连接依赖若干问题的讨论

刘方鑫　鲍剑洋（中国矿业大学）

摘　要

In this paper, a more strict definition about join dependence is presented. Some contradictory which caused by the former definition has been eliminated. The relation between the lossless decomposition and the restrain condition is found. At last, a way to judge whether a join dependence exists in the given relation is proved.

一、引言

关系数据库简单灵活，数据独立性高，理论严密，因此获得人们的普遍重视。在关系数据模型中数据必须满足一些约束条件，且数据的语义性质集中体现在依赖约束上。随着对关系数据理论研究的不断深入，已提出各种形式的数据依赖：函数依赖（FD）、多值依赖（MVD）、层次依赖（HD）、连接依赖（JD）等，并相应地给出第一到第五范式的定义。目前文献中对第一范式到第四范式规范化都有较充分的分析和论证，然而，对如何从第四范式规范为第五范式的过程仍缺乏详尽的合理论证，原因在于对连接依赖的定义不够严密，目前文献[1][3][4][6][7]给出的连接依赖的定义，有时会导出不正确的结果。由于连接依赖是一种强有力的数据约束，从某种意义上讲，FD，MVD，HD等都可视为连接依赖的特殊情况，所以有必要对连接依赖作进一步研究，并给出一种更严格的定义。

二、问题的提出

对于连接依赖（JD），现在通常有如下两类定义：

定义2.1[1][3][6]　关系R满足连接依赖$JD\bowtie(X, Y, \cdots, Z)$的充分必要条件是：R是它在X，Y，…，Z上的投影的自然连接，其中X，Y，…，Z是R的属性集U的某些子集，$X\cup Y\cup\cdots\cup Z=U$

定义2.2[4][7]　如果无论何时，关系R的元组μ_1，μ_2，…，μ_n都是对于结果t在X_1，X_2，…，X_n上可连接的，其中t也是R中的一个元组，则称R满足在$X_1\cup X_2\cup\cdots\cup X_n$上的连接依赖$JD\bowtie(X_1, X_2, \cdots, X_n)$

这里所谓元组μ_1，μ_2，…，μ_n是对于结果t在X_1，X_2，…，X_n上可连接的是指，如果在$\bigcup_{i=1}^{n} X_i$上的关系R的n个元组μ_1，μ_2，…，μ_n（不必是可区别的）。存在一个定义在$\bigcup_{i=1}^{n} X_i$上的映射t使得对所有$1\leqslant i\leqslant n$，$\mu_i[X_i]=t[X_i]$。

相应地，Fagin给出了第五范式的定义：

定义2.3　一个关系R属于第五范式（5NF），当且仅当R中的每个连接依赖都由它的候选关键字所蕴含。

上述关于连接依赖的两个定义虽然形式上不同，但实质上是一致的，它们都不够严密，都忽视对冗余分解的考虑，如果用以进行关系判断，有时会得出一些不正确的结论。

例如：关系S(S#，SNAME，STATUS，CITY)，其候选关键字是S#和SNAME，这个关系通常被认为是属于第五范式[1][6]，但如果按上述连接依赖的定义会导出$S\notin 5NF$的

结果。因为关系S也满足下面的JD $\bowtie$((S#, SNAME), (S#, STATUS), (STATUS, CITY), (S#, CITY)),而这个JD不是由候选关键字S#和SNAME所蕴含的，所以根据定义2.3，关系S不属于5NF，这就产生了矛盾。

实际上，若关系R存在一个连接依赖，根据上述连接依赖定义，则我们几乎总可以找到一个不被R的关键字所蕴含的连接依赖，因此，定义2.1和2.2有关连接依赖的定义是不够严密的。

三、预备知识

定义3.1 给定属性集$U=\{D_1, D_2, \cdots, D_n\}$，如果R是有序的n元组$\langle d_1, d_2, \cdots, d_n\rangle$的集合，其中$d_i \in D_i$ $(i=1, 2, \cdots, n)$，则称R是这个n个集合上的一个关系，记为$R\langle U, \Sigma\rangle$，U是关系R的域，$\Sigma$是R的属性之间的依赖关系集。

由定义3.1可知，一个关系是元组的集合，关系$R\langle U, \Sigma\rangle$可简记为R。在这里元组的有序是指元组中值与属性之间的对应关系，所以关系中属性的顺序是无关紧要的。

