УДК 004.652:004.658

МОДЕЛИРОВАНИЕ ВЕРСИОННЫХ СТРУКТУР С ПОМОЩЬЮ АНАЛИТИЧЕСКОЙ МОДЕЛИ СХЕМЫ ДАННЫХ

Трофимов Б.Ф., Малахов Е.В.
Одесский национальный политехнический университет

Все больше современных технологий позволяют сохранять последовательную промежуточную историю изменения данных. К числу таких относятся транзакционные СУБД, системы контроля версий, текстовые процессоры, менеджеры файловых систем. Однако современные модели представления данных (реляционная, иерархическая и др.) не предназначены для учета их версионности. С этой целью разработана аналитическая модель схемы данных, особенностью которой является ее функциональное представление схемы данных. В рамках модели определены абстрактная модель схемы данных, ее версии, операции над версиями.

Для задания модели введены следующие обозначения и определения:

1. Пусть дано некоторое выражение v над элементами и переменными из X, множество параметров-переменных выражения v обозначено через $|v|$.
2. Функция интерпретирования I_{AC} — функция отображения элементов из некоторого множества A в множество C. I_{AA} — тождественная функция из A в A.

Определение 1. Пусть даны множества A_i, G_i такие, что пары (A_i, G_i) образуют алгебры, при i=1..n. Пусть $I_{Ai\text{-}Aj}$ - функция интерпретирования из A_i в A_j, тогда аналитической моделью схемы данных (далее схемой) называется конечное множество кортежей вида (рис. 1):

$$s=\{ <a_1,v_1>,...,<a_n,v_n>, ...,<a_k, f_1(..., f_2(..., f_n(p_1,...,p_m), ...),...)>,....\},$$

где $f_i \in \{I_{Ai-Aj}\} \cup G_j$, a_i – идентификаторы кортежей, такие что выполнено условие уникальности: $\neg\exists a_i, a_j \in s : a_i = a_j$ при j≠i;

$v_i \in A_1 \cup \ldots \cup A_N$; p_i – идентификаторы, для которых выполнено условие замкнутости: $\forall p_i \exists$ кортеж из s с идентификатором pi..

Определение 2. Классом схем называется множество всех потенциально построенных схем над зафиксированными параметрами $A_1...A_n$, $G_1...G_n$, I_{Ai-Aj} и обозначено через:

$$MS(\{A_1...A_n\},\{G_1...G_n\},\{I_{Ai-Aj}\}) \quad (1)$$

Определение 3. Пусть дано некоторое множество *A,* тогда элементарной схемой называется конечное множество кортежей вида $s=\{(a_1,v_1),(a_2,v_2),...,(a_n,v_n)\}$, где $v_i \in A$, a_i — идентификаторы кортежей с выполнением условия уникальности $\neg\exists a_i, a_j \in s : a_i = a_j$, при $i \neq j$.

Определение 4. Классом элементарных *схем* называется множество всех потенциально построенных элементарных схем над параметром *A* и обозначено через *S(A);* по построению $S(A_n) \subseteq MS(\{A_1,\ldots,A_N\},\{G_1,\ldots,G_N\},\{I_{Ai-Aj}\})$.

Для схемы из (1) предложен метод ее вычисления в элементарную схему, заданный в виде функции интерпретирования $I_{P,MS\to S}$. Для схемы из (1) предложен метод ее вычисления в элементарную схему, заданный в виде функции интерпретирования $I_{P,MS\to S}$.

Определение 5. Версией элемента *a* из схемы *s* называется элемент $a' \in s$, у которого $a \in a'$. Пусть даны $a_1,\ldots,a_n \in S$, такие что a_{i+1} есть версия a_i, тогда схему s можно представить как содержащую всю историю изменения элемента a_1:

$$s=\{<a_1,a>,<a_2,f_1(a_1)>,...,<a_i,f_{i-1}(a_1,...,a_{i-1})>,...,<a_n, f_{n-1}(a_1,...,a_{n-1})>\}.$$

Для оперирования элементами классов схем и элементарных схем определены множества операций. Операции над классом схем *(F)* включают: операцию добавления элемента в схему ($+$), операцию удаления элемента из схемы ($-$), операцию переименования элемента (RENAME), операцию изменения выражения в элементе схемы (CHANGE). Операции над классом элементарных схем *(H)* включают теоретико-множественные

