УДК 004.652/.942

Г.Н. ЖОЛТКЕВИЧ, АХМАД ЮСЕФ ИБРАХИМ ИБРАХИМ

Харьковский национальный университет им. В.Н. Каразина, Украина

ОРГАНИЗАЦИЯ ХРАНИЛИЩА ОБРАЗЦОВ ПОНЯТИЙ СХЕМЫ ДЛЯ ПРЕДСТАВЛЕНИЯ СЛОЖНЫХ СТРУКТУРИРОВАННЫХ ДАННЫХ

В работе рассмотрено множества образцов понятий полусхемы, являющихся структурно-логическими моделями информационных объектов. Введен класс операторов на этих множествах, соответствующих процедурам доступа к структурированным данным и их компонентам. Показано, что замкнутость конечного множества образцов относительно этих операторов является исчерпывающим ограничением целостности корректных состояний информации в информационной системе. Построена реляционная модель хранилища образцов. Описано расширение реляционной алгебры, обеспечивающее выполнение CRUD операций и операций доступа к данным над хранилищем образцов.

информационная система, концептуальная модель, реляционная модель данных, полусхема, образец, реляционная алгебра

Введение

Проектирование логической структуры данных является одним из ключевых этапов создания информационной системы. Одной из основных проблем является обеспечение адекватности информационных моделей объектов предметной области, с одной стороны, и технологичность информационных образов, с другой. Широко используемая в настоящее время реляционная модель данных, обладая высокой технологичностью, достаточно плохо приспособлена для представления сложных структурированных объектов [1, 2]. Усложнение информации, необходимость обеспечения ее семантической интерпретации независимо от специфики их обработки приводит к необходимости хранения и обработки в информационных системах данных со сложной структурой. Существует ряд подходов, направленных на решение проблемы самоинтерпретации информации: онтологии, формальный концептуальный анализ и т.п. Однако эти подходы либо не обладают необходимой для использования в реальных информационных системах технологичностью, либо не обеспечивают решение проблемы проверки корректности данных, что является необходимым условием обеспечения требуемого качества информации [3].

В ряде работ, которые являются результатом исследований, ведущихся в Харьковском национальном университете имени В.Н. Каразина, разрабатывается алгебраический подход к структурно-логическому моделированию предметных областей информационных систем. В основе его лежит введенная авторами категория полусхем, как метамодель данных информационной системы [4, 5]. В работе [6] показано, что всякую полусхему можно представить реляционной базой данных с двумя отношениями, а в работе [7] обоснована возможность проверки корректности представления средствами реляционного исчисления. Таким образом, предложенная модель дает возможность не только корректно описывать метаданные информационной системы, но и представлять и обрабатывать их технологически эффективными методами – средствами реляционной модели данных.

Заметим, что проблема представления данных, соответствующих описанным метаданным, образцов понятий в терминах цитируемых работ оставалась вне внимания авторов. В связи с этим настоящая статья направлена на восполнение отмеченного пробела, что дает возможность использование инвариантной структуры хранения сложной структури-

рованной информации средствами реляционных баз данных.

Целями настоящей работы являются:

– описание структуры множества возможных образцов понятий предметной области информационной системы – информационных объектов;

– выявление ограничений целостности, обеспечивающих корректность представления состояния информации в системе;

– построение реляционной модели хранилища образцов;

– описание расширения реляционного исчисления, обеспечивающего выполнение CRUD-операций и операций доступа к данным над хранилищем образцов.

1. Обзор теории полусхем и постановка задачи

Так же, как и в работах [4 – 7], будем использовать следующие обозначения.

Для пары множеств X, Y обозначим:

$M_+(X,Y)$ – множество частичных, хотя бы где-то определенных, отображений из X в Y;

$dom(f)$ – область определения отображения $f \in M_+(X,Y)$;

$im(f)$ – область значения отображения $f \in M_+(X,Y)$;

xR – подмножество $\{y \in Y \mid (x,y) \in R\}$ для бинарного отношения R между множествами X и Y;

X^* – множество конечных последовательностей элементов множества X, включающее пустую последовательность, которая всегда обозначается символом ε;

$h(x_1 \ldots x_k) = x_1$ – для любой непустой последовательности элементов множества X;

$\mathrm{t}(\mathrm{x}_1\mathrm{x}_2 \ldots \mathrm{x}_\mathrm{k}) = \mathrm{x}_2 \ldots \mathrm{x}_\mathrm{k}$ –для любой непустой последовательности элементов множества X.