为讨论方便起见，我们用符号A,B,C,…代表单个属性，用符号Z, Y, X, …代表属性集，大写字母代表属性，小写字母代表属性值。

如果μ是关系R的一个元组，X是一个属性集，那么$\mu[x]$表示μ元组在X列上的值，关系$R\langle U, \Sigma\rangle$($U=\{X, Y, \cdots, Z\}$)的元组记为$R[X, Y, \cdots, Z]$。

关系中两种最重要的运算是投影和自然连接：

定义3.2 设给定关系$R\langle U, \Sigma\rangle$及$S\langle U', \Sigma'\rangle$，其中$U=\{X, Y\}$，$U'=\{X\}$，如果$S[x]\in S \Leftrightarrow \exists y(R[x, y]\in R)$，则称S是R在U'上的投影，记为$S=\pi_{U'}R$

定义3.3 给定关系$R\langle U_1, \Sigma_1\rangle$，$S\langle U_2, \Sigma_2\rangle$及$T\langle U_3, \Sigma_3\rangle$，$U_1=\{X, Y\}$，$U_2=\{Y, Z\}$，$U_3=\{X, Y, Z\}$，$X\cap Z=\phi$，$B=U_1\cap U_2$，若$T[x, y, z]\in T \Leftrightarrow R[x, y]\in R \wedge S[y, z]\in S$，则称T是R和S的自然连接，记为$T=R\bowtie S$。

容易证明自然连接具有可结合性和可交换性，即：

$$(R_1\bowtie R_2)\bowtie R_3=R_1\bowtie(R_2\bowtie R_3)$$

$$R_1\bowtie R_2=R_2\bowtie R_1$$

每个关系都具有一个不随时间变化的结构，称之为关系模式。一个关系模式由关系的属性集组成，并且应保持数据依赖。一个数据库模式就是关系模式的集合。如果$S=\{X_1, X_2, \cdots, X_n\}$是一个数据库模式，那么我们把$\bigcup_{i=1}^{n} X_i$记为$attr(S)$，把$\bowtie_{i=1}^{n}\pi_{x_i}S$记为$\bowtie[S]$。为了方便起见，我们把$\bowtie[\{X_1, X_2, \cdots, X_n\}]$，写作$\bowtie[X_1, X_2, \cdots, X_n]$，把$\bowtie[S\cup\{X, Y\}]$写做$\bowtie[S, X, Y]$。

在数据库术语中，关系这个词有时是含混不清的，因为它既可用来表示元组的集合，又可用来表示元组集合的结构描述。把两者之间的区别搞清楚是必要的：当我们说元组的集合时将用关系这个词，而在说明元组集合的结构描述时，我们就用关系模式这个词。按照这种说法，关系就是关系模式的一个实例。在以后讨论中，为方便起见，在不发生混淆的情况下，我们有时也把关系模式简称为关系。

四、自然连接和投影的性质

根据自然连接和投影的定义，有如下定理：

定理4.1[1][4] 关系$R\langle U, \Sigma\rangle$，$U=\{X_1, X_2, \cdots, X_n\}$，$r_i=\pi_{x_i}R$。

那么a) $R\subseteq\bowtie[U]$

b) 如果$S=\bowtie[U]$，则$\pi_{x_i}S=r_i$

由定理4.1，可以证得下述定理：

定理4.2 关系$R\langle U, \Sigma\rangle$，$U=\{A_1, A_2, \cdots, A_n\}$，满足无损分解$\{X_1, X_2, \cdots, X_m\}$的充要条件是：

$$\left(\bigwedge_{i=1}^{m}\pi_{x_i}\langle a_1, a_2, \cdots, a_n\rangle\in\pi_{x_i}R\right)\Rightarrow\langle a_1, a_2, \cdots, a_n\rangle\in R$$

证明：(1) 充分性 根据自然连接的定义和性质可知：

若 $\bigwedge_{i=1}^{m} \pi_{x_i}(a_1, a_2, \cdots, a_n) \in \pi_{x_i} R$. 则 $\langle a_1, a_2, \cdots, a_n\rangle$必属于 $\bowtie_{i=1}^{m} \pi_{x_i} R$, 又因为 $\bowtie_{i=1}^{m} \pi_{x_i} R = R$,所以$\langle a_1, a_2, \cdots, a_n\rangle \in R$。

