k(C)\equiv\varphi_i(C)\vee\varphi_j(C)$,对于任意属性 $A\in Y$ 和 $B\in Z$,$\varphi_k(A)\equiv\varphi_i(A)\vee\varphi_j(A)$ 和 $\varphi_k(B)\equiv\varphi_i(B)\vee\varphi_j(B)$;
规则 6	如果 $(R:X\rightarrow Y\mid Z,\ \varphi_i)$ 和 $(R:W\rightarrow Y\mid Z,\ \varphi_j)$ 在 R 上成立,那么 $(R:XW\rightarrow Y\mid Z,\ \varphi_k)$ 在 R 上成立,其中 $\varphi_i(X)\equiv\varphi_k(X)$,$\varphi_j(W)\equiv\varphi_k(W)$,对于属性 $A\in Y$ 和 $B\in Z$,$\varphi_k(A)\equiv\varphi_i(A)\wedge\varphi_j(A)$ 和 $\varphi_k(B)\equiv\varphi_i(B)\wedge\varphi_j(B)$;
规则 7	如果 $(R:XA\rightarrow \cong\mid Z,\ \varphi_i)$ 在 R 上成立,其中 $\varphi_i(A)\equiv$ '_',那么 $(R:X\rightarrow \cong\mid Z,\ \varphi_j)$ 在 R 上成立,其中 $\varphi_i(XZ)\equiv\varphi_j(XZ)$。

图2 内置谓词函数依赖的推理规则集 $\mathscr{A}$

Fig. 2 Inference system $\mathscr{A}$ for FDPs

定理1 推理规则 $\mathscr{A}$ 是可靠且完备的。

证明 由内置谓词函数依赖语义易证推理规则 $\mathscr{A}$ 是可靠的。

下面证明推理规则集 $\mathscr{A}$ 是完备的,即给定一组内置谓词函数依赖 Σ 和一个内置谓词函数依赖 $\Phi=(R:X\rightarrow Y\mid Z,\varphi)$,若 Σ 逻辑蕴涵 ,则 可由 Σ 根据推理规则集 $\mathscr{A}$ 推导得到,记作 $\Sigma\vdash_{\mathscr{A}}\Phi$。

可由 Σ 根据推理规则集 $\mathscr{A}$ 推导得到是指,存在一个内置谓词函数依赖序列,这个序列的最后一个依赖是 ,且序列中每个依赖属于 Σ 或可从序列中位于该依赖前的有限个依赖根据推理规则集 $\mathscr{A}$ 推出。该序列称为 的推导序列。

由命题1可知,CLOSURE 算法是证明内置谓词函数依赖逻辑蕴涵的正确算法。因此欲证推理规则集 $\mathscr{A}$ 是完备的,只需证明 CLOSURE 算法等价于基于推理规则集 $\mathscr{A}$ 的推导序列,于是若 Σ 逻辑蕴涵 ,则存在上述基于推理规则集 $\mathscr{A}$ 的推导序列由 Σ 导出 。

下面采用归纳法证明 CLOSURE 算法等价于基于推理规则集 $\mathscr{A}$ 的推导序列。

CLOSURE 算法首先根据属性- 规则集 $S_{(X,\varphi(X))}$ 对 $fd\_closure$ 和 $nfd\_closure$ 进行初始化,得到 $fd\_closure_0$ 和 $nfd\_closure_0$。对于任意属性 $A_i\in X$,由 $(A_i,\varphi(A_i))\in S^{*}_{(X,\varphi(X))_{fd}}$,且 $(A_i,\varphi(A_i))\in S^{*}_{(X,\varphi(X))_{\neg fd}}$ 可得 $\Sigma\vDash\Phi=(R:X\rightarrow A_i\mid\cong,\varphi_i)$,且 $\Sigma\vDash\Phi=(R:X\rightarrow\cong\mid A_i,\varphi_i)$,其中,$\varphi_i(X)\equiv\varphi(X)$,$\varphi_i(A_i)\equiv\varphi(A_i)$。

