УДК 004.652

О СИНТЕЗЕ УНИВЕРСАЛЬНОЙ ЛОГИЧЕСКОЙ МОДЕЛИ ДАННЫХ

Б.Е. Панченко, *канд. физ.-мат. наук, доцент*
Сумский государственный университет, г. Сумы

В работе предложен новый подход к построению универсальной логической модели данных на основании многозначных зависимостей ключевых атрибутов. Предлагается использовать N-арные таблицы для моделирования связей кардинальности «многие ко многим». Решается проблема модифицируемости структуры реляционного хранилища данных.

У роботі запропонований новий підхід до побудови універсальної логічної моделі даних на підставі багатозначних залежностей ключових атрибутів. Пропонується використовувати N-арні таблиці для моделювання зв'язків кардинальности «багато хто до багато кому». Вирішується проблема модифікованості структури реляційного сховища даних.

ВВЕДЕНИЕ

Одним из подходов к преобразованию концептуальной модели в логическую является замена некоторых не подходящих для промышленных реляционных СУБД структур данных с целью их максимального упрощения. Но такие «упрощения» делают итоговую логическую модель жестко зависимой от предметной области, а программную среду, манипулирующую данными, зависимой от структур данных. В [1] приведен список основних «трудных» структур данных, которые принято упрощать: связи типа «многие ко многим», n-арные и рекурсивные связи сущностей, атрибуты связей, иерархическая зависимость («слабость») сущностей, а также множественные атрибуты. В работе [2] обсуждалась еще одна немаловажная проблема, которая касается не столько проектирования, сколько эксплуатации хранилищ данных – модифицируемость реляционной схемы.

Решает перечисленные проблемы **универсальная логическая модель данных** (УЛМД), полученная в [2] на полной совокупности связей произвольной группы сущностей в смысле [3], которая моделирует произвольную предметную область (ПО).

ПОСТАНОВКА ЗАДАЧИ

Пусть в произвольной ПО имеется N-совокупность сущностей [3]. Проектировщику хранилища данных необходимо предоставить алгоритм, который позволяет:

– получать полную совокупность реляционных отношений, каждое из которых удовлетворяет критериям как минимум 4НФ [4];

– проводить анализ произвольных предметных областей с целью выявления различных противоречий, в том числе и «множественность» атрибутов;

– учитывать начальные иерархические зависимости произвольного подмножества «слабых» сущностей;

– без вспомогательных графических средств моделировать n-арные, в том числе и рекурсивные связи между сущностями, кардинальность которых в общем случае может быть «многие ко многим»;

– интегрировать произвольное количество атрибутов связей;

– минимальным количеством операций модифицировать структуру модели в процессе эксплуатации.

Под модифицируемостью структуры модели понимается [2] добавление дополнительных или удаление существующих подструктур, а также внесение произвольного количества изменений во «внутреннюю» структуру отношений. Это значит, что искомый алгоритм должен быть полным в смысле видов модификаций, а также не должен зависеть от манипулирующего программного обеспечения. То есть добавление или удаление новых подструктур модели не должно затрагивать уже существующие структуры и связи между ними.

Как будет показано ниже, *особые* отношения [3] обладают свойствами, позволяющими говорить, что искомый алгоритм формирования УЛМД может быть получен на совокупности особых отношений.

В рамках настоящей работы для синтеза УЛМД рассмотрим упрощенную произвольную ПО, в которой отсутствуют рекурсивные связи и иерархические зависимости в исходной группе сущностей. Исследование УЛМД на обобщенных ПО будет проведено отдельно.

В работе [3] была сформулирована и доказана теорема о шунтировании МЗ.

Теорема 1 Если в отношении $R(X, Y, Z)$ имеется нетривиальная МЗ $X \to\to Y \setminus Z$, причем X, Y, Z могут быть составными множествами, то есть $X = \{X\}, Y = \{Y\}, Z = \{Z\}$, 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.

Перепроверим полученный в [3] результат. Рассмотрим одну из теорем Фейджина [4].

Теорема Фейджина МЗ $X \to\to Y$ выполняется для отношения $R(X, Y, Z)$ если и только если R является соединением своих проекций $R_1(X,Y)$ и $R_2(X,Z)$.

