Вісник Харківського національного університету
Серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»
№ 605, 2003, с. 33-42

УДК 004.652/.942

К проблеме формализации концептуального моделирования информационных систем

Г.Н. Жолткевич, Т.В. Семенова
Харьковский национальный университет им. В.Н. Каразина, Украина

In the present paper the problem of the information system modelling is considered. The approach to the data modelling using the algebraic structures was made. As a result the mathematical data model has been built. The model obtained consists two levels, these are structural and semantic ones. The mechanism of the correctness verification of the data model was developed.

Характерный для современного общества процесс массового внедрения информационных технологий, базирующихся на обработке информации при помощи вычислительной техники, приводит к росту спроса на информационные системы различного назначения. Такие системы должны соответствовать ряду требований, среди которых:

адекватность – достоверное отражение в информационной системе фактов и закономерностей предметной области;

открытость – обеспечение возможности реструктуризации информационной базы системы в связи с развитием знаний о ее предметной области;

гибкость – обеспечение возможности реализации различных, удобных с точки зрения разных категорий пользователей, представлений фактов и закономерностей предметной области;

эффективность – достижение оптимального для заказчика соотношения между ценой разработки информационной системы и ожидаемым доходом от ее использования.

Применяемые в настоящее время при разработке информационных систем технологические решения в той или иной мере ориентированы на реализацию этих требований. Они базируются на использовании инструментальных средств, позволяющих построить различные модельные взгляды на проектируемую систему. Интеграция этих моделей обеспечивает достаточно адекватное описание разрабатываемой системы. Ярким примером этого подхода являются технологические решения, базирующиеся на языке UML [1-2].

Формализация средств представления модельных взглядов на проектируемую систему открывает путь к использованию в процессе разработки CASE (computer aided software engineering) – средств, например Rational Rose. Применение существующих CASE-средств позволяет значительно повысить эффективность процессов разработки программного обеспечения. Однако указанные средства обеспечивают прежде всего, интеграцию проектных данных, фиксацию текущей проектной ситуации в репозитории проекта, но не поддерживают механизмы принятия проектных решений. Кроме того, реализация в информационной системе пп. 2 и 3 из перечисленных выше требований представляется весьма сомнительной с точки зрения указанных

подходов. Дальнейшее же повышение эффективности разработки информационных систем связано как раз с автоматизацией принятия решений в процессе проектирования.

Таким образом, одной из ключевых проблем при создании CASE-средств, ориентированных на поддержку процесса принятия проектных решений, является проблема разработки строгих математических методов моделирования достаточно широкого класса предметных областей. В настоящее время существует формализм моделирования в терминах реляционной модели данных [4-6]. Однако, как было отмечено авторами настоящей работы в [7], реляционная модель данных неадекватна широкому спектру информационных систем, а именно, информационным системам, ориентированным на поддержку научных исследований. Такое положение стимулировало настоящую работу, результатом которой стала описанная ниже модель.

Пусть N – конечное множество, элементы которого соответствуют именам понятий предметной области, R – конечное множество, элементы которого соответствуют именам ролей, т.е. ссылкам внутри экземпляра понятия на его структурные части.

Обозначим через $M(R,N)$ множество частичных отображений из R в N. Для областей определения отображений $f \in M(R,N)$ будем использовать обозначение $dom(f)$. Будем также обозначать через ε отображение с пустой областью определения.

Кроме того, будем использовать стандартные обозначения R^* – для множества всех слов над R, R^+ – для множества непустых слов над R, e – для пустого слова.

Определение 1.

Полусхемой предметной области назовем тройку $S = (N,R,D)$, где N, R – конечные множества, $D \subset N \times M(R,N)$, для которой выполняются следующие условия:

1. если $(n,\varepsilon) \in D$, где $n \in N$, то $\{f \in M(R,N) \mid (n,f) \in D\} = \{\varepsilon\}$;
2. если для $n \in N$, $f,g \in M(R,N)$, $r \in R$ одновременно выполнено
   2.1. $(n,f) \in D$,
   2.2. $(n,g) \in D$,
   2.3. $r \in dom(f) \cap dom(g)$,
   то $f(r) = g(r)$.

