

С.С. ТАНЯНСКИЙ, канд. техн. наук,
В.В. ТУЛУПОВ, канд. техн. наук,
Д.А. РУДЕНКО, канд. техн. наук

МОДИФИКАЦИЯ ОГРАНИЧЕНИЙ НА ВЕДЕНИЕ ДАННЫХ ДЛЯ ОБЕСПЕЧЕНИЯ ЦЕЛОСТНОСТИ КРУПНОМАСШТАБНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

В статті досліджені методи підтримки цілісності в базах даних і визначені властивості деяких операцій реляційної алгебри при їхньому використанні в розподілених системах. Для розширення можливостей роботи з даними введені узагальнені операції, що дозволяють коректно виконувати запити при наявності невизначених значень у відношеннях бази даних. Для визначення необхідних умов забезпечення цілісності пропонується використовувати алгоритм Грехема, що перевіряє властивість повного з'єднання відношень. Якщо схема бази даних не задовольняє цим умовам, то розглядається варіант її модифікації без порушення схем відношень і зв'язків між ними.

In article the methods of support of integrity in databases are investigated and the properties of some operations of relational algebra are determined at their use in the allocated systems. For expansion of possibilities of work with the data the generalized operations are entered allowing it is correct to carry out searches at presence of indeterminate values concerning a database. For definitions of necessary conditions of provide integrity, it is offered to use algorithm Grahama, checking property complete joining of the relations. If the scheme of a database does not satisfy with it conditions, the variant of its updating without infringement of the schemes of the relations and Relationships between them is considered.

Постановка проблемы. Современные тенденции информационных систем (ИС) состоят в переходе от централизованных вычислительных систем к распределенным системам. Выделяется два вида систем обработки данных: системы распределенных баз данных (БД) и системы распределенной обработки данных. В системе распределенных БД данные распределены между несколькими (возможно территориально разобобщенными) ЭВМ и обеспечены соответствующие возможности для управления этими разделенными частями. В системах распределенной обработки данных, которые имеют распределенные между взаимосвязанными ЭВМ вычислительные мощности и программное обеспечение, данные хранятся на централизованной БД. Системы распределенной обработки данных в дальнейшем будем называть крупномасштабными.

В статье основное внимание уделяется именно крупномасштабным системам. Использование крупномасштабных систем необходимо, с одной стороны, для обеспечения управляемого доступа к распределенным данным, с другой стороны, для независимого функционирования локальных БД при обеспечении согласованности данных во всей системе (глобально).

Существует два типа представлений логической структуры БД: глобальное представление и пользовательское (локальное) представление. При этом с одним и тем же глобальным логическим представлением может быть связано несколько локальных представлений.

Один из способов построения системы управления крупномасштабной системой состоит в интеграции всех БД на основе глобальной схемы. В дальнейшем под глобальной схемой будем понимать схему, состоящую из имен всех атрибутов БД, которые соответствуют как синтаксической, так и семантической уникальности. Схему, отвечающую таким свойствам, будем называть универсальной схемой.

Предположим, что универсальной схемы не существует. На практике это трудно осуществимо, так как необходимо решить следующие задачи:

- определить адреса, где находятся запрашиваемые данные;
- сформулировать запрос, который будет выполняться в разных БД;
- передать результаты на один компьютер для их объединения;
- собрать и объединить результаты;
- выделить из объединенных результатов ответ на исходный запрос.

Эти задачи требуют больших затрат времени, подвержены ошибкам и поэтому являются объектами автоматизации в крупномасштабных системах.

Анализ литературы. Среди работ, посвященных моделированию данных, можно выделить монографии Д. Цикритзиса [1] и М.Ш. Цаленко [2], позволяющие определить основные требования к функционированию ИС и моделированию БД в частности.

Для детального изучения методов проектирования и обработки реляционных БД можно обратиться к работам Д. Мейера [3], Дж. Ульмана [4] и Н.Н. Буслика [5], в которых определены основные свойства операций над отношениями и рассмотрены ограничения на возможные состояния БД.

Несмотря на подробное описание методов обработки данных, в литературе мало внимания уделяется расширению моделей для включения в качестве допустимых значений неопределенностей.

