

СВОЙСТВА РЕЛЯЦИОННОГО КАРКАСА НА МНОЖЕСТВЕ СЕМАНТИЧЕСКИ АТОМАРНЫХ ПРЕДИКАТОВ

Ключевые слова: *реляционный каркас, семантически атомарный предикат, схема реляционной базы данных, кортеж, домен.*

Схемой реляционной базы данных (БД) в традиционном представлении является некоторая фиксированная совокупность реляционных схем R_j , т.е. именованных множеств атрибутов и ключей [1]. Как правило, для построения такой схемы вводится совокупность атрибутов x_i и однозначно соотносимых с ними множеств значений — доменов $D(x_i)$ [2]. При этом совокупности самих атрибутов ассоциированы с объектами или сущностями, а совокупности значений атрибутов — с экземплярами объектов или сущностей. Это — первый шаг к отображению семантики предметной области в схеме БД. Заметим, что множество x_i и совокупность множеств $D(x_i)$ общие 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. $K_p \subset D(x_1) \times \dots \times D(x_i) \times \dots \times D(x_K)$, $x_i \in R_j$.

Разнородные взаимосвязи между множествами атрибутов, выявляемые анализом семантики предметных областей, играют решающую роль в процессе нормализации схемы БД. Можно сказать, что при подходе «атрибуты-домены» вначале задается аморфное «поле», на котором выявленные семантические связи позволяют сформировать (после ряда преобразований, как правило, достаточно длинного для нетривиальных примеров) нормализованную структуру БД. Такое «формирование нормализованной структуры» является одним из наиболее трудоемких этапов конструирования логической схемы БД.

Несмотря на то что существуют развитые техники декомпозиции схем БД [2], подобный подход к конструированию имеет кардинальный недостаток: отсутствие гибкости. Именно изменение совокупности атрибутов и/или взаимосвязей между ними (например, при расширении предметной области БД, что на практике случается чаще всего) вызывает необходимость повторного формирования нормализованной схемы БД. В подобных случаях логическая схема БД либо трудно модифицируема, либо вообще не поддается модификации (поскольку между старой и новой схемами БД нет строгой преемственности). Для отображения в БД новых (дополнительных) объектов, сущностей, связей требуется сформировать расширенную совокупность объектов и связей и затем вновь выполнить процедуру нормализации, т.е. фактически решить задачу декомпозиции заново. При этом корректная процедура нормализации схемы БД должна обеспечивать сохранность информации — как соединение без потерь, так и сохранение наличных зависимостей, что далеко не всегда осуществимо для нормальных форм Бойса–Кодда и выше. Можно утверждать, что подход к построению схемы БД, основанный на декомпозиции, в некотором смысле неустойчив по отношению к исходному множеству атрибутов и связей. Альтернативные алгоритмы синтеза нормализованных реляционных схем по Бернштейну отказывают, если семантика предметной области предполагает наличие многозначных зависимостей или зависимостей соединения.

Нормализация все чаще представляется чем-то несовершенным и требующим кардинального пересмотра [3, 4]. При этом сама реляционная модель эволюционирует, вмещая в себе подходы, основанные на моделях универсального и бинарного отношений, модели «сущность-связь», объектно-ориентированной и семантической моделях [1, 2]. Каждый из этих подходов использует свою «точку зрения» (viewpoint) предметной области и собственное модельное представление, отображаемое в схеме реляционной БД.

© Б.Е. Панченко, И.Н. Писанко, 2009

Здесь и далее реляционная БД рассмотрена как структурированная совокупность фактов. Используем логическую интерпретацию БД [2, 5] как естественно подходящую для синтеза универсальной структуры, вмещающей схемы реляционных БД. Кроме того, это позволит изначально дистанцироваться от подхода «сущности-атрибуты-домены».

Всякий факт формально является n -местным j -предикатом $P_j^n(x_i)$, принимающим истинное значение, $P_j^n(x_i) = 1$ [6]. Здесь x_i — аргументы предиката P , взятые без конкретизации в качестве объектов предметной области или их атрибутов, и тем более, без конкретизации каких бы то ни было взаимосвязей между ними. При этом аргументы x_i предикатов — так называемые предметные (индивидуальные) переменные [7]. Заметим, что появление в БД ложного предиката, $P_j^n(x_i) = 0$, недопустимо и может указывать на некорректное функционирование БД, например, если сама схема БД не нормализована и допускает аномалии вставки, модификации или удаления. Аргументом x предиката может быть любая переменная (или, в частном случае, постоянная) с единичной кардинальностью, выделяемая в предметной области. При этом не уточняется, относится ли семантически аргумент предиката к какому-либо объекту или атрибуту объекта.

