УДК 519.681.3

ТЕОРЕТИЧЕСКИЕ ПРЕДПОСЫЛКИ ФОРМАЛИЗОВАННОЙ НОРМАЛИЗАЦИИ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

И.В. Клименко

ФГБОУ ВПО «Ростовский государственный университет путей сообщения», г. Ростов-на-Дону

Рецензент д-р техн. наук, профессор С.М. Ковалев

Ключевые слова и фразы: задача о наименьшем покрытии; нормализация отношений реляционной модели; нормальные формы; первичный ключ; функциональные зависимости.

Аннотация: Проведен анализ текущего состояния вопроса нормализации отношений реляционной модели. Проанализирован опыт основоположников современной теории баз данных, выявлены причины возникновения противоречивости, избыточности хранимых данных. Обоснованы теоретические предпосылки создания формализованной методики к процессу нормализации отношений реляционной модели. Предложен новый подход к решению задачи нормализации.

В современной теории баз данных наиболее актуальной считается задача нормализации отношений реляционной модели, так как до настоящего времени она не является формализованной [1–3].

Настоящая статья является логически первой в системе статей, описывающих методику решения проблемы формализованной нормализации реляционных баз данных (**БД**). Предлагаемую систему дополняют следующие статьи.

1. Метод построения семейства максимальных транзитивно независимых множеств атрибутов.

2. Модификация алгоритма Балаша для решения задачи нормализации реляционных баз данных.

Рассмотрим теоретические предпосылки возможности формализации процесса нормализации реляционных баз данных.

Цель нормализации – понижение избыточности и повышение достоверности хранимых данных. Кроме того, безызбыточные файлы БД зани-

Клименко Игорь Валерьевич – кандидат технических наук, доцент кафедры «Вычислительная техника и автоматизированные системы управления» e-mail: klimigor67@mail.ru, ФГБОУ ВПО «Ростовский государственный университет путей сообщения», г. Ростов-на-Дону.

мают меньше места на внешних носителях и требуют меньше времени при обработке информационных запросов пользователей.

Любое априорное знание о различного рода ограничениях, накладываемых на совокупности данных, может принести большую пользу для достижения целей нормализации.

Один из способов формализации этих знаний – установление зависимостей между элементами данных (атрибутами). Известно два основных типа таких зависимостей:

1) функциональные;
2) многозначные.

Предметом рассмотрения является нормализация на основе анализа функциональных зависимостей.

Функциональная зависимость, или F-зависимость, имеет место тогда, когда значения кортежа на одном множестве атрибутов единственным образом определяют эти значения на другом множестве атрибутов.

Пусть r – отношение со схемой R, X и Y – подмножества R. Отношение r удовлетворяет функциональной зависимости $X \to Y$, если

$$t_1(X) = t_2(X), \text{ то } t_1(Y) = t_2(Y) \text{ для } \forall \text{ кортежей } t_1 \text{ и } t_2 \text{ в } r.$$

Для отношения $r(R)$ в любой момент существует некоторое семейство F-зависимостей, которым это отношение удовлетворяет. Здесь может возникнуть следующая проблема: одно состояние отношения может удовлетворять F-зависимости, а другое – нет.

Требуется выявить семейство F-зависимостей F, которому удовлетворяют все допустимые состояния r. Чтобы найти F, необходимы семантические знания об отношении r. Поэтому можно считать семейство F-зависимостей заданным в схеме отношения R. В этом случае любое отношение $r(R)$ должно удовлетворять всем F-зависимостям из F. Не всегда ясно, что является первичным: множество допустимых состояний отношения, которое определяет F-зависимости, или F-зависимости накладывают ограничения на схему отношения [3].

Множество функциональных зависимостей, применимых к отношению $r(R)$, конечно, так как существует только конечное число подмножеств множества R. Таким образом, всегда можно найти все F-зависимости, которым удовлетворяет r. Однако этот подход требует большого количества шагов и, соответственно, много времени.

Если известны некоторые F-зависимости из F, то по определенным правилам можно вывести остальные.

Множество F-зависимостей F влечет за собой зависимость $X \to Y$, если каждое отношение, удовлетворяющее всем зависимостям в F, удовлетворяет также зависимости $X \to Y$.