Цель статьи. Свойства БД, основанные на реляционной модели, соответствуют первоначальным предложениям Кодда. Из [6] следует, что наиболее очевидным "узким местом" предложенной реляционной модели данных является неразвитость средств выражения ограничений целостности, которые в общем

случае требуется определить для отношений, связи между которыми нарушены (иными словами, речь идет о недостаточности средств выражения "семантики").

В связи с этим целью статьи является расширение возможностей реляционной алгебры за счет введения в множество возможных значений значения "неопределенно" и соответствующей модификации некоторых операций.

Обобщение реляционных операторов. Операции, предложенные в [3], при использовании неопределенностей не отражают необходимую информацию. Рассмотрим обобщенные операции реляционной алгебры на отношениях с неопределенными значениями. Здесь Rel^\uparrow и Rel обозначают множества всех частичных (допускаются неопределенности) и полных (не содержащих неопределенные значения) отношений соответственно, чьи схемы входят в некоторый фиксированный универсум атрибутов U .

Лякруа и Пирот в [3] предложили обобщенную операцию естественного соединения, обозначаемую \gg^+ , при которой каждая строка отношений-операндов попадает в соединение. Строки, которые не соединяются с другими строками, дополняются неопределенными значениями и присоединяются к результату. Если применять операцию \gg^+ для более чем двух отношений, то закон ассоциативности не выполняется, то есть $(r \gg^+ s) \gg^+ p \neq r \gg^+ (s \gg^+ p)$, и следовательно, операция неоднозначна при любых соединениях отношений.

Заниоло в [3] предложил обобщение операции соединения, в обозначении \gg^Z , при которой строки соединяются, если на всяком общем атрибуте они согласованы, либо в точности одна из них не определена. Если строка не определена на некотором атрибуте, соединенная строка берет значение из другой строки. Если некоторая строка, не соединяется ни с какой другой строкой, то она дополняется неопределенными значениями и добавляется к результату. Однако, операция \gg^Z также не ассоциативна, то есть $(r \gg^Z s) \gg^Z p \neq r \gg^Z (s \gg^Z p)$ и однозначно не выполняется для отношений, при многократных соединениях.

Введем операцию соединения, в обозначении \gg^U , при котором каждая строка попадает в соединение. Отличие \gg^U от рассмотренных операций состоит в следующем: во-первых, отношения, участвующие в соединении, должны принадлежать множеству из Rel ; во-вторых, строки соединяются только тогда, когда они полностью соединимы, несоединимые строки дополняются неопределенными значениями и присоединяются к результату.

Рассмотрим операцию \gg^U для нескольких отношений. Пусть даны отношения $r(A, B)$, $s(B, C)$, $p(C, D)$ представленные на рис. 1. Результат операции $r \gg^U s \gg^U p$ показан на рис. 2.

r	<table border="1"><tr><th>A</th><th>B</th></tr><tr><td>a</td><td>b</td></tr><tr><td>a</td><td>b₁</td></tr><tr><td>a₁</td><td>b₁</td></tr><tr><td>a₂</td><td>b₂</td></tr></table>	A	B	a	b	a	b ₁	a ₁	b ₁	a ₂	b ₂
A	B										
a	b										
a	b ₁										
a ₁	b ₁										
a ₂	b ₂										
s	<table border="1"><tr><th>B</th><th>C</th></tr><tr><td>b</td><td>c</td></tr><tr><td>b₁</td><td>c</td></tr><tr><td>b₁</td><td>c₁</td></tr></table>	B	C	b	c	b ₁	c	b ₁	c ₁		
B	C										
b	c										
b ₁	c										
b ₁	c ₁										
p	<table border="1"><tr><th>A</th><th>C</th></tr><tr><td>a</td><td>c</td></tr><tr><td>a</td><td>c₁</td></tr><tr><td>a₁</td><td>c</td></tr><tr><td>a₂</td><td>c₂</td></tr></table>	A	C	a	c	a	c ₁	a ₁	c	a ₂	c ₂
A	C										
a	c										
a	c ₁										
a ₁	c										
a ₂	c ₂										