Полусхемой предметной области назовем тройку $S = (\mathbf{N}, \mathbf{R}, \mathbf{D})$, где $\mathbf{N}$, $\mathbf{R}$ – конечные множества, $\mathbf{D} \subset \mathbf{N} \times M_+(\mathbf{R}, \mathbf{N})$, для которой выполняется условие:

для $n \in \mathbf{N}$, $f, g \in M_+(\mathbf{R}, \mathbf{N})$, $r \in \mathbf{R}$ таких, что $(n,f) \in \mathbf{D}$, $(n,g) \in \mathbf{D}$ и $r \in dom(f) \cap dom(g)$, верно $f(r) = g(r)$.

Для полусхемы $S = (\mathbf{N}, \mathbf{R}, \mathbf{D})$ понятие $n \in \mathbf{N}$ называется **базовым**, если для всякого $f \in M_+(\mathbf{R}, \mathbf{N})$ выполняется $(n,f) \notin \mathbf{D}$. Множество базовых понятий полусхемы будем обозначать $\mathbf{N}_0$.

Пусть $S = (\mathbf{N}, \mathbf{R}, \mathbf{D})$ является полусхемой и для некоторого $n \in \mathbf{N}$ существует $f \in M_+(\mathbf{R}, \mathbf{N})$ такое, что $(n,f) \in \mathbf{D}$. Тогда будем говорить, что для понятия n задан **вариант определения** f.

Теорию полусхем можно рассматривать как метамодель для структурно-логического моделирования предметных областей информационных систем, в том числе и для использования с целью формализации структуры отчетной информации.

В работе [6] построена реляционная модель, представляющая полусхему: всякая полусхема может быть задана реляционной базой данных с двумя отношениями, которые будут обозначены

$$\mathbf{DFC}(NNM, DNM) \text{ и}$$

$$\mathbf{DEF}(DNM, RNM, IRF).$$

Атрибуты в этих отношениях – NNM, DNM, RNM, IRF – являются именами понятий, вариантов определения, ролей и значений варианта определения на роли соответственно.

Содержательно (см. [6]), $(n,f) \in \mathbf{DFC}$ означает, что для понятия n задан **вариант определения** f, а $(f,r,n) \in \mathbf{DEF}$, что $r \in dom(f)$ и $f(r) = n$.

Поскольку далеко не всякая база данных из двух отношений с заданными схемами представля-

ет какую-либо полусхему, в работе [6] найден критерий того, что база данных есть базой данных некоторой полусхемы: для того чтобы пара отношений r_{DFC} и r_{DEF} со схемами **DFC** и **DEF** соответственно представляли некоторую полусхему необходимо и достаточно выполнения трех следующих условий:

$DNM, RNM \to IRF$ в отношении r_{DEF}; (1)

$(\forall f, g \in r_{DEF})$ (2)

$\left(\pi_{RNM,IRF}\left(\sigma_{DNM=f}\left(r_{DEF}\right)\right) = \pi_{RNM,IRF}\left(\sigma_{DNM=g}\left(r_{DEF}\right)\right) \Rightarrow f = g\right)$;

$NNM, RNM \to IRF$ в отношении $r_{DFC} \triangleright\!\triangleleft r_{DEF}$. (3)

В работе [7] получены способы проверки корректности базы данных полусхемы (условия 1 – 3) средствами реляционной алгебры.

Для того, чтобы от интенсиональной модели предметной области, каковой является полусхема, перейти к экстенсиональной модели в работе [4] определены образцы понятий полусхемы. Определение носит рекурсивный характер.

Элемент (n, w) множества $N \times R^*$ называется терминальной именующей нитью понятия n, если выполнено одно из следующих двух условий 1 или 2:

1. $w = \varepsilon$;
2. $w = r_1 r_2 \ldots r_k$ и найдется последовательность пар $(n_i, f_i) \in D$, где $i = 1, \ldots, k$, причем
   2.1. $n_1 = n$;
   2.2. $r_i \in \mathrm{dom}(f_i), f_i(r_i) = n_{i+1}, i = 1, \ldots, k-1$, а $f_k(r_k) \in \mathbf{N}_0$;
   2.3. $r_k \in \mathrm{dom}(f_k)$.