Пусть семантика предметной области предполагает наличие K таких переменных x . Индекс $i = 1, \dots, K$ является идентификатором каждого аргумента x_i , вне зависимости от привязки к предикатам. Естественно, для предикатов $P_j^n(x_i) \forall j: n \leq K$; например, $P_9^3(x_1, x_5, x_{17})$, $P_2^5(x_3, x_4, x_{11}, x_{12}, x_{108})$. Индекс $j = 1, \dots, M$ идентифицирует n -местный предикат $P_j^n(x_i)$ в схеме БД. Заметим, что для предметной области с K аргументами x_i возможно до 2^K предикатов, различных только по количеству аргументов: одноместных, двуместных и т.д., и единственного K -местного предиката. Однако предикаты из одной группы «вместимости» (арности) $P_A^3(x_1, x_2, x_3)$ и $P_B^3(x_1, x_2, x_3)$ могут быть высказываниями, имеющими отличную одна от другой семантику. Например: «Человек $X(x_1)$ в городе $Y(x_2)$ владеет машиной с номером $Z(x_3)$ » и «Человек $X(x_1)$ в городе $Y(x_2)$ пострадал в ДТП с участием машины с номером $Z(x_3)$ ».

Такое построение соответствует методологии Дейта [2], когда БД задана совокупностью истинных высказываний (но не «экземпляров» объектов), а оперирование БД интерпретируется как получение новых высказываний из имеющейся совокупности. При этом есть прямая аналогия с реляционной моделью: всякий выполнимый предикат $P_j^n(x_i)$ — аналог реляционной схемы R_j , а каждое истинное высказывание для этого предиката — аналог кортежа в схеме R_j . Если для каждого из аргументов x_i предикатов $P_j^n(x_i)$ ввести свою область значений D_i , $x_i \in D_i$, аналогия с кортежами становится очевидной. Здесь допустима привязка к семантике предметной области: «Формулы [предикатов] имеют смысл только тогда, когда имеется какая-нибудь интерпретация входящих в них символов. Под интерпретацией мы понимаем всякую систему, состоящую из непустого множества D , называемого областью интерпретации, и какого-либо соответствия, относящего каждой предикатной букве P_j^n некоторое n -местное отношение в D , каждой функциональной букве p_j^n — некоторую n -местную операцию в D (т.е. $p_j^n: D^n \rightarrow D$) и каждой предметной постоянной x_i — некоторый элемент из D . При заданной интерпретации предметные переменные x_i мыслятся пробегающими область D этой интерпретации, а ... всякое n -местное отношение в D может рассматриваться как некоторое подмножество множества D^n всех n -ок элементов из D » [7, с. 57]. Конечно, истинное высказывание $P_j^n(x_i) = 1$ в БД справедливо только для определенных значений аргументов предиката, т.е. предикат в БД не должен быть тавтологией.

Реляционная модель предполагает наличие связей (relationships). Как правило, связи интерпретируются как особый вид сущностей (entities), которые, в свою оче-

Схема 2^α фактически есть булеан (множество всех подмножеств) ключей семантических атомов. Универсальность схемы состоит в единообразии описания всех возможных отношений между семантическими атомами. Структура схемы 2^α , называемой реляционным каркасом (relational framework), детально рассмотрена в работах [9, 10].

Очевидно, что реальная предметная область содержит только некоторое (актуальное) подмножество $2^{\alpha^+} \subseteq 2^\alpha$ из всех возможных семантических связей. Это так называемая актуальная часть булеана ключей семантических атомов. Связи из 2^{α^+} , не противоречащие семантике предметной области, назовем семантически актуальными, а все остальные — семантически потенциальными. Вся совокупность потенциально возможных связей будет описываться множеством $2^{\alpha^-} = 2^\alpha \setminus 2^{\alpha^+}$. Это — потенциальная часть булеана ключей семантических атомов. Технически такое разбиение можно реализовать индексацией элементов булеана 2^α с последующим введением бинарной переменной δ_m , принимающей значение 0 для потенциальной связи $(2^\alpha)_m$ и значение 1 для семантически актуальной связи $(2^\alpha)_m$. Заметим, что в некотором смысле изложенный подход снимает дилемму Кодда о «выборе между единственным универсальным отношением и множеством бинарных отношений» [1, с. 468].