所以 $(\bowtie_{i=1}^{m} \pi_{x_i} R = R) \Rightarrow \{(\bigwedge_{i=1}^{m} \pi_{x_i}\langle a_1, a_2, \cdots, a_n\rangle \in \pi_{x_i} R) \Rightarrow \langle a_1, a_2, \cdots, a_n\rangle \in R\}$

(2) 必要性

$(\bigwedge_{i=1}^{m} (\pi_{x_i}\langle a_1, a_2, \cdots, a_n\rangle \in \pi_{x_i} R) \Rightarrow \langle a_1, a_2, \cdots, a_n\rangle \in R) \Rightarrow (\bowtie_{i=1}^{m} \pi_{x_i} R = R)$

采用反证法：假设左侧成立，右侧不成立。根据定理4.1(a),知 $\bowtie_{i=1}^{m} \pi_{x_i} R = R$, 又根据假设知，$\bowtie_{i=1}^{m} \pi_{x_i} R \supset R$，那么必有一个元组设为$\langle b_1, b_2, \cdots, b_n\rangle \in \bowtie_{i=1}^{m} \pi_{x_i} R$ 而不属于R，考察$\langle b_1, b_2, \cdots, b_n\rangle$在$x_i$(i=1, 2, …, m)上的各个投影，则必存在一个 j($1 \leqslant j \leqslant m$)，使得 $\pi_{x_j}\langle b_1, b_2, \cdots, b_n\rangle \notin \pi_{x_j} R$，否则根据假设：

$(\bigwedge_{i=1}^{m} \pi_{x_i}\langle b_1, b_2, \cdots, b_n\rangle \in \pi_{x_i} R) \Rightarrow \langle b_1, b_2, \cdots, b_n\rangle \in R$，这就产生了矛盾。

因为 $\pi_{x_j}\langle b_1, b_2, \cdots, b_n\rangle \notin \pi_{x_j} R$，所以 $\pi_{x_j}(\bowtie_{i=1}^{m} \pi_{x_i} R) \neq \pi_{x_j} R$。但由定理4.1(b)知，$\pi_{x_j}(\bowtie_{i=1}^{m} \pi_{x_i} R) = \pi_{x_j} R$，矛盾。所以若 $(\bigwedge_{i=1}^{m} (\pi_{x_i}\langle a_1, a_2, \cdots, a_n\rangle \in \pi_{x_i} R) \Rightarrow \langle a_1, a_2, \cdots, a_n\rangle \in R) \Rightarrow (\bowtie_{i=1}^{m} \pi_{x_i} R = R)$ □

这个定理的重要性在于，它实际上反映了属性间的约束条件和无损分解之间的关系，以后我们就可以从关系的约束条件中，直接找出它的无损分解来，同样也可以从它的无损分解中，得出关系的约束条件。更重要的是，就投影和自然连接来说，由于JD是所有可能的依赖的最一般形式，所以我们可以根据定理4.2将各种形式的依赖条件(由语义定义的FD 除外)化成形如 $\bigwedge_{i=1}^{m} \pi_{x_i}\langle a_1, a_2, \cdots, a_n\rangle \in \pi_{x_i} R \Rightarrow \langle a_1, a_2, \cdots, a_n\rangle \in R$的约束条件。

对于关系R(U, Σ)，我们还可以导出如下定理[6]：

定理4.3 如果$\bowtie[S] = \pi_T R[T = attr(S))$，则记$\bowtie[S] = *[S]$，那么有：

(1) 如果$Y \subseteq attr(S)$，则$*[S] \Rightarrow *[S, Y]$

(2) $*[S, Y, Z] \Rightarrow *[S, YZ]$

(3) 如果$attr(X) = Y$，则$*[S, Y] \wedge *[X] \Rightarrow *[S, X]$

(4) 如果$Z \notin attr(S)$，则$*[S, YZ] \Rightarrow *[S, Y]$

注意，我们对这些定理的看法和文献[4]中的看法并不完全相同，我们认为定理4.3 (1)～(4) 不是连接依赖的，而是关于无损分解的定理，其理由下面将要阐述。

五、连接依赖的一个更加严密的定义

在第二节，我们指出，由于原有的连接依赖的定义不严密，会得到一些不正确的结论。为了消除这些矛盾，下面给出两种等价的连接依赖的更严格定义：

定义5.1 关系$R\langle U, \Sigma\rangle$满足$JD \bowtie[X_1, X_2, \cdots, X_n]$的充要条件是：$\bowtie[X] = R(X = \{X_1, X_2, \cdots, X_n\})$，且对于任何$X' \subset X$，$\bowtie[X'] \neq R$。