Для полусхемы $S = (\mathbf{N}, \mathbf{R}, \mathbf{D})$ и понятия $n \in \mathbf{N}$ его образцом называется $p = \{(n, w_i) \mid i = 1, \ldots, Q\}$ – конечное множество терминальных именующих нитей n, обладающее следующими свойствами:

1. если $n \in N_0$, то $p = \{(n, e)\}$;
2. если $n \notin N_0$, то существует $f \in nD$ такое, что:
   2.1. $dom(f) = \{h(w) \mid (n, w) \in p\}$;
   2.2. $p = \bigcup_{r \in \mathrm{dom}(f)} p_r$, где
   $p_r = \{(n, w) \in p \mid h(w) = r\}$;
   2.3. для $r \in \mathrm{dom}(f)$
   $\{(f(r), t(w)) \mid (n, w) \in p_r\}$ является образцом понятия $f(r)$.

Образцы понятия являются структурными моделями информационных объектов, соответствующих понятию.

В связи с этим в работе [4] отмечено, что корректность структурно-логической модели предметной области эквивалентна наличию у всех понятий полусхемы образцов и приведен алгоритм соответствующей проверки.

Полусхема, для которой каждое понятие имеет хотя бы один образец, называется **схемой**.

В дальнейшем мы ограничимся рассмотрением только схем.

Как уже отмечено выше, образцы понятия являются структурными моделями однотипных информационных объектов, поэтому желательно иметь для них более наглядную реализацию. Такая реализация была получена авторами в [8] в виде маркированного ориентированного дерева. Этот результат подтверждает на уровне модели тезис о том, что всякий сложный структурированный объект конструируется из своих компонентов посредством абстракции агрегации.

2. Хранилище образцов понятий схемы

Множество всех возможных образцов схемы $\mathbf{S} = (\mathbf{N}, \mathbf{R}, \mathbf{D})$ обозначим через $\mathbf{P}(\mathbf{S})$. Тогда состояниям информационной базы системы (состояниям

хранилища образцов понятий) будут соответствовать конечные подмножества $\mathbf{P}(\mathbf{S})$.

Введем на множестве $\mathbf{P}(\mathbf{S})$ семейство частичных операторов $\{\mathbf{A}_r \mid r \in \mathbf{N}\}$. Действие оператора $\mathbf{A}_r$ на образце $p = \{(n, w_i) \mid i = 1, \ldots, Q\}$ определяется следующим образом:

1. если $p = \{n\}$, где $n \in \mathbf{N}_0$, то $\mathbf{A}_r p$ не определено;
2. если $f \in M_+(\mathbf{R}, \mathbf{N})$ и
   $dom(f) = \{h(w) \mid (n, w) \in p\}$, причем
   $r \notin dom(f)$, то $\mathbf{A}_r p$ не определено;
3. если $f \in M_+(\mathbf{R}, \mathbf{N})$ и
   $dom(f) = \{h(w) \mid (n, w) \in p\}$, причем
   $r \in dom(f)$, то
   $\mathbf{A}_r p = \{(f(r), t(w)) \mid (n, w) \in p \wedge h(w) = r\}$.

Из определения образца и определения семейства операторов $\{\mathbf{A}_r \mid r \in \mathbf{N}\}$ следует, что множество $\mathbf{P}(\mathbf{S})$ замкнуто относительно операций семейства, а сами операторы представляют собой операторы доступа к компонентам образца. Таким образом, мгновенному описанию состояния информационной базы предметной области, которое, очевидно, может содержать лишь конечное число образцов, соответствует конечное подмножество $\mathbf{P}(\mathbf{S})$, инвариантное (замкнутое) относительно семейства $\{\mathbf{A}_r \mid r \in \mathbf{N}\}$.

Тем самым, инвариантность мгновенного описания относительно операторов доступа к компонентам является **ограничением целостности хранилища образцов понятий схемы**.

3. Представление хранилища образцов схемы средствами реляционной модели данных

Реляционная модель хранилища образцов схемы представляет собой БД, состоящую из отношений:

уровень метаданных представляют отношения, определенные в разделе 1 $\mathbf{DFC}(NNM, DNM)$ и $\mathbf{DEF}(DNM, RNM, IRF)$;

уровень данных задает отношение

$$\mathbf{DAT}(PID, NNM, DNM, RNM, CRF),$$

атрибуты которого: PID – идентификатор образца; NNM – имя понятия, которому соответствует образец; DNM – имя варианта определения, в соответствии с которым строится образец; RNM – имя роли, определяющей компонент образца; и CRF – ссылка на этот компонент.