Рассмотрим, каким образом «межатомные» связи в их классической интерпретации (т.е. в виде функциональной и многозначной зависимостей, а также зависимостей соединения) переносятся на множество 2^{α^+} . В общем виде механизм такой трансляции связей заключается в следующем. Пусть заданы атомарные предикаты $P_A \in A$ и $P_B \in A$. Пусть между $x_A \subseteq (x_i)_A$ и $x_B \subseteq (x_i)_B$ существует (другими словами, актуальна) некоторая зависимость λ , т.е. $x_A \lambda x_B$. По определению ключа семантического атома имеем $(\alpha \rightarrow x_i)_A$ и $(\alpha \rightarrow x_i)_B$, следовательно, $\alpha_A \rightarrow x_A$ и $\alpha_B \rightarrow x_B$. Поэтому зависимость $x_A \lambda x_B$ между подмножествами аргументов предикатов P_A и P_B транслируется в зависимость $\alpha_A \lambda \alpha_B \in 2^{\alpha^+}$ между ключами α_A и α_B соответствующих семантических атомов.

Из этого примера следует важное уточнение к определению семантически атомарного предиката: предикат будет атомарным тогда, когда никакие подмножества его аргументов не фигурируют ни в каких зависимостях как одно с другим, так и с подмножествами аргументов других предикатов (правило взаимной семантической атомарности). Действительно, пусть $x_{A_1} \subseteq (x_i)_A$, $x_{A_2} \subseteq (x_i)_A$, $x_{A_1} \neq x_{A_2}$ и $x_B \subseteq (x_i)_B$. Пусть также $x_{A_1} \lambda x_B$ и $x_{A_2} \lambda x_B$. В этом случае имеем $\alpha_A \rightarrow x_{A_1}$, $\alpha_A \rightarrow x_{A_2}$ и $\alpha_B \rightarrow x_B$. Пара семантически различных зависимостей $x_{A_1} \lambda x_B$ и $x_{A_2} \lambda x_B$ будет транслироваться в одну зависимость $\alpha_A \lambda \alpha_B \in 2^{\alpha^+}$ между ключами α_A и α_B семантических атомов, т.е. в этом случае актуальная часть 2^{α^+} булеана 2^α не будет способна отобразить указанную пару зависимостей. Из правила взаимной семантической атомарности следует, что $x_A = (x_i)_A$ и $x_B = (x_i)_B$, поэтому никаких затруднений с трансляцией зависимости λ на 2^{α^+} не возникает. Отсюда очевидно, что конструирование класса A семантически атомарных предикатов необходимо начинать с рассмотрения наименьших (по числу аргументов) детерминант зависимостей, выделенных в предметной области.

Синтез реляционного каркаса сводится к построению множества 2^{α^+} на булеане ключей семантических атомов и выражается такой последовательностью действий:

- 1) в предметной области выделяется множество x_i аргументов предикатов;
- 2) формируется совокупность предикатов ${}_A P_j^n(x_i)$, семантически атомарных (так называемое базовое множество предикатов реляционного каркаса);
- 3) для атомарных предикатов ${}_A P_j^n(x_i)$ вводятся ключи α_j ;
- 4) для отображения связей строится булеан 2^α ключей семантических атомов;
- 5) строится актуальная часть 2^{α^+} булеана ключей семантических атомов;
- 6) выполняется построение логической схемы реляционной БД, эквивалентной 2^{α^+} .

Данный алгоритм впервые был изложен в [11]. Возможность использования реляционного каркаса в качестве универсального «носителя» данных для предметных областей с произвольно заданной семантикой обусловлена следующими характеристиками каркаса: единственностью и полнотой, а также устойчивостью к модификации базового множества атомарных предикатов.