Пусть аксиома вывода – это правило, устанавливающее, что если отношение удовлетворяет определенным F-зависимостям, то оно должно удовлетворять и некоторым другим F-зависимостям.

В теории реляционных баз данных сформулированы шесть аксиом вывода F-зависимостей. В этих формулировках используется обозначение r для отношения на R и W, X, Y и Z – для подмножеств R.

F1. *Рефлексивность*: $X \to X$.

F2. *Пополнение*: $X \to Y$ влечет за собой $XZ \to Y$.

F3. *Аддитивность*: $X \to Y$ и $X \to Z$ влечет за собой $X \to YZ$.

F4. *Проективность*: $X \to YZ$ влечет за собой $X \to Y$.

F5. *Транзитивность*: $X \to Y$ и $Y \to Z$ влечет за собой $X \to Z$.

F6. *Псевдотранзитивность*: $X \to Y$ и $YZ \to W$ влечет за собой $XZ \to W$.

Некоторые аксиомы вывода могут быть получены из других. Например, транзитивность F5 является частным случаем псевдотранзитивности F6 при $Z = \varnothing$.

Приведенная система аксиом F1–F6 является полной. Это означает, что каждая F-зависимость, которая следует из множества F, может быть выведена путем одно- или многократного применения к F этих аксиом.

Из аксиом F1, F2 и F6 можно вывести остальные, а значит, они образуют полное подмножество для F1–F6. Аксиомы F1, F2 и F6 являются также независимыми: ни одна из этих аксиом не может быть получена из двух других. Иногда эти три аксиомы называются аксиомами Армстронга [3].

Пусть F – множество F-зависимостей для отношения $r(R)$. Замыкание F, обозначаемое F^+, – это наименьшее содержащее F множество, такое, что при применении к нему аксиом Армстронга нельзя получить ни одной F-зависимости, не принадлежащей F. Так как F^+ должно быть, то можно вычислить его, начиная с F, путем применения F1, F2 и F6 и добавления полученных F-зависимостей к F до тех пор, пока не перестанут получаться новые зависимости. Замыкание F зависит от схемы R.

Из множества F можно вывести F-зависимость $X \to Y$, если $X \to Y$ принадлежит F^+. Так как аксиомы вывода порождают только функциональные зависимости, то F влечет за собой $X \to Y$, если $X \to Y$ выводится из F.

Пример.

Пусть $F = \{AB \to C, C \to B\}$ – множество F-зависимостей на $r(ABC)$.

Тогда: $F^+ = \{A \to A, AB \to A, AC \to A, ABC \to A, B \to B, AB \to B, BC \to B, ABC \to B, C \to C, AC \to C, BC \to C, ABC \to C, AB \to AB, ABC \to AB, AC \to AC, ABC \to AC, BC \to BC, ABC \to BC, ABC \to ABC, AB \to C, AB \to AC, AB \to BC, AB \to ABC, C \to B, C \to BC, AC \to B, AC \to AB\}$.

Для данной схемы отношения R ключ – это подмножество $K \subseteq R$, такое, что для любого допустимого отношения $r(R)$ не существует двух различных кортежей t_1 и t_2 в r, таких, что $t_1(K) = t_2(K)$, и никакое собственное подмножество $K' \subset K$ не обладает этим свойством.

Таким образом, нормализация – формальный метод анализа отношений на основе их первичного ключа (или потенциальных ключей) и существующих функциональных зависимостей.

Нормализация чаще всего выполняется в несколько последовательных этапов, результатом каждого из которых является некоторая нормальная форма с известными свойствами.

Нормальная форма представляет собой ограничение на схему базы данных (отношения), которое избавляет базу данных от некоторых нежелательных свойств.

В теории реляционных баз данных определено несколько нормальных форм (**НФ**), которые подчиняются правилу вложенности (рисунок): каждая нормальная форма является в некотором смысле более ограниченной, но и более желательной, чем предшествующая. Это связано с тем, что (N + 1) нормальная форма не обладает некоторыми недостатками, свойственным N-й нормальной форме. Общий смысл дополнительного условия, налагаемого на (N + 1) нормальную форму по отношению к N-й нормальной форме, состоит в исключении этих недостатков.