Приведенная схема базы данных обосновывается полученным в разделе 2 ограничением целостности хранилища образцов понятий системы, которое можно реализовать за счет сохранения записей об образце независимо от того, являются они компонентами других образцов или нет.

Поскольку отношения **DFC** и **DEF** соответствуют полусхеме, то для них должны выполняться условия 1 – 3, приведенные в разделе 1.

Сформулируем условия, которые необходимо наложить на отношение **DAT** и условия связи всех трех отношений для того, чтобы база данных представляла корректное состояние информации.

Во-первых, каждый образец соответствует только одному понятию и задается только одним вариантом определения.

Это условие выражается парой функциональных зависимостей: $PID \to NNM$ и $PID \to DNM$, что эквивалентно в силу аксиом Армстронга [2] одной зависимости:

$$PID \to NNM, DNM. \qquad (4)$$

Во-вторых, должна сохраняться однозначность определения компонента по роли в рамках варианта определения:

$$DNM, RNM \to CRF. \qquad (5)$$

В-третьих, должен выполняться аналог условия 3 для **DAT**:

$$PID, NNM, RNM \to CRF. \qquad (6)$$

Кроме того, необходимое множество образцов должно быть согласовано с метаданными, хранящимися в отношениях со схемами **DFC** и **DEF**.

Пусть r_{DFC} отношение хранилища со схемой **DFC**, r_{DEF} – отношение со схемой **DEF** и r_{DAT} – отношение со схемой **DAT**.

Во-первых, образцы понятий должны соответствовать введенным понятиям и их вариантам определения:

$$\pi_{NNM,DNM}\left(r_{DAT}\right) \subset r_{DFC}. \qquad (7)$$

Во-вторых, определения понятий должны быть согласованы с метаданными:

$$\left(\forall s \in r_{DAT}\right) \qquad (8)$$

$$\pi_{RNM}\left(\pi_{PID,DNM,RNM}\left(\langle s\rangle\right) \triangleright\triangleleft r_{DAT}\right) =$$

$$= \pi_{RNM}\left(\pi_{DNM,RNM}\left(\langle s\rangle\right) \triangleright\triangleleft r_{DEF}\right)$$

$$\pi_{DNM,RNM,IRF}\left(\qquad (9)\right.$$

$$\delta_{PID\leftarrow IRF}\left(\pi_{DNM,RNM,IRF}\left(\langle s\rangle\right)\right) \triangleright\triangleleft$$

$$\triangleright\triangleleft \pi_{PID,NNM}\left(\delta_{CRF\leftarrow NNM}\left(r_{DAT}\right)\right)\Big) \subset$$

$$\subset r_{DEF}$$

Имеет место следующая **теорема**:

Реляционная база данных, состоящая из трех отношений **DFC**, **DEF** и **DAT** со схемами, определенными выше, является корректным хранилищем образцов понятий тогда и только тогда, когда выполнены условия 1 – 9, причем понятия определены схемой, соответствующей базе данных с отношениями **DFC** и **DEF**.

4. CRUD операции и операции доступа к данным хранилища образцов схемы

В этом разделе приведены алгоритмы доступа к данным, создания и удаления образцов. Все алгоритмы приведены на псевдокоде, расширенном реляционной алгеброй.

Алгоритм локализации образца предназначен для построения отношения

$$\mathbf{SMP}\left(PID, NNM, DNM, RNM, CRF\right),$$

содержащего выборку одного образца.

```
global warehouse; // хранилище образцов
proc select _ sample( pid )
// pid - идентификатор выбираемого образца
   global warehouse;
   local count : integer, work : relation(PID);
   warehouse.create(
     SMP(PID, NNM, DNM, RNM, CRF));
   last _ count := 0;
   work := ⟨pid⟩;
   do
      count := |warehouse.SMP|;
      warehouse.SMP := warehouse.SMP ∪
        work ▷◁ warehouse.DAT;
      work :=
        δ_{PID←CRF}(π_{CRF}(warehouse.SMP));
   while |warehouse.SMP| ≠ count end do;
   return;
end proc;
```

