УДК 004.652

О ШУНТИРОВАНИИ МНОГОЗНАЧНЫХ ЗАВИСИМОСТЕЙ В РЕЛЯЦИОННОЙ МОДЕЛИ ДАННЫХ

Б.Е. Панченко

Институт кибернетики имени В.М. Глушкова НАН Украины,
03680, ГСП, проспект Академика Глушкова, 40,
(044) 526 3603, pgt@ukr.net

Предложен новый подход к анализу многозначных зависимостей атрибутов реляционных отношений. Доказана теорема о шунтировании многозначной зависимости, позволяющая не декомпозировать отношение. Делается вывод о возможности моделирования N-арных связей кардинальностью «многие ко многим» с помощью шунтированных отношений.

New approach to analyze of multivalued dependencies of attributes of relational relations is offered in this work. The theorem about bypassing of multivalued dependency that allows to not decompositionse relation is proved. The conclusion about possibility of modeling N-links with cardinality “many-to-many” with help of by-passed relations is drawn.

Введение

Классическая логическая модель данных подразумевает такой анализ предметных областей, который приводит к постановочно-зависимым структурам хранилищ. Такие структуры существенно отличаются одна от другой в зависимости от специфики предметных областей. Это приводит к большой зависимости управляющего программного обеспечения от всевозможных модификаций структуры хранилищ. Поэтому разработка новой концептуальной модели, позволяющей получить универсальную логическую модель данных, является актуальной задачей.

И хотя классическая реляционная модель основана на очень жестком алгоритме, слабо поддающемся унификации, идея универсальности не нова. Ее проявление состоит в алгоритме синтеза схемы реляционной таблицы, предложенном Ф.А. Бернштейном [1, 2]. Эта же идея есть в выводах работы П.П. Чена [3], как и в самом ее названии. Та же мысль обсуждалась в работах Ю.А. Пергаменцева [4, 5]. К сожалению, метод, предложенный в [4, 5], дальнейшего развития пока не имеет. Однако основной посыл – унификация моделирования данных, по мнению многих исследователей БД, очень перспективен [6, 7]. Технической реализации универсальной логической модели данных посвящены и работы [8, 9] .

Проанализируем, от чего может зависеть решение этой проблемы. Здравый смысл подсказывает, что существование любого единого языка передачи информации, синтаксис которого не зависит от содержания сведений, доказывает то обстоятельство, что и при моделировании данных конструкции должны получаться эквивалентными вне зависимости от особенностей разнообразных предметных областей (ПО). Что же формирует эту особенность? Главным фактором декомпозиции в процедуре нормализации [10, 11, 12] являются функциональные зависимости (ФЗ). Как указано в [1], функциональные зависимости – это не что иное как связи. Выскажем гипотезу: именно связи порождают проблемы при попытке проектирования тождественных структур реляционных таблиц для всех видов ПО.

Рассмотрим последовательно классические процедуры нормализации и причину аномалий [6, 7, 10 – 13].

Вторая нормальная форма

Пример 1.

Пусть есть отношение $Поставки(Код\ Поставщика, Товар, Цена)$. Проектировщик, рассматривая данное отношение, отмечает следующие факты:

1) ключ отношения:

$$Код\ Поставщика + Товар; \tag{1}$$

2) имеется две ФЗ:

$$Код\ Поставщика + Товар \to Цена, \tag{2}$$

$$Товар \to Цена. \tag{3}$$

Известно, что наличие в отношении частичной зависимости неключевого атрибута *«Цена»* от части ключа *«Товар»* приведет к аномалиям включения, удаления и модификации. Устраняется декомпозицией по атрибуту *«Товар»*. Свободными от аномалий будут таблицы: $Поставки(Код\ Поставщика, Товар)$, $Цена\ Товара(Товар, Цена)$.

Причина проблем – наличие неполной ФЗ в отношении.

Третья нормальная форма

Пример 2.