На основании этого утверждения реляционное отношение $R(X, Y, Z, A)$ из теоремы 1 не имеет МЗ, потому что не может быть спроектировано ни на один из своих атрибутов.

Очевидно, что в отношении «экзамен» [3] сумма $X + Y + Z$ – единственный возможный ключ. Значит, такое отношение не может быть декомпозировано ни по одному из атрибутов, так как $(X + Y + Z) \to A_i$.

Как указывалось в [3], никакая часть A_i не зависит ни от какой комбинации частей ключа, кроме как от самого ключа. И сам ключ - так же. Таким образом, получена 4НФ.

И хотя формальным признаком МЗ есть повторяемость экземпляров атрибутов, никаких аномалий в отношении не наблюдается. Очевидно, что любую МЗ можно шунтировать, по крайней мере, суррогатным атрибутом, физический смысл которого – точный момент наступления факта N-арной связи – дата и время вплоть до миллисекунд. Такой атрибут – полный аналог суррогатного ключевого атрибута, вводимого пользователем в качестве уникального порядкового номера экземпляра сущности для получения строгой НФБК. Это дает возможность использования наряду с бинарными реляционными отношениями также и таблицы произвольной арности, что значительно расширяет рамки логической модели данных и тем самым решает практически все проблемы, указанные в [1].

Заметим также, что обобщением многозначной зависимости атрибутов является их декартово произведение. Алгоритм декартового произведения порождает *полную* [5] многозначную зависимость. И как показано в [2],

именно процедура сочетаний «всех из всех» сущностей в совокупности с декартовым произведением ключевых атрибутов этих сущностей лежит в основе алгоритма УЛМД. Полноту многозначной зависимости, которую гарантирует процедура декартова произведения, будем в дальнейшем выделять специальным термином – *декартова зависимость* (ДЗ).

УНИВЕРСАЛЬНАЯ ЛОГИЧЕСКАЯ МОДЕЛЬ ДАННЫХ

На основании доказанной теоремы 1 приведем без доказательства несколько утверждений.

Лемма 1 Отношение $R(X_j, A_i)$, полученное шунтированием ДЗ в ключевых атрибутах X_j, причем так, что $\{X_j \to A_i,\ j = 2, J;\ i = 1, I\}$ и предикат каждой X_j-й части ключа уникален, моделирует связь арностью j и кардинальностью «многие ко многим».

При этом совокупность A_i является атрибутами связи, чем и моделируется специфика ПО. Действительно, если совокупность атрибутов данной связи A_i существует, то существует и сама связь. А из теоремы 1 и теоремы Фейджина следует обратное утверждение.

Лемма 2 Если в отношении $R(X_j, A_i)$ совокупность атрибутов A_i - это пустое множество, а $\{X_j\}$ - непустое множество составного ключа, причем такое, что при $j \geq 3$ в R существует нетривиальная ДЗ и предикат каждой X_j-й части ключа уникален, такое отношение $R(X_j)$ *не является актуальным* для данной предметной области.

Иными словами, такое отношение не моделирует ни сущности, ни связи. И может быть либо исключено из совокупности отношений, либо шунтировано атрибутами. Именно найденные в ПО атрибуты придадут такому отношению смысл.

Отметим, что в практике проектирования хранилищ данных принято использовать такие «пустые» отношения в качестве вспомогательных фильтров-справочников, отражающих лишь вероятность факта связи сущностей. Но отсутствие в ПО естественных атрибутов такой связи говорит о возможных аномалиях использования таких справочников.

Лемма 3 Отношение $R(X_j, A_i)$, полученное шунтированием ДЗ в ключевых атрибутах X_j, причем так, что $\{X_j \to A_i,\ j = 2, J; i = 1, I\}$ и предикат каждой X_j-й части ключа уникален, является подобным особому отношению с составным ключевым атрибутом X_j и тем самым моделирует виртуальную сущность с атрибутами A_i.

Такую виртуальную сущность принято называть «постсвязной» [6] («отглагольным существительным»).

Заметим, что леммы 1 и 3 показывают единство категории связи и категории сущности, что согласовывается с идеей Чена [7]. Это вполне закономерно, потому что и та, и иная категории моделируются совокупностью атрибутов. А они инвариантны. С этой точки зрения, метод «сущностей и связей» [7] можно рассматривать как метод «сущностей». Из этого следует еще одно утверждение.

