УДК 004.652

О ЗАМКНУТОСТИ ПУТЕЙ НОРМАЛИЗАЦИИ РЕЛЯЦИОННОГО КАРКАСА

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

В работе исследованы пути нормализации в универсальном каркасе реляционных баз данных. Изучена топология этих путей. Доказана теорема о замкнутости путей нормализации в реляционном каркасе.

Ключевые слова: *реляционный каркас, универсальная логическая модель данных, модификация схемы хранилища.*

У роботі досліджені шляхи нормалізації в універсальному каркасі реляційних баз даних. Вивчено топологію цих шляхів. Доведено теорему про замкнутість шляхів нормалізації в реляційному каркасі.

Ключові слова: *реляційний каркас, універсальна логічна модель даних, модифікація схеми сховища.*

ВВЕДЕНИЕ

В работах [1,2] предложена общая схема реляционного *ключевого каркаса* универсальной логической модели данных УЛМД, которая строго соответствует схеме множества всех подмножеств комбинаций сумм ключей. То есть, каждая из этих сумм унарных ключей образует отношение, полученое способом сочетания унарных ключевых столбцов [3]. Очевидно, что для проектировщиков промышленных баз данных большинство отношений не будут актуальными в контексте конкретных постановок. Но их актуализация в любой момент является модификацией структуры конкретного хранилища.

Из этого следует, что модификация схемы хранилища сводится к 2 типам операций: деактуализация - аннулирование отношения (реляционной таблицы) и актуализация - восстановление 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.

Там же доказано, что на операторе роста L может быть построена УЛМД, отображающая специфику произвольной предметной области (ПО) из N сущностей на множество реляционных отношений, общее количество которых $S(N)$ определяется формулой числа сочетаний (1):

$$S(N) = \sum_{m=1}^{N} S_m = \sum_{m=1}^{N} \frac{N!}{m!(N-m)!} = 2^N - 1 . \qquad (1.1)$$

Кратко охарактеризуем схему построения предлагаемой каркасной модели [1,2,3].

Рассматривается множество E сущностей x некоторой произвольной предметной области. Каждая сущность $x_j \in E$ предметной области идентифицируется ключом j. В этом множестве выделяются подмножества $G_i \subseteq E$ сущностей, причем сущности одной группы характеризуются своей уникальной совокупностью свойств a_k. Фактически каждая группа G_i определяется свойствами входящих в нее

сущностей $x \in G_i$. Поэтому формирование групп G_i может быть эквивалентно формированию совокупности всех возможных свойств a_k сущностей и выделению из этой совокупности подмножеств, характеризующих группы. Наконец, каждая группа G_i идентифицируется ключом i.

Далее, в соответствии с семантикой выбранной предметной области устанавливаются все возможные связи между сущностями выделенных групп G_i. Это делается для того, чтобы выявить т.н. «иерархические» связи и в соответствии с ними обеспечить «наследование» ключей иерархически связанных групп сущностей. Тем самым выделяются элементарные группы и «сложные» группы, образованные из элементарных. При этом сложные группы идентифицируются составными ключами, образованными ключами элементарных «порождающих» групп. Здесь, помимо простого введения простых и «сложных» (составных) групп объектов, рассматриваются т.н. уровни сложности составных групп, выражаемые количеством «порождающих» групп и соответственно проявляющиеся в количестве элементов составных ключей. На примере составной группы {Дата} подчеркнем упорядоченность совокупности элементарных групп, порождающих составную (например, {Эра}..{День}); это может указывать на требование упорядоченности элементов составного ключа, идентифицирующего «сложную» группу объектов предметной области. Процедура «кластеризации» (т.е. установления семантических связей между объектами различных групп и построения групп «сложных» объектов на основе простых) в некотором смысле является процедурой, обратной требованию 1НФ об атомарности значений элементов кортежей в отношениях (таблицах) создаваемой БД. Например, элементом составной группы {Дата} (т.е. множества всех дат) является совокупность элементов простых групп {Эра}..{День}, т.е. элементы группы {Дата} не являются атомарными. Вообще, элемент любой составной группы по построению является множеством (более строго – упорядоченной последовательностью). Поэтому возможность включения в БД в качестве элементов будущих кортежей сущностей, относящихся к составным группам, нарушает условие 1НФ (а значит и всех последующих НФ). Этот вопрос снимается путем замены каждой составной группы сущностей предметной области условно-простой группой (см. анализ ниже). При этом возникает вопрос о целесообразности идентификации каждой условно-простой группы особым ключом-именем; действительно, все (как простые, так и составные) группы G_i, выделенные в предметной области, уже идентифицированы значениями простого ключа i. Конечно, идентификация группы составных объектов может осуществляться составным ключом, чтобы отразить семантическую зависимость от групп простых порождающих объектов. Фактически существование групп составных объектов в предметной области может интерпретироваться как наличие непустых пересечений между совокупностями свойств (атрибутов), характеризующих группы сущностей.