Условие 2 определения позволяет корректно задать частичное отображение $\tau : N \times R \to N$ следующим образом: $\tau(n,r)$ определено в том и только том

случае, если существует $f \in M(R,N)$, для которого $(n,f) \in D$, $r \in dom(f)$ при этом $\tau(n,r) = f(r)$.

Определение 2.

Именующей нитью понятия n назовем элемент (n,w) из множества $N \times R^*$, который удовлетворяет одному из следующих условий:

1. $w = e$ и $(n,\varepsilon) \in D$;
2. для $w = r_1 r_2 \ldots r_k$ в последовательности $n_0 = n$, $n_i = \tau(n_{i-1}, r_i)$, где $i = 1,\ldots,k$, все члены определены.

Множество именующих нитей понятия n будет обозначаться через $T(n)$. Именующие нити обладают следующим очевидным свойством.

Утверждение 1.

Если $t = (n, r_1 \ldots r_k)$ – именующая нить, то для всякого $1 \le m < k$ $t' = (n, r_1 \ldots r_m)$ также является именующей нитью.

Для именующей нити $t = (n,w)$ длина слова w будет называться длиной именующей нити и обозначаться $|t|$. Можно рассмотреть расширенный вариант частичного отображения $\tau : N \times R^* \to N$, которое определено ниже.

Определение 3.

Пусть $t \in T(n)$, тогда

1. если $|t| = 0$, то $\tau(t) = n$;
2. если $|t| = k$, то $\tau(t) = n_k$, где $n_0 = n$, $n_i = \tau(n_{i-1}, r_i)$, $i = 1,\ldots,k$, а $t = (n, r_1 \ldots r_k)$.

Определение 4.

Понятие $n \in N$ такое, что $(n,\varepsilon) \in D$ будем называть **базовым понятием**. Обозначим через N_0 множество базовых понятий полусхемы S.

Определение 5.

Пусть $n \in N$, $t \in T(n)$. Если $\tau(t) \in N_0$, то t будем называть **терминальной именующей нитью**.

Обозначим через $T_0(n)$ множество терминальных нитей понятия n, а через $F(T_0(n))$ множество всех конечных подмножеств множества $T_0(n)$.

Ключевым в предлагаемой модели является следующее

Определение 6.

Образцом понятия $n \in N$ называется конечное множество p терминальных именующих нитей этого понятия, удовлетворяющее следующему условию: для всякой именующей нити $t = (n, w) \in p$ и для всякого представления w в виде $w_1 r w_2$, где $w_1, w_2 \in R^*$, $r \in R$, найдется единственное отображение $f \in M(R, N)$, для которого $(\tau(n, w_1), f) \in D$, $r \in dom(f)$, и для всех $r' \in dom(f)$ в p найдется нить вида $(n, w_1 r' v_{r'})$ для некоторого $v_{r'} \in R^*$.

Обозначим через $P(n)$ множество образцов понятия $n \in N$, а через $P = \bigcup_{n \in N} P(n)$ – множество всех образцов понятий из N.

Определение 7.

Будем говорить, что образец $p \in P(n)$ **имеет сигнатуру** $f \in M(R, N)$, если выполнены следующие два условия:

1. $(n, f) \in D$
2. $p = \bigcup_{r \in dom(f)} \left\{ t \in p \,\middle|\, \exists (w \in R^*)\, t = (n, rw) \right\}$

Определение 8.

Полусхема $S = (N, R, D)$ называется **схемой**, если для любых $n \in N$ и $f \in M(R, N)$, для которых выполнено $(n, f) \in D$, можно построить хотя бы один образец $p \in P(n)$ с сигнатурой f.

В общем случае не для каждого $n \in N$ может быть построен образец (будем называть такие понятия **плохо определенными**). В связи с этим возникает следующая задача: для заданной полусхемы найти необходимые и достаточные условия существования образцов. Решение этой задачи требует введения ряда вспомогательных конструкций.