根据规则1,对于 $A_i\in X$,可得 $\Sigma\vdash\mathscr{A}(R:X\rightarrow A_i\mid\cong,\varphi_i)$,且 $\Sigma\vdash\mathscr{A}(R:X\rightarrow\cong\mid A_i,\varphi_i)$。

设 CLOSURE 算法第 j 次循环得到属性- 约束集 $fd\_closure_j$ 和 $nfd\_closure_j$,对于任意属性 $A_i\in fd\_closure_j^{(X)}$ 和 $B_j\in nfd\_closure_j^{(X)}$,假设存在基于推理规则集 $\mathscr{A}$ 的推导序列 $\Sigma\vdash\mathscr{A}(R:X\rightarrow A_i\mid B_j,\varphi)$。

第 $(j+1)$ 次循环,CLOSURE 算法选择一组内置谓词函数依赖 $\{\Phi_{k_1}=(R:YA\rightarrow Z\mid W,\varphi_{k_1})\mid k_1=1,2,\cdots,n\}\subseteq\Sigma$,其中,$\{\Phi_{k_1}\}$ 在属性集 Y 上具有等价的内置谓词约束。

若上述依赖满足条件,则对于任意属性 $B\in Z$ 和 $C\in W$,得到 $(B,\vee_{k_1=1}^{n}\varphi_{k_1}(B))\in fd\_closure_{j+1}$,且

$(C, \vee_{k_1=1}^{n} \varphi_{k_1}(C)) \in fd\_closure_{j+1}$。

下面证明这一过程可根据基于推理规则集 $\mathscr{A}$ 的推导序列得出。

根据规则 1,由 $\approx \subseteq W$ 可得一组内置谓词函数依赖 $\{\Phi_{k_2} = (R: W \rightarrow \approx \mid \approx, \varphi_{k_2}) \mid k_2 = 1, 2, \cdots, n\}$,其中,对 W 中的任意属性 D, $\varphi_{k_2}(D)$ 等价于 $\varphi_{k_1}(D)$。

根据规则 4,由 $\{\Phi_{k_1}\}$ 和 $\{\Phi_{k_2}\}$ 可得 $\{\Phi_{k_3} = (R: YA \rightarrow Z \mid \approx, \varphi_{k_3}) \mid k_3 = 1, 2, \cdots, n\}$,其中,对 YAZ 中的任意属性 D, $\varphi_{k_3}(D)$ 等价于 $\varphi_{k_1}(D)$。

根据规则 1,由 $YA \subseteq fd\_dosure_j^{(X)}$, $fd\_dosure_j^{(\varphi)}(Y) \preceq \varphi_{k_1}(Y)$,且 $fd\_closure_j^{(\varphi)}(A) \preceq \vee_{k_1=1}^{n} \varphi_{k_1}(A)$ 可得 $\Phi_{k_4} = (R: fd\_dosure_j^{(X)} \rightarrow YA \mid \approx, \varphi_{k_4})$,其中, $\varphi_{k_4}(fd\_dosure_j^{(X)}) \equiv fd\_closure_j^{(\varphi)}$, $\varphi_{k_4}(Y) \equiv \varphi_{k_1}(Y)$, $\varphi_{k_4}(A) \equiv \vee_{k_1=1}^{n} \varphi_{k_1}(A)$。

根据规则 1,对于任意属性 $B \in Z$,由 $\varphi_{k_1}(B) \preceq \vee_{k_1=1}^{n} \varphi_{k_1}(B)$,可得一组内置谓词函数依赖 $\{\Phi_{k_5} = (R: Z \rightarrow B \mid \approx, \varphi_{k_5}) \mid k_5 = 1, 2, \cdots, n\}$,其中, $\varphi_{k_5}(Z) \equiv \varphi_{k_1}(Z)$, $\varphi_{k_5}(B) \equiv \vee_{k_1=1}^{n} \varphi_{k_1}(B)$。