Рассмотрим оператор $L^k(X_j)$, который в соответствии с индексом $k = 1, N$ формирует полное множество сочетаний сумм суррогатных ключевых атрибутов, а формирование множеств значений атрибутов

осуществляется в соответствии с алгоритмом декартового произведения суррогатных ключевых атрибутов сущностей. Здесь $j = 1, \mathrm{N}$ – номер сущностей; k – текущая арность ключа; $l = 1, L_k$ – номер суммы ключевых атрибутов k-й арности. Общая совокупность имеет вид

$$L^1(X_i) = \{X_1, X_2, X_3, X_4, \ldots, X_N\}, \tag{1}$$

$$L^2(X_i) = \begin{Bmatrix} X_1 + X_2,\ X_1 + X_3, X_1 + X_4, \ldots, \\ X_2 + X_3,\ X_2 + X_4, \ldots, X_{N-2} + X_{N-1},\ X_{N-1} + X_N \end{Bmatrix}, \tag{2}$$

$$L^3(X_i) = \begin{Bmatrix} X_1 + X_2\ X_1 + X_3,\ X_1 + X_2 + X_4, \ldots, \\ X_2 + X_3 + X_4, \ldots, X_{N-2} + X_{N-1} + X_N \end{Bmatrix}, \tag{3}$$

$$L^{N-1}(X_i) = \{X_1 + X_2 + X_3 + X_4 + \ldots + X_{N-1},\ X_2 + X_3 + X_4 + \ldots + X_N\}, \tag{4}$$

$$L^N(X_i) = \{X_1 + X_2 + X_3 + X_4 + \ldots + X_N\}. \tag{5}$$

Тогда каждый элемент множества запишется в виде

$$L_1^1(X_j) = (X_1), \tag{6}$$

$$L_2^1(X_j) = (X_2), \tag{7}$$

$$L_3^1(X_j) = (X_3), \tag{8}$$

$$L_N^1(X_j) = (X_N), \tag{9}$$

$$L_l^2(X_j) = (X_1 + X_2), \tag{10}$$

$$L_{L2}^2(X_j) = (X_{N-1} + X_N), \tag{11}$$

$$L_1^3(X_j) = (X_1 + X_2 + X_3), \tag{12}$$

$$L_{L3}^3(X_j) = (X_{N-2} + X_{N-1} + X_N), \tag{13}$$

$$L_1^{N-1}(X_j) = (X_1 + X_2 + X_3 + X_4 + \ldots + X_{N-1}), \tag{14}$$

$$L_2^{N-1}(X_j) = (X_2 + X_3 + X_4 + \ldots + X_N), \tag{15}$$

$$L_l^N(X_j) = (X_1 + X_2 + X_4 + \ldots + X_N). \tag{16}$$

Общее число S полученных групп ключевых атрибутов определяется выражением числа сочетаний [2]:

$$S = \sum \frac{N!}{k!(N-k)!} = 2^N - 1, \tag{17}$$

где N – число сущностей; k – арность ключа, которая соответствует текущему числу связей каждой сущности с другими группами сущностей, а также совпадает с коэффициентом подобия соответствующих отношений.

Тогда с использованием алгоритма оператора $L_l^k(X_j)$ сформулируем утверждение об универсальной логической модели данных. Приведем его пока без доказательства.

Теорема 2 Совокупность отношений $R_l^k(L_l^k(X_j),A_{li})$, в которой при $k=1$ содержится N отношений, каждое из которых построено на предикате *j-й* сущности и имеет единственный суррогатный ключевой атрибут X_j, а также *j-ю* совокупность неключевых атрибутов A_{li}, а при $k=2,N$ содержится совокупность отношений, каждое из которых моделирует связи *k-й* арности всех со всеми сущностям кардинальности «многие ко многим», причем начиная с тернарных связей декартова зависимость в ключевых атрибутах шунтирована так, что $L_l^k(X_j) \to A_{li}^k$, является **полным множеством отношений**, определяющих специфику произвольной предметной области для *N*-сущностей.

