条件函数依赖与数据质量控制

程录庆

（南京人口管理学院 信息科学系， 南京 210042）

摘要：提高数据质量对于企业管理信息系统意义重大，数据质量的提高涉及数据库的设计、数据的输入、数据的维护及数据清洗全过程。比较了条件函数依赖 CFD 与传统函数依赖 FD；基于条件函数依赖框架对业务规则及领域知识的表达作了初步研究，对脏数据的识别和清洗提供可行的方法和思路。

关键字：数据质量； 条件函数依赖； 业务规则； 数据清洗

1、引言

在关系数据库中，数据依赖是最重要的语义约束表达，决定了关系数据的模式设计。基本的数据依赖形式有函数依赖、多值依赖、连接依赖等。在传统的关系数据库理论研究中，我们是根据数据依赖来设计一个结构良好的关系数据库模式，基本的方式是采用“一事一地”的原则将现实数据分散存储于若干个关系表中。一个设计良好的关系数据结构提供了有效保证数据完整性的基本框架，然而，这还不够，不符合业务规则的数据依然会产生。

如表 1 所示，根据实际语义规则，不存在传统的函数依赖 FD（注：这里记录号仅用于标识记录，不作为表中的字段），所以表 1 符合 BC 范式，是设计良好的关系表。然而，实际关于课程的业务规则可能会有：(1)rule1：任何教师只能上一门研究生课程；（2）rule2：每一门实践类课程的授课地点只能是一个。我们可以看到：表中的数据不违反任何的 FD，但记录 r4 和 r5 不符合 rule1，r1 和 r6 不符合 rule2。表 1 中的数据依然是不符合实际业务规则的，是不完整的，存在“脏”数据。

表 1：课程（C）

记录号	课程名	课程性质	授课对象	授课地点	教师
r1	数据库系统	实践	本科	Lab2	赵明
r2	电子学	课堂教学	专科	2-502	李达
r3	电子学	课堂教学	研究生	1-403	陆海
r4	电子学	课堂教学	研究生	1-402	李达
r5	数字电路	课堂教学	研究生	1-506	李达
r6	数据库系统	实践	函授	Lab5	陈新

这也就表明，传统的数据依赖还不能充分表达应用领域的业务规则，事实上，rule1 表达的也是一个函数依赖{教师→课程名}，只不过这个函数依赖需要在一定的条件[授课对象=“研究生”]下才能成立。为此，本文引入，并探讨条件函数依赖与领域知识的表达及数据质量控制的关系。

2、条件函数依赖

在此，引入文献[1][2]的条件函数依赖（CFD，conditional functional dependancy）。

定义 1 一个 CFD ϕ 可以表达为：(R: $X \rightarrow Y$, Tp)，其中（1）R 是关系模式；（2）以 attr(R)表示 R 的属性，X、Y 是 attr(R)子集；（3）$X \rightarrow Y$ 是 R 上传统的函数依赖；（4）Tp 是型表（pattern tableau），属性由 $X \cup Y$ 组成，Tp 的元组（tuple）由常量或未命名的变量“_”构成。

例子：表 1 课程的两个业务规则 rule1 和 rule2 可以表示成 CFD，分别为 $\phi 1$， $\phi 2$

$\phi 1$：课程([授课对象=“研究生”，教师]→课程名)，或

（课程(授课对象，教师)→课程名，Tp1）.

ϕ2：课程([课程性质=“实践”，课程名]→授课地点)，或
（课程(课程性质，课程名)→授课地点，Tp2)

Tp1		
授课对象	教师	课程名
研究生	_	_
函授	李新	英语

Tp2		
课程性质	课程名	授课地点
实践	_	_

如果还存在这样的约束：李新给函授生只上英语课。则可以将引约束表示成一行 Tp 元组{函授，李新，英语}插入到 Tp1 中去。

传统的函数依赖可表示成 CFD 的一个特例，假如存在：{课程名，授课对象→授课地点 }，则可表示成（课程(课程名，授课对象)→授课地点，Tp3)，Tp3 仅由一行未命名的变量“_”构成。

Tp3		
课程名	授课对象	授课地点
_	_	_

使用符号“≈”表示匹配（match)，“≈”规则如下：

a≈b，当（1）a or b 取值为“_”；（2）a、b 均为常量，且 a=b。两个元组 t1、t2 的匹配可记为：t1≈t2。例如：(a,b)≈(a,_)，但和(a,c)不匹配。

定义 2 一个关系的实例 r 满足一个 CFD（R：$X \rightarrow Y$, Tp)：设 t1、t2 是 r 中的任意元组，tp 为型表 Tp 中的任意元组，如果 t1[X]=t2[x]≈tp[X]，则 t1[Y]=t2[Y]≈tp[Y]。

一个关系的实例 r 满足一个 CFDϕ记为 r ⊨ϕ。设Σ为ϕ的集合，若 r 满足Σ中的每一个ϕ_i，则 r ⊨Σ。显然，表 1 中的元组 r4、r5 和 r6 违背ϕ1，r1 和 r6 违背ϕ2。