Алгоритм создания образца предназначен для вставки нового образца заданного понятия, определенного по заданной схеме из заданных компонентов.

```
global warehouse; // хранилище образцов
proc insert _ sample( pid, nnm, dnm,
  comps : relation(RNM, CRF))
// pid,nnm, dnm - идентификаторы образца,
//  понятия и варианта определения
// comps - отношение содержащее роли и
//  ссылки на образцы, соответствующие ролям
   if |⟨pid⟩ ▷◁ warehouse.DAT| > 0 then
      return "Error : Sample alredy exists";
   end if;
   if |⟨nnm, dnm⟩ ▷◁ warehouse.DFC| = 0 then
      return "Error : Illegal type or definition case";
   end if;
   if π_{RNM,IRF}(⟨dnm⟩ ▷◁ warehouse.DFC) ≠
     δ_{IRF←NNM}(π_{RNM,NNM}(
       δ_{PID←CRF}(comps) ▷◁ π_{PID,NNM}(
         warehouse.DAT))) then
      return "Error : Illegal sample structure";
   end if;
   warehouse.DAT := warehouse.DAT ∪
     ⟨pid, nnm, dnm⟩ ▷◁ comps;
end proc;
```

Алгоритм удаления образца обеспечивает удаление записи, соответствующей образцу понятия из хранилища образцов, а также записей прямо или косвенно на нее ссылающихся.

Его сигнатура имеет вид:

global *warehouse*; *// хранилище образцов*
proc *delete _ sample*(id)
// pid - идентификатор образца

Этот алгоритм близок к алгоритму локализации образца, поэтому его текст приводиться здесь не будет. Однако отметим, что основным источником итерации в алгоритме является необходимость выбора всех образцов, которые ссылаются на исходный прямо или опосредствованно.

Выводы

Таким образом, в настоящей статье впервые

– предложен подход к построению хранилища сложных структурированных информационных объектов, базирующийся на теории полусхем (раздел 2);

– сформулированы ограничения целостности, обеспечивающие корректность представления состояния информационной базы системы (раздел 2);

– построена реляционная модель хранилища, для которой ограничения целостности выражены на языке реляционной модели данных (раздел 3);

– описано расширение языка реляционной алгебры для реализации CRUD операций и операций доступа к данным над хранилищем.

Литература

1. Дейт К.Дж. Введение в системы баз данных, 2-е изд. – М.: Мир, 1982. – 846 с.

2. Мейер Д. Теория реляционных баз данных. – М.: Мир, 1987. – 608 с.

3. Eppler M.J. Managing Information Quality. – Berlin: Springer, 2003. – 302 p.

4. Жолткевич Г.Н., Семенова Т.В. К проблеме формализации концептуального моделирования информационных систем // Вісник Харк. нац. ун-ту ім. В.Н. Каразіна. – Х., 2003. – № 605: Математичне моделювання. Інформаційні технології. Автоматизовані системи управління. – Вип. 2. – С. 33-42.

5. Семенова Т.В. Морфизмы полусхем и их приложения // Вісник Харк. нац. ун-ту ім. В.Н. Каразіна. – Х., 2005. – № 703: Математичне моделювання. Інформаційні технології. Автоматизовані системи управління. – Вип. 4. – С. 198-206.

6. Жолткевич Г.Н., Семенова Т.В., Федорченко К.А. Представление полусхем предметных областей информационных систем средствами реляционных баз данных. – Вісник Харк. нац. ун-ту ім. В.Н. Каразіна. – Х., 2004. – № 629: Математичне моделювання. Інформаційні технології. Автоматизовані системи управління. – Вип. 3. – С. 11-24.

7. Жолткевич Г.Н., Федорченко К.А. Проверка корректности спецификации концептуальной модели предметной области средствами реляционной алгебры // Вестник Херсонского нац. техн. ун-та. - 2005. – № 22. – С. 138-142.

8. Жолткевич Г.Н., Ибрахим Ахмад Юсеф Ибрахим. О возможности представления образцов понятий полусхем маркированными деревьями // Системи обробки інформації. – Х.: ХУ ПС, 2006. – Вип. 2 (51). – С. 20-26.

Поступила в редакцию 28.04.2006

Рецензент: д-р техн. наук, проф. А.Ю Соколов, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.