УДК 004.652

О ТОПОЛОГИИ ПУТЕЙ НОРМАЛИЗАЦИИ В РЕЛЯЦИОННОМ КАРКАСЕ

Б.Е. ПАНЧЕНКО, И.Н. ПИСАНКО

Исследованы пути нормализации в универсальном каркасе реляционных баз данных (БД) и топологии этих путей. Доказана теорема замкнутости путей нормализации в реляционном каркасе. Теорема позволяет применять реляционный каркас как уникальный носитель схем БД, нормализованных до высоких форм, а также анализировать существующие и внедренные БД на предмет их аномалий и влияния на приложения в процессе эксплуатации.

ВВЕДЕНИЕ

Получение безаномальных логических схем является одной из ключевых задач проектирования логической структуры реляционных баз данных (БД). Известно, что аномалии вставки, модификации и удаления кортежей связаны с наличием зависимостей (функциональных, многозначных, а также зависимостей соединения) между подмножествами столбцов таблиц БД [1, 2, 3]. Такие зависимости выражают ограничения, накладываемые на значения, хранимые в кортежах таблиц. В свою очередь, эти ограничения являются формальным отображением семантики, присущей конкретным предметным областям. Для предметных областей с нетривиальной семантикой, характеризуемых многообразными и сложными ограничениями на хранимые данные, избыточность представления данных в таблицах является типичным недостатком, обуславливающим аномалии логических схем БД.

Традиционная методика устранения таких аномалий состоит в декомпозиции проблемных таблиц БД. Согласно Дейту, всякая БД есть хранилище истинностных высказываний (фактов) о так называемых «сущностях» заданной предметной области и их взаимосвязях [2]. Поэтому декомпозиция является не чем иным как представлением некоторого хранимого высказывания в виде эквивалентной совокупности других (более простых) высказываний. При этом эквивалентность между исходным высказыванием с семантическим ограничением, с одной стороны, и совокупностью высказываний без семантического ограничения, с другой стороны, обеспечивается целым рядом определений и теорем о декомпозиции без потерь [3, 4]. Таким образом, посредством декомпозиции происходит нормализация представления семантически сложного факта в БД за счет его разбиения на более простые факты, рассматриваемые в их совокупности. При этом для сложных фактов

с семантическими ограничениями, как правило, имеется более чем один вариант такого разбиения. Можно заметить некоторую аналогию между указанным процессом разбиения фактов БД и функциональной декомпозицией.

Традиционно схемой реляционной БД является некая фиксированная совокупность реляционных схем R_j, т.е. именованных множеств атрибутов и ключей [3]. Для построения такой схемы вводится совокупность атрибутов x_i и однозначно соотносимых с ними множеств значений — доменов $D(x_i)$ [4]. 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. $K_p \subset D(x_1)\times...\times D(x_i)\times...\times D(x_K)$, $x_i \in R_j$.

В работах [5, 6] показано, что на операторе роста L может быть построена универсальная логическая модель данных (УЛМД), отображающая специфику произвольной предметной области (ПО) из N сущностей на множество реляционных отношений, полное количество которых $S(N)$ определяется формулой:

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

На рисунке изображена общая схема *ключевого каркаса* реляционной УЛМД. Очевидно, что большинство отношений не будут актуальными в контексте конкретных постановок. Но их актуализация в любой момент яв-

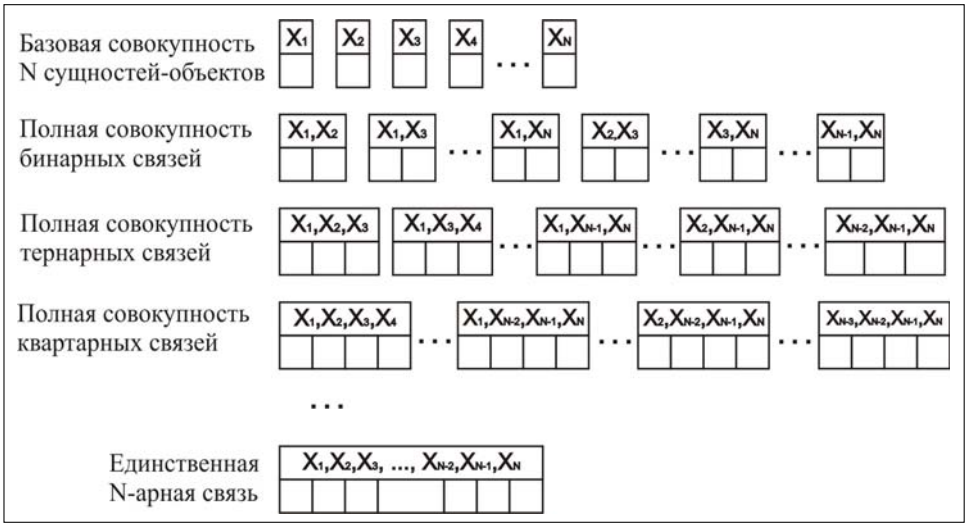

Рисунок. Ключевой каркас реляционной УЛМД для N сущностей-объектов

ляется модификацией структуры конкретного хранилища.

Из этого следует, что модификация схемы хранилища сводится к двум типам операций актуализации: аннулирование отношения (реляционной таблицы) и 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.

ПОСТАНОВКА ЗАДАЧИ

Цель работы — исследовать схемы БД, получаемые на универсальном реляционном каркасе (в дальнейшем, просто *каркасе*), на принадлежность той или иной нормальной форме (НФ). Для этого рассмотрим пути нормализации каркаса.