与传统的函数依赖不同，单个的元组也有可能违背条件函数依赖约束，如表 1 中的 r6 不满足ϕ1。

3、条件函数依赖与数据质量控制

在数据库技术日益成熟的背景下，人们越来越多地关注数据质量的提高，对于一个企业管理信息系统，数据质量的低下可能滞约企业的发展。错误、冗余、不一致的数据将导致市场动作的低效率，影响企业的客户关系，误导企业的战略决策。据估计，“脏数据”每年会给美国带来 600 亿美元的经济损失[3]，有理由相信，这种的情形在中国也同样存在。

脏数据是指错误的、冗余的、不一致的数据，一般是由于违背了数据库本应遵从的完整性约束而产生，而完整性约束意味着准确、一致的数据所应遵循的规则，是由具体的商业规则得来的，商业规则限定了关系数据的属性值应符合其所反映现实的上下文。例如，一个企业的客户关系数据库可能有这样的规则：（1）一个新的客户在第一次购买时享有 15%的折扣；而一个 VIP 客户在任何时候购买任何产品享有 25%的折扣；（2）一个地址为美国的客户，“街道”、“城市”、“州”字段应函数确定“邮政编码”。提高数据质量的首要任务是找出一套正确反映企业政策和业务规则的完整性约束[4]。

在数据库中，识别脏数据是进行数据清洗（提高数据质量的一个过程）第一步。数据的错误或者不一致有些是很容易识别的，如：课程性质为“实践”的课程，对应的授课地点若为“1-403”，则是明显不正确的。而有些潜在的数据不一致就不那么容易发现了，如“李达”老师给研究生上了二门课（表 1 中的“电子学”和“数字电路”)，这时，就要求根据业务规则和领域知识来发现潜在的脏数据。

条件函数依赖（CFD）事实上就是一种新的表达数据约束的方法。首先，传统的 FD 对数据的约束是全局的，而 CFD 是表达一定条件下数据的约束，是局部的，可以说，在这一点上，CFD 比 FD 对数据约束的表达要精细得多，这也使得 CFD 更适应新的数据库技术

要求，如数据集成。另外，与传统的数据依赖主要用于数据库的结构设计不同，CFD 则可广泛用于数据的输入控制、数据的维护及数据的自动清洗。数据质量的提高涉及数据库的设计、数据的输入、数据的维护及数据清洗全过程，为提高数据的质量，数据库的拥有者会采用大量的人力通过手工的方式来控制数据的输入、脏数据的识别以及数据的纠错，而 CFD 这种新的数据约束表达方式的出现提供了利用计算机程序自动完成上述工作的可行方法。Wenfei fan 等人对基于条件函数依赖进行数据的自动清洗作了理论和实践上的深入研究[1][2]。

值得注意的是，条件函数依赖本身也存在不一致的问题，即 CFD 本身是“脏”的。CFD 表达的是数据库应遵循的一套完整性约束，其本源还是反映数据库拥有者所制定的业务规则和领域知识，也就是说，CFD 是不一致的，则必然是业务规则存在相互矛盾的地方。这也应验了一个道理：“输入是垃圾，输出的也是垃圾。”，数据是管理活动（或其他业务过程）产生的，糟糕的业务管理会导致数据质量的低下，反过来，提高数据质量还存在一个可行的途径是提高管理的质量。

4、结语

数据的质量与数据约束（即数据应遵循的规则）有着密切的关系，传统的函数依赖（还有其他数据依赖形式）主要服务于数据库的结构设计，也是数据完整性约束的表达方式。条件函数依赖是数据约束表达方式的重大扩展，适应新的数据库技术要求，可广泛服务于数据的输入控制、数据的维护及数据的自动清洗。很多数据库研究者接受这样一个事实，即数据质量的好坏很大程度上取决于领域知识和实际的业务规则[5]。本文基于条件函数依赖对数据库反映的实际业务规则的表达作了初步的研究，探讨了提高数据质量与条件函数依赖的关系，进一步的研究将包括探讨基于条件函数依赖框架下新的数据依赖形式，如条件包含依赖（CIND），以及对业务规则和领域知识的表达作深入研究。

参考文献：

[1]P.Bohannon,W.Fan,F.Geerts,X.Jia,and A.Kementsietsidis.“Conditional functional dependencies for data cleaning”[C]. In ICDE, 2007.

[2]WENFEI FAN，FLORIS GEERTS，XIBEI JIA，ANASTASIOS KEMENTSIETSIDIS. “Conditional Functional Dependencies for Capturing Data Inconsistencies”[R].ACM Transactions on Database Systems, Vol.V,No.N,April 2008, Pages 1-44

[3]Gao Cong,Wenfei Fan,Floris Geerts,Xibei Jia,Shuai Ma. “Improving Data Quality: Consistency and Accuracy”[C]. In VLDB,September 2328,2007.

[4]Fei Chiang, Ren´ee J. Miller.“Discovering Data Quality Rules”[C]. VLDB ‘08, August 2430,2008,

[5]Philip Bohannon,Eiman Elnahrawy. “Putting Context into Schema Matching”[C]. VLDB ‘06, September 12-15, 2006