Рис. 1. Пример трех не полностью соединимых отношений

r'	<table border="1"><tr><th>A</th><th>B</th><th>C</th></tr><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>a</td><td>b₁</td><td>c</td></tr><tr><td>a</td><td>b₁</td><td>c₁</td></tr><tr><td>a₁</td><td>b₁</td><td>c</td></tr><tr><td>a₂</td><td>b₂</td><td>⊥</td></tr><tr><td>a₂</td><td>⊥</td><td>c₂</td></tr></table>	A	B	C	a	b	c	a	b ₁	c	a	b ₁	c ₁	a ₁	b ₁	c	a ₂	b ₂	⊥	a ₂	⊥	c ₂
A	B	C																				
a	b	c																				
a	b ₁	c																				
a	b ₁	c ₁																				
a ₁	b ₁	c																				
a ₂	b ₂	⊥																				
a ₂	⊥	c ₂																				

Рис. 2. Пример операции \gg^U при многократном соединении

Сформулируем утверждение ассоциативности операции \gg^U .

Утверждение: Пусть $r(R)$, $s(S)$ и $p(P)$ – несогласованные отношения из Rel , тогда оператор \gg^U ассоциативен, то есть $r \gg^U (s \gg^U p) = (r \gg^U s) \gg^U p$.

Справедливость утверждения следует из определения полностью соединимых отношений, дополняя недостающие значения при соединении неопределенными значениями. Отношения $S = \{r_1, \dots, r_n\}$ называются полностью соединимыми, если каждый кортеж в каждом отношении является элементом некоторого списка соединимых на S кортежах. Другими словами, отношения являются полностью соединимыми, если выполняется условие $r_i = \pi_{R_i}(S)$, где R_i – схема отношения r_i , а π – операция проекции.

Например, воспользовавшись отношениями, представленными на рис. 1, применяя операцию \gg^U , получим результат, соответствующий рис. 3.

$r \gg^U (s \gg^U p)$	<table border="1"><tr><th>A</th><th>B</th><th>C</th></tr><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>a</td><td>b₁</td><td>c</td></tr><tr><td>a</td><td>b₁</td><td>c₁</td></tr><tr><td>a₂</td><td>⊥</td><td>c₂</td></tr><tr><td>a₂</td><td>b₂</td><td>⊥</td></tr></table>	A	B	C	a	b	c	a	b ₁	c	a	b ₁	c ₁	a ₂	⊥	c ₂	a ₂	b ₂	⊥
A	B	C																	
a	b	c																	
a	b ₁	c																	
a	b ₁	c ₁																	
a ₂	⊥	c ₂																	
a ₂	b ₂	⊥																	
$(r \gg^U s) \gg^U p$	<table border="1"><tr><th>A</th><th>B</th><th>C</th></tr><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>a</td><td>b₁</td><td>c</td></tr><tr><td>a₁</td><td>b₁</td><td>c</td></tr><tr><td>a₂</td><td>b₂</td><td>⊥</td></tr><tr><td>a₂</td><td>⊥</td><td>c₂</td></tr></table>	A	B	C	a	b	c	a	b ₁	c	a ₁	b ₁	c	a ₂	b ₂	⊥	a ₂	⊥	c ₂
A	B	C																	
a	b	c																	
a	b ₁	c																	
a ₁	b ₁	c																	
a ₂	b ₂	⊥																	
a ₂	⊥	c ₂																	

Рис. 3. Пример ассоциативности операции \gg^U

Введем модифицированную операцию проекции для отношений с неопределенными значениями в обозначении π_X^U .

Пусть r – отношение из $Rel\uparrow$ со схемой R и $X \subset R$, тогда $\pi_X^U(r)$ есть отношение $r'(X)$, полученное вычеркиванием столбцов, соответствующих $R-X$, где $R, X \subseteq U$ и $X \subseteq R$, а также исключением из оставшихся повторяющихся строк и строк с неопределенными значениями.

Проекция отношения r' (см. рис. 2) на схемы $r_1'(A, B)$, $r_2'(B, C)$, $r_3'(AC)$ с помощью обобщенного оператора имеют вид, соответствующий рис. 1.