Пусть задана полусхема $S = (N, R, D)$. Определим последовательность подмножеств множества N следующим образом:

положим $N^{(0)} = N_0$, и для $k > 0$, определим

$$N^{(k)} = N^{(k-1)} \cup \left\{ n \in N \setminus N^{(k-1)} \mid \exists (f \in M(R, N)) \mid (n, f) \in D \wedge im\, f \subset N^{(k-1)} \right\}.$$

Очевидно, что $\left\{ N^{(k)} \mid k \geq 0 \right\}$ является возрастающей по вложению последовательностью подмножеств множества N. В силу конечности

последнего она, очевидно, стабилизируется, т.е. найдется такое m, что $N^{(m)} = N^{(m+1)} = \ldots$. Обозначим через $N_\infty = \bigcup_{k\geq 0} N^{(k)}$.

Теорема 1.

Множество N_∞ состоит из тех и только понятий, которые имеют хотя бы один образец.

Доказательство.

Докажем сначала, что любое понятие из N_∞ имеет хотя бы один образец.

Пусть $n \in N_\infty$, тогда обозначим через $\mathrm{m}(n) = \min\left(k \mid n \in N^{(k)}\right)$.

Доказательство проведем индукцией по $\mathrm{m}(n)$.

Если $\mathrm{m}(n) = 0$, то утверждение очевидно.

Предположим, что утверждение верно для случая $\mathrm{m}(n) < k$, и покажем, что оно верно для $\mathrm{m}(n) = k$.

Действительно, в случае $\mathrm{m}(n) = k$ для n выполняется следующее соотношение

$$\exists\left(f \in M(R,N)\right)\left|(n,f) \in D \wedge im\, f \subset N^{(k-1)}\right.$$

Используя это отображение f, построим образец p для такого n. Выберем $r \in dom(f)$. В силу предположения индукции $f(r)$ имеет хотя бы один образец. Пусть $p_r = \left\{t_1, t_2, ..., t_{s(r)}\right\}$ – любой образец $f(r)$, где $t_i = \left(f(r), w_i\right)$. Построим множество $p^{(r)} = \left\{(n, rw_i) \mid i = 1, \ldots, s(r)\right\}$. Определим теперь $p = \bigcup_{r\in dom(f)} p^{(r)}$. Очевидно, что p - образец n.

Докажем обратное утверждение: любое понятие из N, для которого можно построить образец, принадлежит множеству N_∞, т.е. $\mathrm{m}(n) < \infty$.

Пусть $n \in N$ и существует образец p для n.

Так как p – конечное множество терминальных именующих нитей, то в этом множестве существует нить $t = (n, w)$ максимальной длины. Обозначим соответствующую длину через m. Очевидно, что в силу построения последовательности $\left\{N^{(k)} \mid k \geq 0\right\}$ выполняется неравенство $\mathrm{m}(n) \leq m$, что и завершает доказательство теоремы.

Теорема 2.

Пусть задана полусхема $S=(R,N,D)$, тогда тройка $S'=(R',N',D')$, где $N'=N_\infty$, $D'=D\setminus\{(n,f)\in D \mid f\in F\}$, $R'=\bigcup\limits_{f\in M(R,N)\setminus F} dom(f)$, где F определено соотношением: $F=\{f\in M(R,N)\mid \exists(r\in dom\, f) f(r)\in N\setminus N_\infty\}$, является схемой.

Доказательство.

Заметим, что определение D' может быть переписано следующим образом:

$$D'=\{(n,f)\in D \mid im\, f\subset N_\infty\}.$$

Пусть $n\in N'$, тогда в силу построения N' выполнено

$$\exists(f\in M(R,N))\,|\,(n,f)\in D\wedge im\, f\subset N_\infty,$$

а значит может быть построен хотя бы один образец n с сигнатурой f. Так как это верно для любого отображения f такого, что $(n,f)\in D'$ в силу определения D', то S' является схемой. Доказательство завершено.

Теорема 3.

Пусть задана полусхема $S=(R,N,D)$ и по ней построена схема $S'=(R',N',D')$, тогда множества образцов у S и S' совпадают.

Доказательство.

Включение $P'\subset P$ очевидно.

Докажем обратное включение, используя метод доказательства от противного.