Пусть есть отношение

Лечение (*Код Врача, Код пациента, Дата, Лечение, Лекарство, Побочный Эффект*):

1) ключ отношения: $Код\ Врача + Код\ Пациента + Дата$;
2) имеется две функциональные зависимости:

$$Код\ Врача + Код\ Пациента + Дата \to Лекарство, \quad (4)$$

$$Лекарство \to Побочный\ эффект. \quad (5)$$

Известно, что наличие в отношении транзитивной зависимости типа $X \to Y,\ Y \to Z$ (но при этом $Y - I -> X$) приводит к аномалиям:

1) нельзя включить в приведенное отношение сведений о лекарстве, не назначенном ни одному больному;
2) при изменении сведений о побочном эффекте лекарства требуется пересмотреть все отношения для корректного изменения данных.

Алгоритм нормализации тот же. Устранив транзитивные ФЗ и ФЗ атрибутов от части ключа, Кодд получил 3НФ [11, 12]. Причина проблемы – транзитивность ФЗ.

Нормальная форма Бойса – Кодда

Пример 3.

Пусть имеется несколько иное отношение

$$Поставка\ (Код\ Поставщика, Имя\ Поставщика, Код\ Товара, Объем\ покупки). \quad (6)$$

При условии, что $Имя\ Поставщика$ – уникальный атрибут, имеем два ключа:

1) ключи отношения:

$$Код\ Поставщика + Код\ Товара, \quad (7)$$

$$Имя\ Поставщика + Код\ Товара. \quad (8)$$

2) имеется избыток ФЗ:

$$Код\ Поставщика \to Имя\ Поставщика, \quad (9)$$

$$Имя\ Поставщика \to Код\ Поставщика. \quad (10)$$

Наличие этих ФЗ означает, что в отношении имеются два детерминанта. Но каждый из детерминантов не является полностью ключом. Аномалии коснутся изменений имени поставщика. Нормализация – проекция по атрибуту «Имя Поставщика».

Причина проблемы – зависимость части ключа от неключевых детерминантов. Несмотря на то, что детерминантами являются части ключей, они сами по себе ключами не являются. То есть, причина все та же – «проблемные» ФЗ [13].

Таким образом, в каждом из приведенных случаев устранение нежелательных ФЗ приводит к нормализации отношений и к большей их схожести. Но, тем не менее, исходной таблицей для последовательной нормализации является отношение в 1НФ [10, 11]. Как известно, такая таблица получается неким бессистемным формированием совокупности атрибутов. Поэтому, несмотря на универсальность подхода нормализации [11,12], результирующие нормализованные отношения не гарантируют универсальности схемы хранилища данных для произвольных ПО.

Постановка задачи

Многие утверждения в сфере проектирования хранилищ данных сложно доказывать с использованием формального математического аппарата в рамках классической концептуальной модели данных. Поэтому для решения вопроса построения универсальной логической модели данных применим инженерный подход Чена [3]. Если утверждения укладывается в логику реляционных схем, а также, если в результате цепочки этих утверждений получаются адекватные нормализованные отношения, которые в свою очередь будут подобны или даже полностью идентичны для различных ПО, утверждения приобретут характер технических рекомендаций. Выскажем некоторые из них. Большинство приведем без доказательств.

Гипотеза 1. Причиной отличий схем реляционных отношений для различных предметных областей являются функциональные зависимости.

Исходя из этого утверждения построим некоторую новую концептуальную модель данных. Для нее определим основные категории.

Сущностями будем называть абстрактные множества, каждый элемент которых имеет общий набор параметров, которые соответствуют общему предикату. Этот предикат и объединяет все элементы в одно множество. Данное определение упрощает понимание отличий сущностей от еще одного важного фактора концептуальной модели, схожего с сущностью – от атрибута.

Атрибутом будем называть характеристику сущности, один из тех ее параметров, набор которых соответствует единому предикату. Дополнительным признаком атрибута является тот факт, что атрибут, в отличие от сущности, не имеет никаких атрибутов.

Связь – функциональная или многозначная зависимость между сущностями, а также функциональная зависимость атрибутов от своих сущностей.

Произвольная предметная область – произвольная совокупность сущностей, каждая из которых имеет произвольное количество атрибутов.