Рассмотренная модификация операций проекции и соединения позволяет поддерживать некоторый вид ограничений, в частности функциональные зависимости (ФЗ) между атрибутами, путем соединения не полностью соединимых отношений, получив при этом отношение, подобное универсальному представлению U , но с неопределенными значениями в несоединимых кортежах. Если правая часть ФЗ однозначно определяет левую или является неопределенным значением, то неопределенность в правой части зависимости недопустима. Для анализа возможных вариантов использования модифицированных операций необходимо рассмотреть "поведение" ФЗ на отношениях с неопределенными значениями.

Слабоуниверсальное отношение и функциональные зависимости. При рассмотрении интегрированной БД потребуем, чтобы атрибут имел одинаковый смысл всюду, где он появляется, то есть, чтобы он играл определенную роль некоторого класса сущностей. Такое ограничение будем называть предположением об универсальной реляционной схеме (ПУРС). Следствием ПУРС является то, что строки с данным множеством атрибутов имеют одинаковый смысл, следовательно, БД, удовлетворяющая ПУРС, не может содержать двух отношений с одной и той же схемой. Множество атрибутов определяет единственную семантическую связь между ними.

Рассматривая БД как семантическое целое, будем допускать отношения со схемами, являющимися подсхемами других отношений. В таких случаях для того, чтобы БД удовлетворяла ПУРС, определим ограничения на такие отношения. Пусть $r(R)$ и $s(S)$ – отношения БД, причем $R \subset S$. По ПУРС множество атрибутов должно однозначно определять семантическую связь между ними. Поскольку $R \subset S$, то, какова бы ни была связь между его атрибутами, она должна быть частью связи между атрибутами S . То есть, если t – строка в S , то $t(R)$ должна быть строкой в r . Другими словами, в r должна входить $\pi_R(s)$. Это ограничение будем называть условием вхождения схем.

Рассмотрим условие вхождения при обновлении БД. При выполнении условия вхождения может показаться, что разрешение на обновление s влечет за собой разрешение на обновление r . С другой стороны, обновление r накладывает ограничение на обновление s , поскольку строка t не может быть добавлена к s , если $t(R) \notin r$. Введенное ограничение на ПУРС необходимо для поддержки ограничений на зависимости существования [7], так как эти ограничения накладываются на тип связи. Условие вхождения схем используется также при поддержке ФЗ. При проектировании БД по средствам нормализации, поддержка зависимостей в каждом локальном отношении гарантирует выполнение этих зависимостей и в U , то есть глобальное выполнение ФЗ. Для поддержки ФЗ при интеграции некоторого множества локальных БД, не всегда возможно получить U из-за несогласованности локальных БД. На основе введенных ранее модифицированных операций определим понятие слабоуниверсального отношения для несогласованных БД.

Пусть d – БД со схемой $R = \{R_1, \dots, R_n\}$, причем $U = R_1 \cup R_2 \cup \dots \cup R_n$. Отношение s со схемой U будем называть слабоуниверсальным отношением для d , если $\pi_{R_i}(s) \supseteq r(R_i)$ для каждой схемы $R_i \in R$.

Слабоуниверсальное отношение является результатом применения модифицированной операции естественного соединения $\succ\langle^U$. Такие отношения необходимы для поддержки глобального выполнения зависимостей в БД.

Пусть дана база $d(R)$ и множество атрибутов U , а также множество ФЗ над U . База d глобально удовлетворяет множеству ФЗ, если для него существует слабоуниверсальное отношение $s(U)$, и проекции $\pi_{R_i}(s)$ удовлетворяют данному множеству ФЗ. Здесь под операцией проекции понимается модифицированная операция проекции.

Для базы d и множества ФЗ над U отношение $s(U)$ будем называть F -слабоуниверсальным отношением (в обозначении $F-U$) для d , если s есть слабоуниверсальное отношение для d , удовлетворяющее множеству ФЗ.

Утверждение: База d имеет $F-U$ если d глобально удовлетворяет множеству ФЗ.

Справедливость утверждения следует из определения F -слабоуниверсального отношения.

В случае, когда атрибуты одной схемы являются подмножеством атрибутов другой схемы, условие вхождения значительно сократит временные затраты на поддержку глобального выполнения ФЗ.

Если множество атрибутов X – ключ отношения R_i соответствует некоторому множеству атрибутов отношения R_j , то в любой момент времени каждое значение X , существующее в R_j , должно также существовать в R_i . Такое ограничение будем называть ограничением на чужой ключ. Трудность поддержки такого ограничения заключается в поиске отношений, которые содержат атрибуты из множества X .