Рассмотрим некоторое отношение $\{C_k\}$ и его экземпляр $[C_k]$ с зависимостями Φ_k. Пусть $K = \{K_1, K_2, K_3, K_4, K_5, ...\}$ — упорядоченная совокупность традиционных НФ универсального реляционного каркаса. Пусть дана индексированная совокупность реляционных схем $\{C_k\}$, $k = 1, 2, \ldots, S(N)$, образующих каркас отношений [6] для множества из N сущностей в смысле [5] 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 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 БД и соответствующих путей нормализации отношений, в дальнейшем будет снято при рассмотрении динамических схем.

Пример 1. Пусть для источника $A = \{a, b, c, d\}$ ($N = 4$) сформирован каркас [6], состоящий из $2^4 - 1 = 15$ отношений $\{C_k\}$, а также образованы соответствующие экземпляры $[C_k]$ этих отношений. Множество зависимостей Φ_k может состоять, например, из: функциональных зависимостей между атрибутами отношений 1-го 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-го уровня (для действия L_A^2); из зависимостей вида $ab \to c$, $a \to\to bd$ между атрибутами отношений 3-го 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 $\widetilde{\Phi}_k \subseteq \Phi_k$, актуальное для конкретного экземпляра $[C_k]$ отношения $\{C_k\}$.

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

$$\psi_n = \psi(C_n),\ \psi_n \in K. \qquad (2)$$

Учитывая взаимосвязь $K_1 \subset K_2 \subset K_3 \subset ...$ 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 БД. 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 БД 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 элементов каркаса схем БД, которую можно интерпретировать как путь нормализации [6]. Формально путь нормализации $Q(j_0, j_1, ..., j_k)$ представляет собой последовательность индексов схем БД, которая описывает переход от начального элемента $D^{(0)}$ к конечному элементу $D^{(k)}$ каркаса схем БД, 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

Рассмотрим теорему замкнутости: для заданных источника A и совокупности $\{\Phi_k\}$ зависимостей между атрибутами в экземплярах $[C_k]$ каркаса отношений существует путь нормализации $D^{(0)} \to ... \to D^{(k)}$, решение которого находится в требуемой НФ $\max\{\Psi_k\}$ из совокупности K. Это непосредственно следует из полноты каркасов отношений и схем БД [6].

Пример 2. Вновь рассмотрим источник $A = \{a, b, c, d\}$ ($K = 4$), для которого сформирован каркас из пятнадцати отношений $\{C_k\}$. Пусть $D^{(0)}$ — начальная схема реляционной БД, содержащая среди прочих единственное

отношение $C = \{abcd\}$ 4-го уровня (т.е. синтезированное действием L_A^4), $C \in D^{(0)}$. Пусть семантика ПО такова, что для экземпляра $[C]$ выполняется следующее множество функциональных и многозначных зависимостей Φ: $abc \to d$, $a \to\to b$, $a \to\to c$. Следует выяснить, 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 1НФ. 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 2НФ, 3НФ 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 4НФ. Поэтому $\Psi = K_4$. Для дальнейшего увеличения Ψ на единицу, т.е. для получения 4НФ, 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-го уровня (т.е. синтезированных действиями $L_A^{m\geq 3}$), т.е. когда отношение имеет минимум три атрибута. В силу насыщения оператора роста наибольшим уровнем является 4-й, где и находится отношение C. Ясно, что решение $D^{(1)}$, для которого $\Psi = K_5$ (4НФ), не должно содержать отношения C. Можно попробовать редуцировать схему $D^{(1)}$ до совокупностей отношений, порождаемых действиями $L_A^{m<3}$, а именно применить так называемое 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-го уровня на совокупность отношений 2-го 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 2 видно, что исходное отношение C 4-го уровня $\{abcd\}$ после декомпозиции на совокупность отношений 2-го 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] 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 (обеспечение соединения без потерь), сохранение имеющихся зависимостей, а также соответствие крите-

риям «целевой» нормальной формы (в примере 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 2, а также 5НФ указанная тройка критериев может быть выполнена не во всех случаях: такая ситуация изложена в [3] 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, получен в [5] с использованием теоремы о шунтировании многозначной зависимости. Ключевой каркас является частным случаем реляционного каркаса, описанного в [6]. Частный случай обеспечивается уникальным множеством суррогатных ключевых атрибутов набора сущностей. Как отмечалось в [5], 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 БД, позволит выработать единый универсальный метод определения решений для задач синтеза и/или модификации логических структур реляционных БД.

ЛИТЕРАТУРА

1. *Codd E.F.* The Relational Model For Database Management, Version 2, Reading Mass. — NY: Addison-Wesley Publishing Co, 1990. — 538 p.
2. *Дейт К.Дж.* Введение в системы баз данных. — 7-е изд. — М.: Вильямс, 2001. — 1072 с.
3. *Мейер Д.* Теория реляционных баз данных. — М.: 1987. — 608с.
4. *Ульман Дж.* Основы систем баз данных. — М.: Финансы и статистика, 1983. — 334 с.
5. *Панченко Б.Е.* О синтезе универсальной логической модели данных // Вестн. СумГУ. Сер. Технические науки. — 2009. — № 2. — С. 60–66.
6. *Панченко Б.Е., Писанко И.Н.* О полноте и единственности универсального каркаса в реляционной модели данных // Системні дослідження та інформаційні технології. — 2010. — № 3. — С. 25–35.
7. *Carver A., Halpin T.* Atomicity and normalization // Proceedings of the 13-th International Workshop on Exploring Modelling Methods for Systems Analysis and Design, June, 29, from CEUR–WS archive. — Montpellier, France. — 2008. — **337**. — P. 40–54.
8. *Silberschatz A., Korth H.F., Sudarshan S.* Database system concepts // PRC edition, McGraw-Hill Higher Education Press, 2006. — 1129 p.

Поступила 01.06.2009