Специфика произвольной предметной области – совокупность связей между сущностями произвольной кардинальности – от «один к одному» до «многие ко многим». При этом предполагается, что все многообразие атрибутов сущностей, их тип, размерность и иные характеристики не влияют на специфику произвольной ПО.

Суррогатный атрибут – искусственно внесенный в отношение атрибут (например, ключевой), имеющий гарантированные свойства. Если суррогатный атрибут вносится в качестве ключа, то обеспечивается его уникальность для любого экземпляра сущности, а также минимальная достаточность размера значения, пропорционального проектному объему экземпляров отношения.

Простейшей будем называть схему реляционного отношения $R(X,\ A)$, когда $X \to A$, X – простой унарный ключ отношения, A – простой неключевой атрибут.

Особой будем называть схему реляционного отношения $R(X_j,\ A_i)$, когда $X_j \to A_i$, $\{X_j,\ j=1,\ J\}$ – в общем случае составной ключ отношения, форма – не ниже 4НФ [13], причем так, что ни один атрибут сам по себе не является ни ключом, ни детерминантом, $\{A_i,\ i=1,\ I\}$ – множество неключевых атрибутов, ни один из которых не является детерминантом и все атрибуты отношения – и ключевые, и не ключевые – входят в множество общего предиката, определяющего $R(X_j,\ A_i)$.

Подобными будем называть схемы реляционных отношений, в которых структуры ключей и функциональных зависимостей аналогичны. При этом тип, количество и размерность неключевых атрибутов на подобие схем не влияют.

Коэффициентом подобия схемы отношений будем называть целую часть от деления большей арностей ключей к меньшей. При этом, если в отношениях сущствует несколько ключей, то сравнению подлежит старшая арность. Для сравнения с простейшим отношением смысл коэффициента подобия – это превышение унарности. То есть коэффициент подобия k схемы с тернарным ключом в отношении $R(X,\ Y,\ Z,\ A_i)$, где $X+Y+Z$ – тернарный ключ, A_i - множество неключевых атрибутов, и простейшей схемы, будет равен 3.

В дальнейшем знаком «+» будем обозначать объединение экземпляров столбцов множества атрибутов (суммирование в строке). В литературе [6, 7, 11, 14, 15, 16] это действие часто обозначают постановкой атрибутов радом: *XY* или запятой между ними: *X, Y*. Но для конкретизации действия будем пользоваться символом суммирования. Иными словами, подобными будут схемы, у которых одинаковые структуры функциональных зависимостей и одинаковые структуры ключей. Отличия могут быть лишь в арности ключей. Соотношение арности и будем называть коэффициентом подобия.

Внешний атрибут (след внешней сущности) – атрибут (возможно, ключевой или даже суррогатный), предикат которого отличается от предиката отношения (т. е., предиката большинства атрибутов этого отношения). Используется в отношении среди атрибутов различных сущностей в качестве внешнего ключа связи.

Внешняя связь (внешняя ФЗ или *МЗ)* – связь, предикат атрибутов которой отличается от предиката большинства атрибутов отношения, находящегося в форме, не ниже 4НФ [14]. К этой категории относятся и ненормализованные функциональные или многозначные зависимости, которые присутствуют в реляционном отношении, находящемся в одной из форм ниже НФБК [13]. Очевидно, что, например, функциональная зависимость совокупности неключевых атрибутов от составного ключа отношения не является внешней связью.

Подобие реляционных схем

Сформулируем утверждение, вытекающее из приведенной гипотезы 1.

Теорема 1. Все реляционные отношения, имеющие особую схему, подобны друг другу, подобны простейшей схеме и не зависят от специфики предметной области.