В большинстве фундаментальных трудов [3] отмечается достаточность и целесообразность целям практики достижения третьей нормальной формы (**3НФ**).

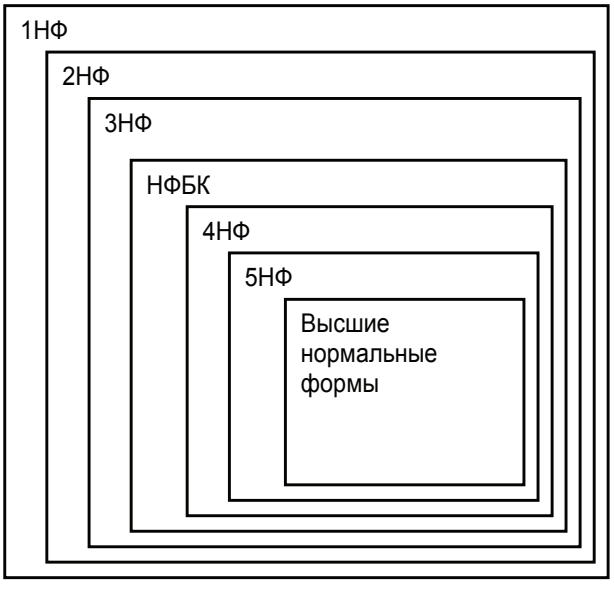
Отношение БД находится в 3НФ относительно множества функциональных зависимостей F, если оно удовлетворяет условиям 1НФ и ни один из непервичных атрибутов не является транзитивно зависимым от ключа [3].

Так как в наборе атрибутов любого отношения можно выделить несколько потенциальных ключей, то целью нормализации в 3НФ является устранение транзитивных зависимостей между атрибутами, входящими в одно отношение.

Пусть дан граф $G(X, F)$, где X – множество атрибутов интересующей предметной области, F – множество функциональных зависимостей между ними.

Определим транзитивно независимое множество (**ТНМ**) атрибутов как подграф графа G такой, что, применяя к нему аксиому F5 правил вывода функциональных зависимостей, вид подграфа не меняется.

Транзитивно независимое множество называется максимальным (**МТНМ**), когда нет другого ТНМ, в которое оно бы входило.



Вложенность нормальных форм

Каждое МТНМ интерпретируется как совокупность атрибутов интересующей предметной области, входящих в одно отношение, находящееся в 3НФ (как минимум).

Очевидно, для оптимизации процесса нормализации необходимо иметь полное семейство МТНМ. Задача построения полного семейства МТНМ может быть решена на основе применения алгоритма Брона–Кэрбоша.

Следующий этап – определение схемы нормализованной базы данных (как совокупности схем отношений). Для этого предстоит решить задачу наименьшего покрытия (**ЗНП**) *G* максимальными ТНМ.

Задача наименьшего покрытия своим названием обязана следующей теоретико-множественной интерпретации [4]. Даны множество $X = \{x_1,...,x_M\}$ и семейство $\Im = \{S_1,...,S_N\}$ множеств $S_j \subset R$. Любое подсемейство $\Im' = \{S_{j_1}, S_{j_2},..., S_{j_k}\}$ семейства $\Im$, такое, что

$$\bigcup_{i=1}^{k} S_{j_i} = X, \qquad (1)$$

называется покрытием множества *X*, а множества S_{j_i} называются покрывающими множествами. Если в дополнение к соотношению (1) $\Im'$ удовлетворяет условию

$$S_{j_h} \cap S_{j_l} = \varnothing, \quad \forall h, l \in \{1, ..., k\}, \quad h \neq l, \qquad (2)$$

то есть множества S_{j_i} ($i = 1, ..., k$) попарно не пересекаются, то $\Im'$ называется разбиением множества *R*.

Если каждому $S_j \in \Im$ поставлена в соответствие (положительная) стоимость c_j, то ЗНП формулируется так: найти покрытие множества *X*, имеющее наименьшую стоимость, причем стоимость семейства $\Im' = \{S_{j_1}, S_{j_2},..., S_{j_k}\}$ определяется как $\sum_{i=1}^{k} C_{j_i}$.