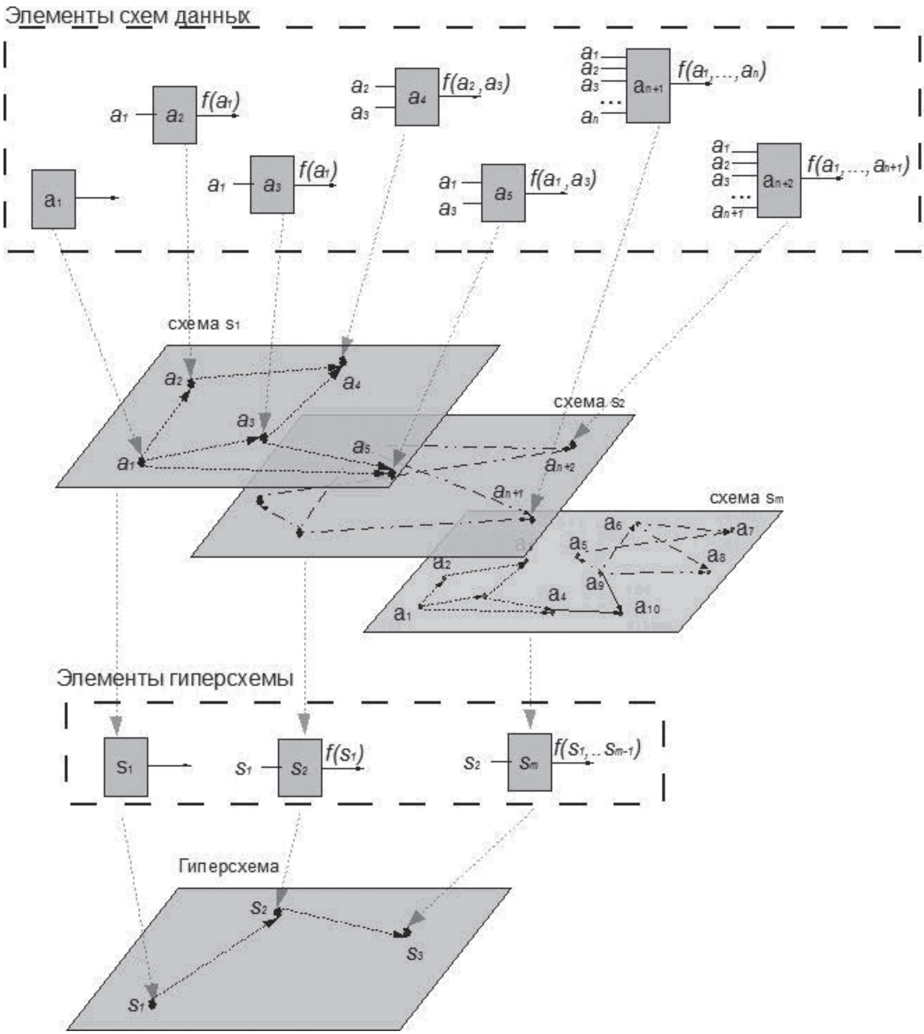

Рисунок 1 - Структурное представление и взаимосвязь элементов схемы, схемы и гиперсхемы

операции объединения ($\cup$), пересечения ($\cap$), вычитания ($\backslash$), а также операцию выборки элементов из элементарной схемы по некоторому предикату (GET).

Для представления версионности схем предложен класс схем на основе (1) с определенными параметрами:

$$MS(\{MS', S'\}, \{F,H\},\{I_{ms'-s'},I_{s'-ms'}\}), \quad (2)$$

где $MS'=MS(\{A_1...A_n\},\{G_1...G_n\},\{I_{Ai-Aj}\})$, $S'=S(A_n)$.

Определение 6. Элементы множества (2) называются гиперсхемами (рис. 2), а (2) – классом гиперсхем.