Введем дополнительное требование для ограничения на чужой ключ. Пусть $d(R)$ – БД со схемой R и K – ключ в R . Пусть F – множество ФЗ, которые полностью определяются ключами из R . БД d удовлетворяет модифицированному ограничению на чужой ключ (МОЧК) относительно F , если для каждого $R_i \in R$ и

$t \in r(R_i)$ найдется такая строка t' в R_i^+ , что $\forall R_j \subseteq R_i^+ \exists t'(R) \in r(R_j)$, и, в частности, $t'(R_i) = t$. Если в каждом отношении БД поддерживаются свои ключи и удовлетворяется ограничение на МОЧК относительно полностью определяемого этими ключами множества ФЗ, то БД глобально удовлетворяет ФЗ. Этот факт подтверждает следующее утверждение, доказательство которого можно найти в [3].

Утверждение: Пусть даны БД d и множество ФЗ F , полностью определяемое ключами из d . Если d удовлетворяет МОЧК относительно F и каждое отношение удовлетворяет своим ключам, то d глобально удовлетворяет F .

Соблюдение ограничения МОЧК все же не исключает просмотра нескольких отношений при обновлении одного отношения, так как замыкание атрибутов одного отношения может включать почти все отношения БД. Таким образом, процесс поддержки ФЗ сводится к рассмотрению слабоуниверсального отношения.

Синтаксические условия на схему слабоуниверсального отношения. Для слабоуниверсальных отношений рассмотрим один класс схем, а именно – класс ациклических схем, обладающих рядом свойств, необходимых для обеспечения глобальной целостности. Важным свойством при проверке целостности данных для этих схем является возможность обойтись без U . Определим эти свойства. Пусть $R = \{R_1, R_2, \dots, R_p\}$ – схема БД и $d = \{r_1, r_2, \dots, r_p\}$ – экземпляр БД. Говорят, что d согласована в целом (СЦ), если r_1, r_2, \dots, r_p обладают свойством полноты соединения. Базу d будем называть попарно-согласованной (ПС), если свойством полноты соединения обладает произвольная пара отношений r_i и r_j . Сложность проверки свойства ПС полиномиально зависит от количества отношений в БД. При этом СЦ с необходимостью влечет за собой ПС, но ПС не всегда достаточно для СЦ [3].

ПС влечет СЦ (ПС \rightarrow СЦ) для БД со схемой $R' = \{r_1'(ABC), r_2'(BCD), r_3'(CDF)\}$. Покажем, что каждый кортеж из r_2 входит в $r_1 \times r_2 \times r_3$. Пусть t_2 – кортеж из r_2 . Поскольку r_1 и r_2 обладают свойством полноты соединения, то r_1 содержит кортеж t_1 , соединимый с t_2 . Аналогично, r_3 содержит кортеж t_3 , также соединимый с t_2 . Эти три кортежа соединимы, поскольку $t_1(C) = t_2(C) = t_3(C)$ и C – атрибут, общий для t_1 и t_3 . С другой стороны, для БД над схемой $R'' = \{r_1''(ABC), r_2''(BCD), r_3''(CE), r_4''(DE)\}$ свойство ПС не достаточно для СЦ.

Видно, что свойство ПС \rightarrow СЦ зависит от того, каким образом атрибуты отношений пересекаются между собой. Используя синтаксис схемы, используем алгоритм, проверяющий, выполняется ли свойство ПС \rightarrow СЦ. Алгоритм редукции Грэхема [3] состоит в последовательном выполнении шагов, на каждом из которых к схеме БД применяется одно из двух правил редукции – до тех пор, пока не окажется, что ни одно из них применить больше нельзя. Правила редукции состоят в следующем:

- (УС) удаление схемы отношения; если схема отношения собственным образом содержится в некоторой другой схеме отношения;
- (УА) удаление атрибута; атрибут принадлежит не более чем одной схеме отношения.