Доказательство. Рассмотрим группу реляционных отношений со схемами

$$R_0(X_0,\ A_i), \tag{11}$$

$$R_1(X_1,\ A_i), \tag{12}$$

$$R_2(X_1+X_2,\ A_i), \tag{13}$$

$$R_3(X_1+X_2+X_3,\ A_i), \tag{14}$$

$$R_j(X_j,\ A_i), \tag{15}$$

причем, каждое из них находится в 4НФ [14] и атрибуты X_j являются различными частями ключей отношений (кроме, конечно $X_{0,1}$ в $R_{0,1}$, которые является полностью ключом). Атрибут A и множество A_i – различные неключевые атрибуты, где $j=1,J;\ \ i=1,I_j$, причем j – номер отношения и количество ключевых атрибутов (частей ключа) в каждом j-м отношении, а I_j – количество неключевых атрибутов каждого j-го отношения.

1. Проверим подобие. Необходимо показать, что из (11) следует (13) и формально $R_2 \sim 2\times R_0$ и $R_3 \sim 3\times R_0$, где символом $n\times$ будем обозначать коэффициент подобия, т. е. арность.

То, что $R_1 \sim R_0$ с коэффициентом подобия $k = 1$, следует из определения подобия. Ключевые атрибуты X и X_1 унарны, а структура неключевых атрибутов не влияет на схему отношения.

Далее, поскольку реляционные отношения находятся в НФБК, то каждый из X_i сам по себе не может быть детерминантом – это противоречит условию НФБК [13]. То есть, каждый из этих атрибутов может быть лишь частью суммарного ключа. Но и среди A_i также нет детерминантов. Это значит, что в R_2 и R_3 имеет место единственная ФЗ $X_j \to A_i$, т. е., в частности, для R_3 зависимость имеет вид $(X_1 + X_2 + X_3) \to A_i$.

Значит, при замене суммарного атрибута $X_1 + X_2 + X_3$ на атрибут W той же размерности и с просуммированным набором значений, составляющие схемы отношения не изменятся. То есть выполняется условия подобия. Коэффициент подобия определится отношением наибольшей арности ключей. Поскольку в исходном и анализируемом отношениях есть только один ключ, отношение арности составляет 2 и 3 для R_2 и R_3 соответственно.

Вывод о подобии важен для сравнения составляющих отношения в особой форме и отношений с МЗ.

2. Независимость особой формы от специфики произвольной ПО, вообще говоря, также очевидна из определения, как и подобие. Действительно, под спецификой произвольной ПО понимаем совокупность связей между сущностями. Как известно, признаком связи между сущностями является наличие в отношении следов сущностей – их атрибутов. Поскольку все множество атрибутов искомого отношения соответствует единому предикату, в нем нет внешних атрибутов. А это значит, что данное отношение свободно от следов внешних сущностей. Поскольку начальные ограничения на искомое отношение устанавливались в общем виде, этот вывод касается любой ПО.

Теорема о шунтировании многозначной зависимости

Для выяснения вопроса, какие типы отношения моделируют специфику произвольной ПО, сформулируем и докажем теорему. Она является базовой для дальнейшего построения универсальной логической модели данных.

Теорема 2. Если в отношении $R(X,\ Y,\ Z)$ имеется нетривиальная МЗ $X \to\to Y \setminus Z$, причем $X,\ Y,\ Z$ могут быть составными множествами, т. е. $X = \{X\},\ Y = \{Y\},\ Z = \{Z\}$, то при добавлении в это отношение дополнительных неключевых атрибутов $\{A\}$, причем так, что $(X + Y + Z) \to A$, зависимость в отношении $R(X,\ Y,\ Z,\ A)$ перестаёт быть многозначной.

Доказательство. Предположим обратное, что в отношении $R(W,\ Y,\ Z)$, где $W = \{X + A\}$, имеется МЗ $W \to\to Y/Z$. Пусть R – реляционная схема, W и Y – не пересекающиеся подмножества R. И пусть $Z = R - (W + Y)$. Следуя [15, 16], отношение R удовлетворяет МЗ $W \to\to Y$, если для любых двух кортежей t_1 и t_2 из R, для которых $t_1(W) = t_2(W)$, в R существуют еще и кортеж t_3, для которого выполняется соотношение $t_3(W) = t_1(W),\ t_3(Y) = t_1(Y),\ t_3(Z) = t_1(Z)$.