定义5.2 如果关系$R\langle U, \Sigma\rangle$无论何时，R的元组$\mu_1, \mu_2, \cdots, \mu_n$都是对于结果t在$X_1 X_2, \cdots, X_n$上可连接的，并且，若除去任何一个元组$\mu_i(1 \leqslant i \leqslant n)$，$\mu_1, \cdots, \mu_{i-1}, \mu_{i+1}, \cdots, \mu_n$都是对于结果t在$X_1, \cdots, X_n$上不可连接，则称R 满足在 $\bigcup_{i=1}^{n} X_i$ 上的连接依赖 JD $\bowtie[X_1, X_2, \cdots, X_n]$。

通过定理4.2，我们可以清楚地看出定

义5.1和5.2实质上是完全等价的。

让我们来考察关系模式R〈U, Σ〉, U={A, B, C, D}，其约束条件是：如果元组〈a, b, —, —〉, 〈—, b, c, —〉, 〈—, —, c, d〉, 〈a, —, —, d〉属于R，则元组〈a, b, c, d〉也属于R，其中"—"代表只出现一次的符号。

根据定理4.2可知，R可无损分解为X={AB, BC, CD, AD}，但根据定理4.3(1)知，R也可无损分解为X′={AB, BC, CD, AD, AC}。

如果根据定义2.1，则R既满足JD⋈〔AB, BC, CD, AD〕，也满足JD⋈〔AB, BC, CD, AD, AC〕。但若根据本节定义5.1，则R只满足JD⋈〔AB, BC, CD, AD〕，而不满足JD⋈〔AB, BC, CD, AD, AC〕，因为对于X′来说，存在一个X′的子集X={AB, BC, CD, AD}，使得⋈〔X〕=R。

现在来分析一下这个关系。根据定理4.3(1)知：*〔AB, BC, CD, AD〕⇒*〔AB, BC, CD, AD, AC〕。又根据定理4.2知：无损分解X′={AB, BC, CD, AD, AC}的约束条件为：若〈a, b, —, —〉, 〈—, b, c, —〉, 〈—, —, c, d〉, 〈a, —, —, d〉, 〈a, c, —, —〉属于R，则〈a, b, c, d〉也属于R，但比较这两个约束条件，可以看出：〈a, —, c, —〉这个约束对于JD⋈〔AB, CD, BC, AD〕显然是多余的，对连接依赖JD⋈〔AB, BC, CD, AD〕也不起任何约束作用。

这样，我们可以看出，如果关系的连接依赖中，允许存在冗余的分解，则必然存在冗余的约束条件，而这个冗余的条件对关系的连接依赖来说是不起作用的，所以我们必须在连接依赖中除去冗余的分解，也就是除去冗余的约束条件，这样才真正体现出属性间的依赖关系。

现在回头看一看第二节中提出的问题，实际上这个矛盾是由于定义2.1中允许存在冗余的分解所造成的，因为根据定义2.1，会导出如下的事实：

若关系R〈U, Σ〉中存在一个JD⋈〔X_1, X_2, …, X_n〕，我们几乎总可以找出U的一个真子集U′，使之不含R的候选关键字（除非R中的任何原子属性都是R的候选关键字）。因为⋈〔X〕=R(X={X_1, X_2, …, X_n})，根据定理4.3(1)知，⋈〔X, U′〕=R，由定义2.1，关系R也满足JD⋈〔X_1, X_2, …, X_n, U′〕，而这个连接依赖不是由R的候选关键字所蕴含的。

这样，只有两种情况的关系属于5NF：

1.不含有任何连接依赖；

2.存在连接依赖，且每个原子属性都是关系的候选关键字。

但这样5NF的范围就太窄了，绝大多数的关系将不属于5NF。

如定义5.1那样，如果在连接依赖的定义中消除冗余的分解，则可以避免这个矛盾。根据定义5.1，关系S($S_\#$, SNAME, STATUS, CITY)不满足JD⋈〔($S_\#$, SNAME), ($S_\#$, STATUS), (STATUS, CITY), ($S_\#$, CITY)〕，因为对S来说，X={($S_\#$, SNAME), ($S_\#$, STATUS), (STATUS, CITY), ($S_\#$, CITY)}，存在一个子集X′={($S_\#$, SNAME), ($S_\#$, STATUS), ($S_\#$, CITY)}，使得⋈〔X′〕=S，实际上，也就是存在一个冗余的分解(STATUS, CITY)。可以验证，根据定义5.1，S中的任何连接依赖，都是由S的关键字所蕴含的，所以S∈5NF。