Говорят, что алгоритм Грэхема завершился успешно, если в результате его применения к атрибутам схем БД получается пустое множество. В противном случае формируется набор атрибутов, образующих связный блок, то есть когда каждый атрибут принадлежит более чем одной схеме, и ни одна схема не входит собственным образом в другую. В дальнейшем такие схемы будем называть циклическими, а схемы, для которых алгоритм Грэхема завершается успешно – ациклическими. Таким образом, для ациклических схем ПС влечет СЦ, а для циклических схем это свойство не выполняется. Для успешного завершения алгоритма Грэхема необходимо, чтобы схема была ациклическая, в противном случае ее необходимо привести к ациклическому виду.

Пусть $d = \{r_i, \dots, r_p\}$ – БД со схемой $R = \{R_1, \dots, R_p\}$, и пусть $d' = \{r_i, \dots, r_j\}$ – связной блок со схемой $R' = \{R_i, \dots, R_j\}$, тогда отношение $b = r_i \times \dots \times r_j$ со схемой $B = R_i \dots R_j$, в обозначении $b(B)$, будем называть сильно-связным блоком для $d(R)$. Замена в циклической схеме связных блоков на сильно-связные приведет ее к ациклическому виду. Это следует из того, что все схемы связного блока будут собственным образом входить в схему сильно-связного блока, и, таким образом, алгоритм Грэхема завершится успешно. Ациклические схемы с сильно связным блоком будем называть блочно-ациклическими.

Пусть база d имеет схему $R = \{AB, BCD, CDE, DEK, EKM, KMN\}$. В результате применения алгоритма Грэхема остается связный блок со схемой $R' = \{BCD, CDE, DEK\}$. Естественное соединение отношений со схемой R' порождает сильно-связный блок $B = \{BCDEK\}$. Таким образом, преобразованная исходная схема будет иметь вид $R_B = \{AB, BCDEK, EKM, KMN\}$ и является ациклической.

Итогом полученных результатов является следующее утверждение.

Утверждение: Для того чтобы циклическая схема удовлетворяла свойству ПС \rightarrow СЦ, необходимо и достаточно заменить все связные блоки сильно-связными. Справедливость утверждения следует из редуктивного алгоритма Грэхема и определения сильно-связного блока.

Выводы. Проанализировав операционную спецификацию реляционной модели данных можно отметить, что при использовании неопределенности в качестве допустимого значения реляционного отношения не позволяет поддерживать целостность традиционными методами. Таким образом, научной новизной статьи является модификация некоторых операций реляционной алгебры для их применения к отношениям БД с неопределенными значениями. С другой стороны, для повышения эффективности

поддержки целостности при интеграции данных в статье рассматривается модификация схемы БД, позволяющая использовать попарное соединение для проверки согласованности в целом.

Практическая значимость полученных результатов состоит в возможности использовать несогласованные отношения локальных БД в интегрированной среде, обеспечивая однозначность представления данных. При этом приведение циклической схемы к блочно-ациклической значительно снижает время проверки целостности интегрированной информации.

Рассмотренные вопросы определяют дальнейшие направления исследований в области поддержки данных в интегрированной среде обработки информации, направленные на разработку прикладных средств, расширяющих традиционные возможности СУБД и систем управления распределенными БД.

Список литературы: 1. Цикритзис Д., Лоховски Ф. Модели данных. – М.: Финансы и статистика, 1985. – 344 с. 2. Цаленко М.Ш. Семантические и математические модели баз данных. – М.: ВИНТИ, 1985. – 208 с. 3. Мейер Д. Теория реляционных баз данных. – М.: Мир, 1987. – 608 с. 4. Гарсиа-Молина Г., Ульман Дж., Уидом Дж. Системы баз данных. Полный курс. – М.: Издательский дом "Вильямс", 2003. – 1088 с. 5. Буслик М.М. Моделі і структури даних: Навч. посібник – Харків: ХТУРЕ, 2000. – 136 с. 6. Codd E. A relational model of data for large shared data banks // SACM 13. – 1970. – № .6. – P. 1958–1982. 7. Буслик Н.Н. Глобальные схемы реляционных баз данных: концепция и методы построения. Монография. – Харьков: ХНУРЭ, 2002. – 68 с. 8. Ульман Дж. Основы систем баз данных. – М.: Финансы и статистика, 1983. – 334 с.

Поступила в редакцию 31.03.2006