Для доказательства этих характеристик используем представление об операторе роста [10] применительно к ключам α_j семантических атомов. В общем виде оператор роста L действует как $\{C_k\} = L_A(B)$, $C_k = B + a$, $a \in A \setminus B$.

Оператор L продуцирует индексированную совокупность $\{C_k\}$ множеств C_k , содержащих все элементы множества B и один из элементов множества A , не принадлежащий B . Множество A называют источником, а множество B — базовым множеством. Многократное (m раз) действие оператора L на базовое множество B обозначаем как $L_A^m(B)$. Заметим, что базовое множество может быть совокупностью множеств (тогда оператор роста действует на каждое из множеств базовой совокупности). Поэтому при многократном действии оператора роста его аргументом будет совокупность множеств, полученных в результате предыдущих действий оператора. Таким образом, сам оператор роста L определяется индуктивно.

Детальный анализ свойств оператора роста L указывает на то, что структуры, порождаемые L на множествах любой природы, эквивалентны схеме $\mathbf{B}(\alpha)$, или 2^α , т.е. собственно реляционному каркасу. Так, в качестве элементов базового множества и источника выступают ключи семантических атомов.

Теорема 1 (о полноте и единственности реляционного каркаса). Каркас, порождаемый оператором роста на конечном множестве атомарных предикатов, единствен и полон.

Докажем вначале полноту реляционного каркаса. Пусть заданы источник A и базовое множество B . Пусть также существует такое множество $C = \{c_i\}$, что $C \supseteq B$ и $\exists c \in A \setminus B$, но $\forall m: C \neq L_A^m(B)$, т.е. допустим, что не существует действия, порождающего множество C . Тогда из определения оператора роста следует $C \neq B \cup C_m^*$, где $\forall m: C_m^* \subseteq A \setminus B$, значит, $\exists c \in C_m^*$. Но это противоречит допущению $\exists c \in A \setminus B$. Следовательно, действие $L_A^m(B)$, порождающее множество C , действительно существует. Поэтому схема 2^α потенциально реализуема и полна, являясь фактически булеаном отношений между семантическими атомами.

Доказательство единственности реляционного каркаса основано на его полноте. В силу полноты каждого из каркасов $\{C_k\} = L_A^m(B)$ и $\{C_k^*\} = L_A^m(B)$ имеем $\{C_k\} \setminus \{C_k^*\} = \emptyset$ и $\{C_k^*\} \setminus \{C_k\} = \emptyset$, следовательно, $\{C_k\} = \{C_k^*\}$, т.е. каркасы идентичны. Это означает, что на конечном множестве семантических атомов оператор роста порождает единственный реляционный каркас.

Практическое значение имеет модифицируемость логических схем БД при изменении предметных областей, например при добавлении новых сущностей (не их экземпляров!) или удалении существующих (встречается гораздо реже). В этом контексте важна следующая теорема.

Теорема 2 (об устойчивости реляционного каркаса). Если выполняется правило взаимной семантической атомарности, то изменение базового множества атомарных предикатов на единицу не влияет на исходную структуру реляционного каркаса.

Действительно, пусть к базовому множеству $B = \{P_j\}$ атомарных предикатов P_j , на котором построен каркас 2_B^α , добавляется некоторый новый предикат $X \notin B$.

Согласно правилу взаимной семантической атомарности каждый из элементов нового базового множества $B + X$ является атомарным. Поэтому на этом новом базовом множестве реализуем (единственным образом!) новый реляционный каркас 2_{B+X}^α . Устойчивость каркаса при добавлении нового атомарного предиката состоит в независимости исходной структуры каркаса от дополнения, определяемого предикатом X , т.е. выражается условием $2_{B+X}^\alpha = 2_B^\alpha + 2_X^\alpha$, причем $2_B^\alpha \cap 2_X^\alpha = \emptyset$. Здесь каркас 2_B^α является полным, или «насыщенным» [10], а каркас 2_X^α состоит из единственного элемента, т.е. $2_X^\alpha = L_X(\emptyset)$. Допустим, что ключ α_X содержится в каркасе 2_B^α , т.е. $2_B^\alpha \cap 2_X^\alpha \neq \emptyset$. Согласно теореме о полноте реляционного каркаса (применительно к

каркасу 2_B^α) для некоторого базового множества B^* существует источник A^* , содержащий ключ $\alpha_X: 2_B^\alpha = L_{A^* \supset \alpha_X}(B^*)$. Это означает, что предикат X является элементом базового множества, на котором построен каркас 2_B^α , т.е. $X \in B$. Но это противоречит исходному допущению о добавлении нового предиката: $X \notin B$. Следовательно, $2_B^\alpha \cap 2_X^\alpha = \emptyset$, т.е. дополнение реляционного каркаса при увеличении базового множества атомарных предикатов на единицу не имеет общих элементов с исходным каркасом и поэтому никак не влияет на исходную структуру отношений между атомарными предикатами.