То есть, поскольку в отношении $R(W,\ Y,\ Z)$ имеются кортежи

$$r_1 = (w,\ y,\ z_1), \tag{16}$$

$$r_2 = (w,\ y_1,\ z), \tag{17}$$

то из МЗ должны следовать и кортежи

$$r_3 = (w,\ y,\ z), \tag{18}$$

$$r_4 = (w,\ y_1,\ z_1). \tag{19}$$

Но эквивалентность кортежей в МЗ нарушается, так как

$$r_1 = (x + a_1,\ y,\ z_1), \tag{20}$$

$$r_2 = (x + a_2\ y_1,\ z), \tag{21}$$

$$r_3 = (x + a_3,\ y,\ z), \tag{22}$$

$$r_4 = (x + a_4, y_1, z_1). \tag{23}$$

Значит, в полном отношении отсутствует МЗ. Теорема доказана.

Из доказанного следует, что если отношение имеет МЗ, ее можно "шунтировать", т. е., перекрыть ее влияние добавлением нового элемента – естественного или суррогатного атрибута, "физический" смысл которого – характеристика связи. Поэтому в дальнейшем будем называть доказанное выше утверждение теоремой о шунтировании МЗ, а таблицы с «перекрытой» МЗ – шунтированными таблицами.

В качестве примера рассмотрим ПО из [16]. Пусть «абитуриенты» сдают «вступительные экзамены» по списку «предметов» в «институты». При этом дополнительным условием является разрешение любому абитуриенту подавать документы в любой институт и выбирать при этом день сдачи экзамена. Данное упрощение мы вводим лишь для того, чтобы не загромождать пример дополнительными ограничениями, которые лишь отвлекут от сути. Заметим, однако, что данный пример не упрощает ситуацию, а наоборот усложняет её. В реальности кардинальность связи «многие ко многим» по произвольному числу сущностей встречается крайне редко. А это значит, что появляются дополнительные ФЗ в ключевых атрибутах, которые лишь упрощают шунтирование МЗ.

X – «Абитуриенты» – $X = \{1, 2, 3\}$.

Y – «Предметы» – $Y = \{1, 2, 3\}$.

Z – «Институты» – $Z = \{1, 2, 3\}$.

$A_{1,2}$ – оценка на экзамене, дата или время его проведения и т. п.

В таблице приведено полное отношение исходя из максимально возможного объема кортежей, полученное декартовым произведением атрибутов [8, 9, 17], чтобы наглядно показать основной вывод теоремы. Очевидно, на практике многие кортежи вообще не появились бы. Условно «абитуриент 2» не явился бы на экзамен «2+2+3» и не получил бы двойку, зная, что он не готов к «предмету 3».

Заметим, что никакая часть A_i не зависит ни от какой комбинации частей ключа, кроме как от самого ключа, сам ключ – так же. Получена форма, *подобная* БКНФ [13], но без МЗ. Как известно из [14], это есть 4НФ.

Таблица. Отношение с шунтированной МЗ

X	Y	Z	A_1	A_2
1	1	1	3	12.10
1	1	2	5	16.30
1	1	3	2	10.30
1	2	1	3	11.00
1	2	2	5	15.00
1	2	3	4	12.10
1	3	1	2	10.30
1	3	2	3	11.00
1	3	3	3	15.00
2	1	1	5	14.10
2	1	2	3	12.30
2	1	3	5	11.45
2	2	1	5	12.10
2	2	2	4	10.30
2	2	3	2	16.30
2	3	1	3	11.00
2	3	2	5	15.00
2	3	3	2	13.00
3	1	1	5	12.10
3	1	2	3	11.00
3	1	3	4	11.45
3	2	1	2	10.15
3	2	2	3	15.00
3	2	3	2	11.00
3	3	1	5	10.30
3	3	2	2	12.10
3	3	3	5	16.30