根据规则 3,由 $\{\Phi_{k_3}\}$ 和 $\{\Phi_{k_5}\}$ 可得 $\{\Phi_{k_6} = (R: YA \rightarrow Z \mid \approx, \varphi_{k_6}) \mid k_6 = 1, 2, \cdots, n\}$,其中, $\varphi_{k_6}(YA) \equiv \varphi_{k_1}(YA)$, $\varphi_{k_6}(B) \equiv \vee_{k_1=1}^{n} \varphi_{k_1}(B)$。

根据规则 5,由 $\{\Phi_{k_6}\}$ 可得 $\Phi_{k_7} = (R: YA \rightarrow B \mid \approx, \varphi_{k_7})$,其中, $\varphi_{k_7}(Y) \equiv \varphi_{k_1}(Y)$, $\varphi_{k_7}(A) \equiv \vee_{k_1=1}^{n} \varphi_{k_1}(A)$, $\varphi_{k_7}(B) \equiv \vee_{k_1=1}^{n} \varphi_{k_1}(B)$。

根据规则 3,由 Φ_{k_4} 和 Φ_{k_7} 可得 $\Phi_{k_8} = (R: fd\_closure_j^{(X)} \rightarrow B \mid \approx, \varphi_{k_8})$,其中, $\varphi_{k_8}(fd\_closure_j^{(X)}) \equiv fd\_closure_j^{(\varphi)}$, $\varphi_{k_8}(B) \equiv \vee_{k_1=1}^{n} \varphi_{k_1}(B)$。

因此,CLOSURE 算法计算 $(B, \vee_{k_1=1}^{n} \varphi_{k_1}(B)) \in fd\_closure_{j+1}$ 的步骤等价于基于推理规则集 $\mathscr{A}$ 的上述推导序列。

同理,由上述规则和规则 6、规则 7 可得,CLOSURE 算法计算 $(C, \vee_{k_1=1}^{n} \varphi_{k_1}(C)) \in fd\_closure_{j+1}$ 的步骤等价于基于推理规则集 $\mathscr{A}$ 的推导序列。

重复上述过程,直至 CLOSURE 算法结束,可知 CLOSURE 算法等价于基于推理规则集 $\mathscr{A}$ 的推导序列。进一步,根据规则 2,可得若 Σ 逻辑蕴涵 ,则存在基于推理规则集 $\mathscr{A}$ 的推导序列 $\Sigma \vdash_{\mathscr{A}}$,因此推理规则集 $\mathscr{A}$ 是完备的。

证毕。

推理规则集 $\mathscr{A}$ 是可靠且完备的保证基于推理规则集 $\mathscr{A}$ 推导得到的内置谓词函数依赖被逻辑蕴涵,同时逻辑蕴涵的内置谓词函数依赖一定可以通过基于推理规则集 $\mathscr{A}$ 的推导序列推导得到。

4 结 论

内置谓词函数依赖通过内置谓词约束指定数据取值关联,增强函数依赖的表达能力,并且具有可靠且完备的推理规则集。内置谓词函数依赖可以检测到现实应用中大量的不一致数据,因此用于数据清洗,修复不一致数据,为实现数据不一致性的自动检测、减少清洗过程的人工干预、提高大规模数据清洗的效率提供了基础。下一步将研究如何基于内置谓词函数依赖修复不一致数据。

致谢 感谢英国爱丁堡大学信息学院樊文飞教授对作者的指导。

参 考 文 献:

[1] Codd E F. Relational Completeness of Data Base Sublanguages[R]. Englewood Cliffs, N.J.: PrenticeHall, 1972.
[2] Fan W F, Geerts F, Jia X B, et al. Conditional Functional Dependencies for Capturing Data Inconsistencies [J]. ACM Transactions on Database Systems (TODS), 2008, 33(2).
[3] Maher M J. Constrained Dependencies[J]. Theoretical Computer Science, 1997, 173(1): 113-149.
[4] Loreto B, Fan W F, Geerts F, et al. Increasing the Expressivity of Conditional Functional Dependencies without Extra Complexity[C]//The 24th International Conference on Database Engineering (ICDE), 2008.