Как описываются сущности x_j в предложенной схеме? Вводится дополнительное отношение «ключ – атрибуты сущности», где ключом является j, а атрибутами являются свойства сущности, которые «индуцируют» разбиение всей предметной области на группы G_i. Пока опустим классическую проблему соотнесения сущноти и атрибутов, выражающаяся в вопросах: что в каждом конкретном случае следует считать сущностью и что – атрибутом? Если группы G_i строить на основе

некоторых (для простоты атомарных) атрибутов a_k, то сущность x_j, находящей отражение в физической реальности, можно назвать любой кортеж $(a_1,\ldots,a_p)$, $p \le K$, где K - количество атрибутов предметной области. В предельном случае каждый атрибут a_k может рассматриваться в качестве сущности. Тогда под «реализацией» сущности следует понимать кортеж $(a_1,\ldots,a_p)$, заполненный значениями задействованных в нем атрибутов a_k, т.е. фактически наполнение соответствующего кортежа контентом.

Традиционно схемой реляционной базы данных (БД) является некая фиксированная совокупность реляционных схем R_j, т.е. именованных множеств атрибутов и ключей [4]. Для построения такой схемы вводится совокупность атрибутов x_i и однозначно соотносимых с ними множеств значений-доменов $D(x_i)$ [5, 6]. При этом совокупности самих атрибутов ассоциируются с «объектами» или «сущностями», а совокупности *значений* атрибутов - с экземплярами объектов или сущностей; это является первым шагом к отображению семантики предметной области в схеме БД. Заметим, что и множество x_i, и совокупность множеств $D(x_i)$ являются *общими* для схем R_j в том смысле, что отдельный атрибут может принадлежать нескольким схемам. Наконец, экземпляр каждой реляционной схемы R_j представляется в виде совокупности кортежей K_p - упорядоченных последовательностей значений атрибутов x_i схемы R_j, т.е. $K_p \subset D(x_1)\times\ldots\times D(x_i)\times\ldots\times D(x_K)$, $x_i \in R_j$.

ПОСТАНОВКА ЗАДАЧИ

Рассмотрим пути нормализации универсального реляционного каркаса (в дальнейшем - просто *каркаса*). Пусть дана индексированная совокупность реляционных схем $\{C_k\}$, $k = 1,\ 2,\ \ldots,\ S(N)$, образующих каркас отношений [2] для множества из N сущностей в смысле [1] в некоторой произвольной предметной области. Внесение реального контента в каркас отношений (т.е. заполнение реляционных схем сведениями о значениях реальных экземпляров) приводит к формированию совокупности экземпляров (instances) отношений. Обозначим текущий экземпляр отношения $\{C_k\}$ символом $[C_k]$. Для *каждого* из экземпляров $[C_k]$ имеет смысл говорить о множестве Φ_k функциональных либо многозначных зависимостей между атрибутами (или множествами атрибутов) соответствующего отношения $\{C_k\}$. Как правило, множество зависимостей Φ_k считается *независимым* от экземпляра $[C_k]$, т.е. любая модификация $[C_k]$ не меняет Φ_k. Это характерное условие, подразумевающее статичность схемы реляционной базы данных и соответствующих путей нормализации отношений, в дальнейшем будет снято при рассмотрении *динамических* схем.