Выводы

Таким образом, получен алгоритм построения 4НФ не с помощью декомпозиции [14], а путем синтеза. Причем повторяемость значений частей ключевого атрибута, которая остается от многозначной зависимости, не является аномалией. Дело в том, что такие отношения не самостоятельны, а входят в совокупность реляционных таблиц. В таком множестве таблицы с шунтированной МЗ и составным ключом, каждая часть которого соответствуют своему уникальному предикату, моделируют связи произвольной арности между соответствующими сущностями. А каждое отношение с единым предикатом моделирует хранилище характеристик соответствующей сущности. Поэтому процедура каскадного обновления, вставки или удаления, редактирующая ключевые атрибуты в таблицах сущностей, может учесть и взаимосвязи этих ключевых атрибутов со своими «потомками» в таблицах связей. Тем самым можно отследить целостность данных в этих подчиненных таблицах.

Совокупность таблиц-сущностей и таблицы связи между этими сущностями в общем случае моделирует связь из реальной ПО кардинальностью «многие ко многим». Каждому кортежу из таблицы сущности соответствует конечное подмножество кортежей из таблицы связи. С этой позиции декомпозиция многозначной зависимости приводит к потере этих связей. Видимо, именно это и является причиной того, что большинство проектировщиков, как утверждается во всех учебниках по БД [6, 7, 16], игнорируют рекомендации Фейджина [14] и оставляют МЗ в БКНФ-отношениях.

1. *Bernstein P., Swenson J., Thichritzis D. A.* Unified Approach to Functional Dependencies and Relations // Proc. International Conference on the Management of Data. – ACM SIGMOD, 1975. – P. 237–245.
2. *Bernstein P.A.* Synthesizing third normal form relation from functional dependencies // ACM Transactions on Database Systems. – 1976. – 1. – № 4. – P. 277–298.
3. *Chen P.P.* The Entity-Relationship Model: toward a unified view of data // ACM Trans. on Data base systems. – 1976. – 1. – 1. – P. 9 – 36.
4. *Пергаменцев Ю.А.* Проектирование БД на основе универсальной модели данных // Материалы седьмой технической конференции «Корпоративные базы данных'2002», 16–17 апреля 2002 года, М.: 2002.
   http://www.citforum.ru/seminars/cbd2002/
5. *Есин В.И., Пергаменцев Ю.А.* Технология проектирования модели предприятия на основе универсальной модели данных. – М.: CIT-Forum, 2008.
   http://www.citforum.ru/database/articles/udm
6. *Ульман Д.Д., Уидом Д.* Основы реляционных баз данных. – М.: 2006 – 374 с.
7. *Кузнецов С.Д.* Базы данных: модели и языки. Учебник. – М.: 2008. – 720 с.
8. *Панченко Б.Е.* Способ расположения данных в компьютерном хранилище, обеспечивающий модифицируемость его структуры // Патент Украины № 63036, 2001 г.
9. *Панченко Б.Е., Писанко И.Н.* Свойства реляционного каркаса на множестве семантически атомарных предикатов// Кибернетика и системный анализ. – Киев. – 2009. – № 6. – С. 120–129.
10. *Codd E.F.* A Relational Model of Data for Large Shared Data Banks // Comm. ACM, 13, 6 (jun), 1970. – P. 377 – 387.
11. *Codd E.F.* Normalised Data Base Struchture: a Brief Tutorial // Proc. ACM, SIGFIDET, Workshop, San Diego, Calif., Nov. 1971. – P.1 – 18.
12. *Codd E.F.* Further Normalization of the Data base Relational Model // Data Base Systems.– N.J.: Prentice–Hall, 1972. – P. 33–64.
13. *Codd E.F.* Recent investigations in relational database systems // In Proc. IFIP Congress – 74, North–Holland, Amsterdam, 1974. – P. 1017–1021.
14. *Fagin R.* Multivalued dependencies and a new normal form for relational databases // ACM Transactions on Database Systems, 1977. – 2. – N 3. – P. 262–278.
15. *Мейер Д.* Теория реляционных баз данных. – М.: 1987. – 608 с.
16. *Пасичник В.В., Шаховская Н.Б.* Хранилища данных. Учебное пособие. – Львов: 2006. –492 с.
17. *Курош А.Г.* Общая алгебра. – М: 1979. – 150 с.