В матричной форме, когда строки ($M \times N$)-матрицы $[t_{ij}]$, состоящей из нулей и единиц, покрываются столбцами, ЗНП может быть сформулирована как задача линейного программирования: минимизировать

$$z = \sum_{j=1}^{N} c_j \xi_j,$$

при ограничениях

$$\sum t_{ij} \xi_j \geq 1,$$

где $c_j > 0$,

$$\xi_j = \begin{cases} 1, \text{ если } S_j \in \Im'; \\ 0, \text{ если } S_j \notin \Im', \end{cases} \quad \text{и} \quad t_{ij} = \begin{cases} 1, \text{если } x_i \in S_j; \\ 0, \text{ если } x_i \notin S_j. \end{cases}$$

Вследствие особой природы ЗНП, часто удается сделать при ее исследовании определенные, хорошо известные заранее выводы и упрощения [4].

Основные из них следующие:

1) если для некоторого элемента x_i из X справедливы соотношения $x_i \notin S_j \ \forall\, j = 1, \ldots, N$, то x_i покрыть нельзя и, следовательно, задача не имеет решения;

2) если $\exists\, x_i \in X$ такое, что $x_i \in S_k$ и $x_i \notin S_j$, $\forall\, j \neq k$, то S_k должно присутствовать во всех решениях и задачу можно свести к «меньшей», положив $X = X - \{x_i\}$ и $\Im = \Im - \{S_k\}$;

3) пусть $V_i = \{j \mid x_i \in S_j\}$; тогда, если $\exists\, p, q \in \{1, \ldots, M\}$ такие, что $V_p \subseteq V_q$, то x_q можно удалить из X, поскольку любое множество, которое покрывает x_p, должно также покрывать x_q, то есть x_p доминирует над x_q;

4) если для некоторого семейства множеств $\overline{\Im} \subset \Im$ справедливы соотношения

$$\bigcup_{S_j \in \overline{\Im}} S_j \supseteq S_k \quad \text{и} \quad \sum_{S_j \in \overline{\Im}} c_j \leq c_k$$

для любых $S_k \in \Im - \overline{\Im}$, то S_k может быть вычеркнуто из $\Im$, поскольку $\bigcup_{S_j \in \overline{\Im}} S_j$ доминирует над S_k.

Когда все упрощения выполнены (если они возможны), исходная ЗНП переформулируется в неприводимой форме и может быть решена одним из эффективных методов целочисленного программирования, например, алгоритмом Балаша [5].

Список литературы

1. Голенищев, Э.П. Информационное обеспечение систем управления : учеб. пособие / Э.П. Голенищев, И.В. Клименко. – Ростов н/Д : Феникс, 2010. – 352 с.

2. Клименко, И.В. Метод формальной нормализации отношений реляционной модели / И.В. Клименко, А.В. Лозовский // Науч. мысль Кавказа. Прил. № 5. – 2004. – С. 115–119.

3. Мейер, М. Теория реляционных баз данных / М. Мейер. – М. : Мир, 1987. – 608 с.

4. Кристофидес, Н. Теория графов. Алгоритмический подход : пер. с англ. / Н. Кристофидес. – М. : Мир, 1978. – 432 с.

5. Рейнгольд, Э. Комбинаторные алгоритмы. Теория и практика / Э. Рейнгольд, Ю. Нивергельт, Н. Део ; пер. с англ. Е.П. Липатова ; под ред. В.Б. Алексеева. – М. : Мир, 1980. – 480 с.

Theoretical Prerequisites of Formalized Normalization of Relational Database

I.V. Klimenko

Rostov State University of Railways, Rostov-on-Don

Key words and phrases: functional dependency; normal forms; normalization of relations; primary key; problem of the least coverage; relational model.

Abstract: The analysis of the current status of the issue of the normalization of relations of relational model is made. The experience of the founders of modern database theory is analyzed; the cause of the inconsistency and redundancy of stored data is identified. The theoretical prerequisites for the creation of a formalized methodology to the process of normalization of relations of relational model are substantiated. A new approach to solving the problem of normalization is offered.