Пример 1.1. Пусть для источника $A = \{a, b, c, d\}$ ($N = 4$) сформирован каркас [2], состоящий из $2^4 - 1 = 15$ отношений $\{C_k\}$, и пусть образованы соответствующие экземпляры $[C_k]$ этих отношений. Множество зависимостей Φ_k может состоять, например, из: функциональных

зависимостей между атрибутами отношений 1-уровня (для действия L_A^1), т.е. из $a \to a$, $c \to c$, которые всегда будут тривиальными; из зависимостей вида $a \to b$, $c \to\to d$ между атрибутами отношений 2-уровня (для действия L_A^2); из зависимостей вида $ab \to c$, $a \to\to bd$ между атрибутами отношений 3-уровня (для действия L_A^3) и т.д. Важно, что *все* элементы множества зависимостей Φ_k между атрибутами (и множествами атрибутов) отношения можно перечислить комбинаторными методами, вводя тем самым *полное* множество зависимостей Φ_k и его подмножество $\tilde{\Phi}_k \subseteq \Phi_k$, актуальное для конкретного экземпляра $[C_k]$ отношения $\{C_k\}$.

Рассмотрим некоторое отношение $\{C_k\}$ и его экземпляр $[C_k]$ с зависимостями Φ_k. Пусть $K = \{K_1, K_2, K_3, K_4, K_5, \ldots\}$ - упорядоченная совокупность традиционных нормальных форм (НФ), а также всех возможных их модификаций. Пусть ψ - множество классических критериев, относящих C_n к НФ из совокупности K:

$$\psi_n = \psi\left(C_n\right),\ \psi_n \in K. \qquad (1.2)$$

Учитывая взаимосвязь $K_1 \subset K_2 \subset K_3 \subset \ldots$ между НФ из совокупности K, обозначим через $\Psi_k = \max \psi_k$ наибольшую НФ, в которой находится экземпляр $[C_n]$ отношения $\{C_n\}$.

Рассмотрим некоторое «начальное» подмножество $D^{(0)}$ каркаса отношений. Следует отметить, что $D^{(0)}$ является элементом каркаса схем баз данных. *Состоянием* схемы $D^{(0)}$ назовем совокупность $\{\Psi_k\}$ НФ экземпляров $[C_k]$ всех отношений $C_k \in D^{(0)}$. Ясно, что НФ, в которой будет находиться схема $D^{(0)}$ (и в общем случае, весь каркас отношений), определяется величиной $\max\{\Psi_k\}$. Именно поэтому *все* отношения схемы баз данных приводятся (как правило, путем декомпозиции) к одной, наибольшей НФ. Хотя целесообразность достижения схем отдельных отношений критериям высоких НФ остается вопросом дискуссионным.

В результате такой нормализации мы получаем последовательность (2.2) элементов каркаса схем баз данных, которую можно интерпретировать как путь нормализации. Формально *путь нормализации* $Q\left(j_0, j_1, \ldots, j_k\right)$ представляет собой последовательность индексов схем баз данных, которая описывает переход от начального элемента $D^{(0)}$ к конечному элементу $D^{(k)}$ каркаса схем баз данных, причем этот переход осуществляется *только декомпозицией* отношений, принадлежащих элементам пути нормализации. Конечный элемент $D^{(k)}$ пути нормализации мы будем называть *решением*. Далее, *топологией* путей нормализации для заданной начальной схемы $D^{(0)}$ будем называть совокупность решений $D^{(k)}$, которые можно получить путем декомпозиции при заданных зависимостях $\{\Phi_k\}$ между атрибутами отношений. Ясно, что топология будет определяться как $D^{(0)}$, так и $\{\Phi_k\}$.