В качестве абстракции управления версиями схемы из (1) в составе гиперсхем из (2) предложены следующие сценарии:

1. Создание версии схемы a в гиперсхеме s_1: s_1=ADD(s_1, a)
2. Замена версии схемы с идентификатором *id* на новую a в гиперсхеме s_1: $s_1=CHANGE(s_1,id, a)$
3. Удаление версии схемы с идентификатором *id* из гиперсхемы s_1 : s_1=DEL(s_1,id)
4. Модификация данных в схеме с идентификатором id_{i-1} в гиперсхеме s: добавление, редактирование, удаление эле-

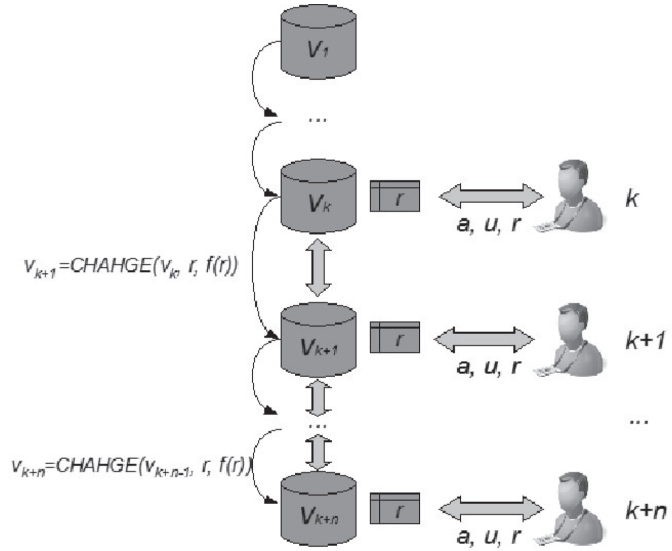

Рисунок 2 Представление реляционного отношения r из версий $v_k,...,v_{k+n}$

ментов схемы. Любая модификация существующих схем в составе гиперсхемы приводит к созданию новой версии схемы: $CHANGE'(s,id_{i-1}, f(id_{i-1})>)=ADD(ADD(s ,<\_id_i, f(\_id_{i-1}) >),<id_i, I_{P,MS-S}(\_id_i) >)$

5. Сокращение версии, упрощение гиперсхемы.

Определение 7. Пусть даны гиперсхемы *s* и *s'* из (2) для некоторых зафиксированных $\{A_1...A_n\},\{G_1...G_n\},\{I_{ai-Aj}\}$. Гиперсхема *s* является эквивалентной *s'*, если $I_{MS\to S}(s)= (s')$.

Определение 8. Высотой гиперсхемы *s* называется длина максимальной последовательности $a_1,...,a_n$ из *s*, где a_{i+1} является версией a_i. Рассмотрен подкласс гиперсхем с высотой равным 2 (*двухуровневые гиперсхемы*).

Показано, что для любой произвольной гиперсхемы существует эквивалентная ей двухуровневая гиперсхема.

Определение 9. Схема *s называется корректной,* если функция интерпретирования $I_{MS\to S}$ определена на *s (*существует $I_{MS\to S}(s)$*)*. Схема *s* называется *слабо-корректной,* если функция интерпретирования $I_{MS\to S}$ вычислима на *s (*для алгоритма вычисления $I_{MS\to S}$ на *s* всегда существует останов, однако не обязательно существует $I_{MS\to S}(s)$*)*.

Предлагается классификация ошибок, нарушающих корректность схем из класса (1):

- ошибки 1 рода: для схемы *s* существует последовательность идентификаторов $id_1,..id_n$, таких, что id_i является версией $id_{((n+i-2) \bmod n)+1}$;
- ошибки 2 рода: для схемы s $\exists id_i \in s$ такой что при известных параметрах из $|id_i|$ выражение id_i не вычислимо;
- ошибки 3 рода: для схемы *s* $\exists id_i \in s$ такой что при известных параметрах из $|id_i|$ выражение id_i вычислимо, но не определено.

Доказаны два критерия взаимосвязи слабо-корректных схем и ошибок 1-3 рода.

Критерий 1.

Пусть дано подмножество схем *S* из класса схем (1) с зафиксированными параметрами $\{A_1...A_n\},\{G_1...G_n\},\{I_{Ai-Aj}\}$, у которого любой $s \in S$ не содержит ошибок 2 рода.

Схема $s \in S$ слабо-корректна тогда и только тогда, когда не содержит ошибок 1 рода.

Следствие 1.

Слабо-корректная схема $s \in S$ может содержать ошибки 3 рода.

Следствие 2.

Схема $s \in S$ корректна тогда и только тогда, когда она слабо-корректна и не содержит ошибок 3 рода.