Смысл индексов описан выше. Для $k=1$ индекс l совпадает с индексом *j*, а индекс $i=1,I_l$ – это номер неключевого атрибута для l-го отношения. Очевидно, что параметры N и I_l произвольны и определяются конкретной постановкой моделируемой ПО. Повторно отметим, что *полнота множеств*, построенных на процедуре декартового произведения, обоснована в [5] .

Это означает, что на совокупности $R_l^k(L_l^k(X_j),A_{li})$ может быть построена УЛМД, отображающая специфику произвольной ПО из N сущностей на множество реляционных отношений, количество которых S определяется формулой (17).

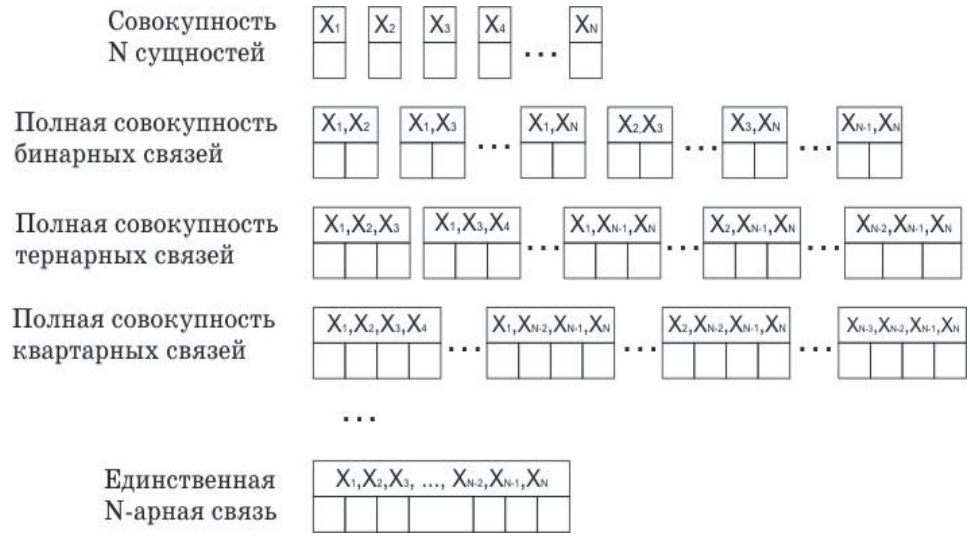

Рисунок 1 – Ключевой каркас реляционной УЛМД для N-сущностей

На рисунке 1 показана общая схема ***ключевого каркаса*** реляционной УЛМД. Очевидно, что большинство отношений не будут актуальными в контексте конкретных постановок информационных задач. Но их актуализация в любой момент и является модификацией структуры конкретного хранилища. Из этого следует, что модификация схемы хранилища сводится к 2 типам операций: актуализация - аннулирование отношения (реляционной таблицы) и актуализация-аннулирование произвольного множества неключевых атрибутов в произвольной группе отношений. При этом целостность хранилища сводится прежде всего к

целостности ключевых атрибутов и их строгого соответствия в различных, но логически связанных отношениях. «Ключевая» целостность отслеживается каскадными функциями, построенными на соответствующих бинарных индексах («b-деревьях») групп ключевых атрибутов. Такие функции должны отрабатываться в режиме реального времени.

Эта концепция модифицирования позволяет исключить *удаление таблицы* из перечня операций по изменению структуры хранилища. Присвоив группе таблиц статус неактуальности, проектировщик лишь снижает количество используемых отношений. Поэтому все обновления, вносимые администратором в структуру такого хранилища, не влияют на хранимые данные.

ВЫВОДЫ

Таким образом, все многообразие предметных областей моделируется совокупностями N-арных сочетаний ключевых атрибутов, каждый из которых отвечает за уникальную сущность. Очевидно, что особенность приведенных отношений заключена именно в том, что они образовывают *полный* [5] каркас всех схем реляционных отношений произвольной ПО. Повторим, что такая система в [2] была получена на основании сочетаний декартовых произведений [5] ключей простейших отношений по принципу „все на все".