Устойчивость реляционного каркаса относительно модификации базового множества атомарных предикатов при «сжатии» предметной области, т.е. при удалении из базового множества B атомарных предикатов одного из них, доказывается аналогично. Если предметная область изменяется не на один, а сразу на несколько атомарных предикатов, то такая модификация сводится к последовательности единичных изменений.

Кардинальное отличие реляционного каркаса, синтезируемого на семантически атомарных предикатах, от множества существующих концепций организации данных (в которых традиционно формируется совокупность сущностей с присущими им атрибутами, затем для них определяется множество отношений-связей, после чего задается некоторая «начальная» схема БД, anomalies которой устраняются путем декомпозиции) заключается в том, что для реляционного каркаса прежде всего формируется предметная область (как множество аргументов предикатов) и задается ее семантика (как множество зависимостей между этими аргументами). На этой основе определяется семантически атомарных предикатов. Реляционный каркас, синтезируемый (единственным образом) на базовом множестве атомарных предикатов, выступает в качестве «носителя» данных о фактах предметной области. Такое свойство каркаса, как его полнота, гарантирует, что все (актуальные и потенциальные) отношения между парами, тройками, четверками и т.д. атомарных предикатов можно единообразно описать в рамках одной структуры — реляционного каркаса. А такое свойство каркаса, как его устойчивость к модификациям базового множества атомарных предикатов, гарантирует, что при изменении предметной области (добавлении новых атомарных предикатов или удалении существующих) не возникнет потребность в тотальной реорганизации всей структуры. Поэтому семантически атомарные предикаты выступают в качестве оптимальных «сущностей» для представления данных о предметных областях, а сам реляционный каркас — в качестве оптимального «носителя» данных о взаимосвязях между этими сущностями.

СПИСОК ЛИТЕРАТУРЫ

1. Codd E. F. The relational model for database management. Version 2. — New York: Addison-Wesley, 1990. — 538 p.
2. Date C. J. Date on database: writings 2000–2006. — New York: Apress, 2006. — 539 p.
3. Fatache M. Why normalization failed to become the ultimate guide for database designers // Soc. Sci. Res. Network. — 2006. — ID 905060. — 230 p.
4. Carver A., Halpin T. Atomicity and normalization // Proc. of the 13th Intern. Workshop on Exploring Modelling Methods for Systems Analysis and Design (EMMSAD08). — Montpellier (France), 2008. — 337. — P. 40–54.
5. Morales-Luna G. Simple epistemic logic fir relational database // Lect. Notes Artif. Intellig. — 2002. — 2313. — P. 234–241.
6. Barwise J., Etchemendy J. Language, proof and logic. — New York: Seven Bridges Press, 1999. — 587 p.
7. Мендельсон Э. Введение в математическую логику. — М.: Наука, 1976. — 320 с.
8. Jackson D. Software abstractions: logic, language, and analysis. — S.l.: MIT Press, 2006. — 350 p.
9. Панченко Б. Е. О синтезе универсальной логической модели данных // Вестн. СумГУ. Сер. «Техника». — Сумы: СумГУ, 2009. — № 2. — С. 60–66.
10. Панченко Б. Е. Теорема о полноте и единственности реляционного каркаса // Вестн. СНАУ. — Сер. «Механизация и автоматизация производ. процессов». — 2009. — № 1(20). — С. 67–76.
11. Пат. 63036. Спосіб розміщення даних у комп'ютерному сховищі із забезпеченням модифікаційності його структури / Б.С. Панченко // Промислова власність. — 2004. — № 1. — С. 3.134 (Заявл. 11.15.2001).

Поступила 08.07.2009