Предложена методика построения орграфа для любой схемы из (1). Доказан критерий связи орциклов в орграфе и ошибок 1 рода.

Критерий 2.

Схема из (1) имеет ошибку 1 рода тогда и только тогда когда соответствующий ей ограф имеет орцикл.

На основе определения 9, критерия 1, его следствий и критерия 2 построен метод определения корректности схемы, который основан на сведении к задаче определения орцикла в орграфе.

Для определения орцикла в орграфе предложена модификация метода временного моделирования. В качестве узловой функции выбрана функция конъюктивности входящих в узел сигналов: $F(x_i) = \bigwedge_{i=1}^{n} f(e_{ij})$. Показано, что отсутствие орциклов в орграфе связано с истинностью $F = \bigwedge_{j=1}^{k} F(x_j)$, где x_j – вершины, чье множество исходящих ребер пусто.

На основе аналитической модели схемы данных и ее методов была разработана модель многоверсионного представления реляционной БД, которая позволяет моделировать БД с несколькими версиями структуры. Для этого введены следующие обозначения:

1. SIG — множество всех потенциальных сигнатур реляционных отношений (РО);

2. R — множество всех потенциальных РО над SIG; R_0 — подмножество R, у которых множество кортежей пусто;
3. $I_{SIG\text{-}R}: SIG \rightarrow R$ — функция интерпретирования по сигнатуре из SIG в элемент из R_0.
4. O – множество операций реляционной алгебры $\{\pi, \sigma, \gamma, x, \alpha\}$.

Предлагается расширение набора реляционных операций O над R:

$$O' = \{a, u, r, a_s, u_s, r_s, c_s, m_s\}, \tag{3}$$

где $a, u, r, a_s, u_s, r_s, c_s, m_s$ – операции соответственно: добавления кортежа, изменение кортежа, удаление кортежа, добавление поля в сигнатуру РО, изменение домена поля в сигнатуре РО, удаление поля из сигнатуры РО, копирование поля сигнатуры из одного РО в другое, перенос поля сигнатуры из одного РО в другое.

Версию структуры БД предлагается представить в виде схемы данных из (1) с параметрами $R, SIG, O, O', I_{SIG\text{-}R}$.:

$$MS' = MS\left(\{R, SIG\}, \{O \cup O'\}, \{I_{SIG-R}\}\right). \tag{4}$$

С учетом (4-6) модель БД с поддержкой многоверсионной структуры БД (ММП) предлагается представить в виде гиперсхемы из класса гиперсхем (2):

$$MS'' = MS\left(\{MS', S'\}, \{F, H\}, \{I_{ms'-s'}, I_{s'-ms'}\}\right), \tag{5}$$

где $S' = S(R)$.

Показано, что модель ММП является расширением традиционной модели БД.

В следствие многоверсионного определения структуры БД произвольное реляционное отношение r может быть по разному представлено в нескольких версиях БД $v_k, \ldots, v_{k+n}$ (рис. 2). Для управления представлением кортежей между несколькими версиями БД разработаны соответствующие методы представления.

В качестве абстракции управления многоверсионной БД определены сценарии (выражены через сценарии управления схемой данных):

- чтение данных кортежей реляционных отношений;
- модификация кортежей реляционных отношений;
- модификация сигнатур реляционных отношений;
- модификация версии БД (добавление, удаление нового реляционного отношения);
- работа с версиями БД (сокращение, упрощение, приведение к двухуровнему виду).

Разработанные модели и методы управления позволили построить информационную технологию многоверсионного представления и управления БД для повышения эффективности модернизации информационных систем.

Литература

[1] Ю.Л. Ершов. Математическая логика [текст] / Ю.Л. Ершов, Е.А. Палютин — М.:Наука, 1987.

[2] К. Дж. Дейт. [текст] Введение в системы баз данных — 8-е изд. — М.: «Вильямс», 2006. — С. 1328.

[3] Малахов Е.В. Обзор модели расширения реляционных схем [текст] / Малахов Е.В., Трофимов Б.Ф // Восточно-европейский журнал передовых технологий. — Харьков: 3(33) — , 2008. — С. 35.

[4] Catharine M. Wyss Relational Languages for Metadata Integration [текст] / Catharine M. Wyss, Edward L. Robertson // ACM Transactions on Database Systems — June 2005 — Vol. 30, No. 2 — C. 624–660.