Как следует из леммы 3, все таблицы этой совокупности имеют форму не ниже 4НФ [4]. Действительно, *особые отношения* [3] обладают НФБК [8]. Но по теореме о шунтировании [3] все отношения, построенные на ключевых атрибутах каркаса, не имеют нетривиальных МЗ. Это значит, что получена совокупность отношений в 4НФ.

Рассмотрим физический смысл изложенного. Если ключевые атрибуты отношений отвечают одному предикату, то физический смысл таких таблиц - это хранилища данных соответствующих сущностей. Такие отношения в практике проектирования баз данных принято называть «справочниками», указывая на то их свойство, что они не зависят от управляющих программных систем и от времени. В нашей терминологии они не зависят от специфики ПО. Это значит, что совокупность таких отношений моделирует весь объем информации обо всех сущностях, входящих в эту совокупность.

Если же каждая часть ключа особого отношения отвечает различным предикатам, эти отношения являются хранилищем характеристик связей. То есть отношения с шунтированной ДЗ, полученные декартовым произведением следов связывающихся сущностей, моделируют произвольные связи между сущностями любой кардинальности – от «один к одному» до «многие ко многим». Как уже отмечалось, совокупность атрибутов такого отношения является атрибутами этой связи.

Как предложил П.П. Чен в [7], моделировать связи различной кардинальности между сущностями можно с помощью отдельных виртуальных сущностей, в которых функции связи реальных сущностей он назвал *ролями*. Очевидно, что тогда множество $R_l^k(L_l^k(X_j),A_{li})$ есть ни что иное, как полная совокупность сущностей и их ролей. Такая модель позволяет, в частности, реализовывать гибкие в смысле [2] структуры операционных баз данных OLTP-информационных систем [9].

Отметим, что в настоящей работе в описанной выше УЛМД показана часть алгоритма, позволяющая проектировать лишь одноразовые связи каждой сущности с произвольной группой сущностей. Отсутствие ограничений на количество *видов связей* на любом уровне арности в каждой конкретной группе сущностей, свойственное такой сущности,

как, например, «люди» и приводящее к *рекурсивным* связям в изложенном алгоритме учитывается с помощью *копий* сущностей, дополнительных таблиц, смысл которых - *маски для ролей* в связях сущностей. Эта возможность пока не рассматривается. Не рассматривается также и подобие N-арных таблиц связей независимых сущностей M-арным таблицам, моделирующим иерархически зависимые («слабые») сущности. Очевидно, что учет этих особенностей [1] не внесет принципиальных изменений в структуры УЛМД.

SUMMARY

ABOUT THE SYNTHESIS OF UNIVERSAL LOGICAL MODEL OF DATA

B.E. Panchenko
Sumy State University

New approach to the creation of universal logical model of data that is based on multivalued dependencies of key attributes is offered in this work. N-tables are offered to be used for modeling links of cardinality "many-to-many". Problem of modifiability of structure of relational data warehouses is solved.

СПИСОК ЛИТЕРАТУРЫ

1. Малыхина М.П. Базы данных: основы, проектирование, использование, - СПб., 2006. – 528 с.
2. Панченко Б.Е. Способ расположения данных в компьютерном хранилище, обеспечивающий модифицируемость его структуры // Патент Украины № 63036, 2001.
3. Панченко Б.Е. К вопросу о многозначных зависимостях в универсальной логической модели данных // Вестник СумГУ. - № 2. -2009.
4. Fagin R. Multivalued dependencies and a new normal form for relational databases // ACM Transactions on Database Systems. – 1977. - Vol. 2, No. 3. - P. 262-278.
5. Курош А.Г. Общая алгебра. – М., 1979. – 150 с.
6. Ульман Д.Д., Уидом Д. Основы реляционных баз данных. – М.,2006. – 374 с.
7. Chen P.P. The Entity-Relationship Model: toward a unified view of data // ACM Trans. on Data base systems. –1976. - V.1, № 1. - P.9 – 36.
8. Codd E.F. Recent investigations in relational database systems // In Proc. IFIP Congress, 74. - North-Holland, Amsterdam, 1974. - P. 1017-1021
9. Чекалов А.П. Базы данных: от проектирования до разработки приложений. – СПб., 2003. – 384 с.

Поступила в редакцию 6 апреля 2009 г.