ТЕОРЕМА О ЗАМКНУТОСТИ ПУТЕЙ НОРМАЛИЗАЦИИ РЕЛЯЦИОННОГО КАРКАСА

Имеет место следующая теорема о замкнутости путей нормализации: для заданных источника A и совокупности $\{\Phi_k\}$ зависимостей между атрибутами в экземплярах $[C_k]$ каркаса отношений существует путь нормализации $D^{(0)} \to \ldots \to D^{(k)}$, решение которого находится в *требуемой* НФ $\max\{\Psi_k\}$ из совокупности K. Это утверждение непосредственно следует из полноты каркасов отношений и схем баз данных [2].

Рассмотрим источник $A = \{a, b, c, d\}$ ($K = 4$), для которого сформирован каркас из 15 отношений $\{C_k\}$. Пусть $D^{(0)}$ - начальная схема реляционной базы данных, содержащая среди прочих единственное отношение $C = \{abcd\}$ 4-уровня (т.е. синтезированное действием L_A^4), $C \in D^{(0)}$. Пусть семантика предметной области такова, что для экземпляра $[C]$ выполняется следующее множество функциональных и многозначных зависимостей Φ: $abc \to d$, $a \to\to b$, $a \to\to c$. Каким может быть решение $D^{(1)}$ для простейшего, одношагового пути нормализации?

Поскольку все атрибуты из A являются атомарными, то $\psi(C) = K_1$, т.е. C находится в НФ1. Далее, для функциональной зависимости $abc \to d$ множество атрибутов abc является суперключом, поэтому $\psi(C) = \{K_1, K_2, K_3, K_4\}$, т.е. C находится также в НФ2, НФ3 и НФБК. Поскольку для каждой из многозначных зависимостей $a \to\to b$ и $a \to\to c$ в отношении C ни a ни b не являются суперключами, $\psi(C) \neq K_5$, т.е. C не находится в НФ4. Поэтому $\Psi = K_4$. Для дальнейшего увеличения Ψ на единицу, т.е. для получения НФ4, можно было бы потребовать, чтобы в отношениях искомого решения $D^{(1)}$ отсутствовали *нетривиальные* многозначные зависимости, такие, как $a \to\to b$ и $a \to\to c$.

Заметим, что многозначная зависимость может существовать для отношений не ниже 3-уровня (т.е. синтезированных действиями $L_A^{m\geq 3}$), т.е. когда отношение имеет минимум 3 атрибута. В силу насыщения оператора роста наибольшим уровнем является 4-й, где и находится отношение C. Ясно, что решение $D^{(1)}$, для которого $\Psi = K_5$ (НФ4), не должно содержать отношения C. Можно попробовать редуцировать схему $D^{(1)}$ до совокупностей отношений, порождаемых действиями $L_A^{m<3}$, т.е. применить т.н. ограничение каркаса отношений сверху. Выполняя декомпозицию отношения C, получаем

$$\{abcd\} \mapsto \{ab\} \cup (\{acd\} \mapsto \{ac\} \cup \{ad\}) = \{ab\} \cup \{ac\} \cup \{ad\},$$

т.е. производится замена отношения 4-уровня на совокупность отношений 2-уровня, которые и будут присутствовать в решении $D^{(1)}$. При этом для отношений $\{ab\}$ и $\{ac\}$ многозначные зависимости $a \to\to b$ и $a \to\to c$ будут уже тривиальными, т.е. критерий 4НФ будет выполняться.