六、连接依赖的判断

如何判断一个已知关系R中是否存在连接依赖呢？根据自然连接的性质，我们有以下定理：

定理6.1 关系R〈U, Σ〉, U={A_1, A_2, …, A_n}，则R存在连接依赖的充分必要条件是：$\bowtie_{i=1}^{n} \pi_{X_i} R = R(X_i = U - A_i,\ i=1,2,\cdots,n)$。

证明：(1) 充分性　因为若 $\overset{n}{\underset{i=1}{\bowtie}}\pi_{X_i}R=R$，我们总可以在 $X=\{X_1, X_2, \cdots, X_n\}$ 中选取若干项 $X'=\{X_1', X_2', \cdots, X_m'\}$ $(m\leqslant n)$，使之 R 满足 $JD\bowtie[X_1', X_2', \cdots, X_m']$。

(2) 必要性　设 $Y=\wp U-\{\phi\}-\{\{A_1, A_2, \cdots, A_n\}\}$，这里 $\wp U$ 是 U 的幂集，$X=\{X_1, X_2, \cdots, X_n\}$。

若R中存在一个 $JD\bowtie[Z_1, Z_2, \cdots, Z_m]$，$Z=\{Z_1, Z_2, \cdots, Z_m\}$。考虑 $P=\bowtie[Y]$，则 $P=\bowtie[Y]=(\bowtie[X])\bowtie(\bowtie[W])$ $(W=Y-X)$

$\because \forall W'\exists X'(W'\in W\wedge X'\in X\wedge W'\subseteq X')$

∴根据自然连接的性质，逐项连接可知：

$$P=\overset{n}{\underset{i=1}{\bowtie}}\pi_{X_i}R$$

另外，$P=\bowtie[Y]=(\bowtie[Z])\bowtie(\bowtie[V])$ $(V=Y-V)$

$$=R\bowtie(\bowtie[V])$$

根据自然连接的性质，逐项连接可知：$P=R$

∴如果R存在一个连接依赖，则 $\overset{n}{\underset{i=1}{\bowtie}}\pi_{X_i}R=R$　□

我们把关系R中的 $\overset{n}{\underset{i=1}{\bowtie}}\pi_{X_i}R$ 称为R的基本判定式，记为 JP。根据定理 6.1，我们可以得到如下推论：

推论1　关系R(U, Σ)，若 $JP\neq R$，则 $R\in 5NF$。

推论2　关系R(U, Σ)，R是全码，若 $JP=R$，则 $R\notin 5NF$。

从推论1可知，关系R的基本判定式 $JP\neq R$ 是 $R\in 5NF$ 的充分条件，但不是 $R\in 5NF$ 的必要条件。

类似地，我们也可以推出关于嵌入型连接依赖（EJD）的判断方法：

推论3　关系R(U, Σ)，$JP\neq R$，若存在一个 $U'\subset U, R'=\pi_{U'}R$ 且 $JP'=R'$，则R中存在一个嵌入型连接依赖。

七、结束语

为了消除连接依赖的原有定义造成的一些矛盾，本文提出了一个更为严格的新定义，并对连接依赖的判断方法和5NF作了一些讨论。由于就投影和自然连接来说，连接依赖是所有依赖的最一般形式，所以，我们只有对连接依赖作更进一步的研究，才能真正弄清数据之间的依据关系。

参考文献

1) 萨师煊、王珊，数据库系统概论，第四版 高等教育出版社，1986

2) 刘方鑫、鲍剑洋，连接依赖的更加严格定义和判断，全国第八届数据库大会

3) C. Delobel，关系数据理论概述，计算机科学，No.2，1982.pp14～28

4) J.D. Ullman, Principles of Database System, 2nd Edit Pitman Publish ltd, 1982.

5) E. Sciore, A Complete Axiomatization of Full Join Dependencies, J. ACM, Vol. 29, No.2, April 1982 PP373～393

6) Data, C.J., A Introdution to Database System, Vol.1, 4th Edit, ADDISON-Wensley Publish Company, 1986.

7) David Maier, On the Complexity of Testing Implications of Function and Join Dependecies, J. ACM, Vol.28, No.4, October 1981, PP680～695