ВЫВОДЫ

Из примера 1.2 видно, что исходное отношение C 4-уровня $\{abcd\}$ после декомпозиции на совокупность отношений 2-уровня уже не

содержит функциональной зависимости $abc \to d$, т.е. информация об этой зависимости для $\{ab\} \cup \{ac\} \cup \{ad\}$ теряется. Если под сохранностью информации понимать обеспечение соединения без потерь [3, 4, 5] и одновременно обеспечение сохранения зависимостей, то такую декомпозицию можно считать декомпозицией с потерей информации. Имеем тройку критериев, которые в случае корректной декомпозиции должны быть выполнены одновременно: возможность восстановления кортежей (т.е. обеспечение соединения без потерь), сохранение имеющихся зависимостей, а также соответствие критериям «целевой» нормальной формы (в примере 1.2 это 4НФ) [7]. Если между критериями из этой совокупности возникает противоречие, т.е. если невозможно одновременно выполнить хотя бы два критерия из трех, декомпозиция будет некорректной. При получении желаемой нормальной формы такая некорректность может быть выражена либо искажениями при соединении дочерних отношений, либо искажениями в зависимостях между атрибутами. Известно, что для «высоких» нормальных форм - НФБК, 4НФ, как в примере 1.2, а также 5НФ указанная тройка критериев может быть выполнена не во всех случаях: такая ситуация изложена в [4] для НФБК, в [7] для 4НФ, а также в [8] - для 4НФ и 5НФ. Искажения при соединении считаются недопустимыми, поэтому искажения в зависимостях рассматриваются в качестве приемлемого, хотя и нежелательного, «артефакта» при нормализации реляционных схем.

Отметим, однако, что *ключевой каркас,* приведенный на рисунке 1, получен в [1] с использованием теоремы о шунтировании декартовой зависимости. Ключевой каркас является частным случаем реляционного каркаса, описанного в [2]. Частный случай обеспечивается уникальным множеством *суррогатных ключевых атрибутов* набора сущностей. Как отмечалось в [1], ключевой каркас строго соответствует критериям 4НФ. Проделав над множеством ключевых атрибутов процедуры, аналогичные изложенным выше, несложно из реляционного каркаса получить ключевой. На множестве суррогатных ключевых атрибутов он по-прежнему будет полным и единственным.

В заключение можно предположить, что дальнейший анализ топологии путей нормализации, использующий особенности структуры универсального каркаса схем реляционных баз данных, позволит выработать единый универсальный метод определения решений для задач синтеза и/или модификации логических структур реляционных баз данных.

SUMMARY

ABOUT CLOSURE OF NORMALIZATION PATHS OF THE RELATIONAL FRAMEWORK

B.E. Panchenko
Sumy State University, Sumy

In this paper normalization paths in the universal framework for relational databases have been analyzed. Theorem about closure of normalization paths in a relational framework has been proved

Key words: *a relational skeleton, universal logic model of the data, updating of the scheme of warehouse.*

СПИСОК ЛИТЕРАТУРЫ

1. Панченко Б.Е. О синтезе универсальной логической модели данных / Б.Е. Панченко // Вестник СумГУ. Серия Технические науки. - 2009. - № 2 – С. 60-66.
2. Панченко Б.Е. Теорема о полноте и единственности реляционного каркаса / Б.Е. Панченко // Вестник СумНАУ. Серия Механиз. и автоматиз. произв. проц. - 2009. – Вып. 1 (20). – С. 67-76.

3. Панченко Б.Е. Способ расположения данных в компьютерном хранилище, обеспечивающий модифицируемость его структуры / Б.Е. Панченко // Патент Украины. - 2001. - № 63036.
4. Мейер Д. Теория реляционных баз данных / Д. Мейер. - М., 1987. – 608 с.
5. Дейт К.Дж. Введение в системы баз данных / К.Дж. Дейт. – 7-е изд. - М.: Вильямс, 2001. - 1072 с.
6. Ульман Дж. Основы систем баз данных / Дж. Ульман. - М.: Финансы и статистика, 1983. - 334 с.
7. Carver A., Halpin T., Atomicity and normalization // Proceedings of the 13th International Workshop on Exploring Modelling Methods for Systems Analysis and Design (EMMSAD08). - Montpellier, France, 2008. - P. 40-54.
8. Silberschatz A., Korth H.F., Sudarshan S. Database system concepts // PRC edition, McGraw-Hill Higher Education Press. – 2002.

Поступила в редакцию 28 сентября 2009 г.