Предположим, что $\exists(n\in N)\exists(p\in P(n)) p\notin P'(n)$. В силу определения S' такое возможно только в одном из двух случаев:

1. $n\notin N_\infty$,
2. образец p имеет сигнатуру f, для которой $(n,f)\in D$ и $f\in F$.

Если $n\notin N_\infty$, то для понятия n нельзя построить ни одного образца в силу теоремы 2, что противоречит определению схемы.

Если образец p имеет сигнатуру f такую, что $f\in F$, тогда $\exists(r\in dom(f)) f(r)\notin N_\infty$. Следовательно, для $f(r)$ не может быть построено ни одного образца, что влечет за собой невозможность построения образца и для n, что также противоречит определению схемы.

Таким образом, $P\subset P'$, что и доказывает теорему.

Определение 9.

Функцию $f \in M(R,N)$ будем называть **сигнатурой понятия** $n \in N$, если выполнено соотношение $(n,f) \in D$.

Определение 10.

Будем говорить, что два понятия являются **структурными синонимами**, если наборы их сигнатур совпадает.

Для большинства предметных областей характерно наличие отношения обобщения, которое в предлагаемой модели может быть введено следующим образом.

Определение 11.

Будем говорить, что понятие n **обобщает понятие** m, если $\forall (f \in M(R,N))\big((m,f) \in D \Rightarrow (n,f) \in D\big)$.

Обозначение: $n \supset m$.

Очевидно, что определенное таким образом отношение обобщения является рефлексивным и транзитивным.

Утверждение 1.

Если $n \supset m$ и $m \supset n$, то понятия m и n имеют одинаковый набор сигнатур, т.е. являются структурными синонимами. При этом отношение "быть структурными синонимами" является отношением эквивалентности.

Доказательство непосредственно следует из определений.

Отметим, что для построенной математической модели существует простой механизм проверки ее корректности.

В первую очередь речь идет об алгоритмах, определяющих, является ли заданная полусхема схемой, и выделяющих множества плохо определенных понятий в противном случае.

Алгоритм 1.

Вход: полусхема $S = (N,R,D)$.

Выход: множество плохо определенных понятий N_b.

{

$N^* = N_0$

do {

$N_{temp} = \left\{ n : N \middle| \exists \left(f : M\left(R,N\right)\right)\left(n,f\right) \in D \wedge \forall \left(r : domf\right) f\left(r\right) \in N^* \right\}$;

$N^* = N^* \cup N_{temp}$;

} while ($N^* != N_{temp}$);

$N_b = N \setminus N^*$;

}

Обозначения:

N – множество понятий предметной области;

N_b – множество плохо определенных понятий;

N_0 – множество базовых понятий;

N_{temp} – вспомогательное множество;

N^* – итерационно формируемое множество понятий, для которых можно построить хотя бы один образец.

Сходимость алгоритма обеспечивается за счет конечности множества N.

Если в результате работы алгоритма $N_b = \varnothing$, то исходная полусхема S является схемой.

В процессе определения схемы часть понятий могут находиться в отношении обобщения. Для нахождения таких понятий можно воспользоваться следующим алгоритмом.

Алгоритм 2.

Вход: понятия n и m.

Выход: сообщение о взаимосвязи понятий.

{

$F_1 = \{f : M(R,N) | (n,f) \in D\}$

$F_2 = \{f : M(R,N) | (m,f) \in D\}$

$F = F_1 \cap F_2$

if ($F = F_1$) { return "m обобщает n";}

else

if($F = F_2$) { return "n обобщает m";}

else { return "понятия не находятся в отношении обобщения";}

}

Обозначения:

F_1 - множество сигнатур понятия n;

F_2 - множество сигнатур понятия m;

F - множество сигнатур, принадлежащих и понятию n, и понятию m.

На этапе конструирования модели важно уметь выделять структурные синонимы. Для этого можно использовать изложенный выше алгоритм с небольшими изменениями.

Алгоритм 3.

Вход: два понятия n и m.

Выход: сообщение о наличии структурной синонимии понятий

$F_1 = \{ f : M(R,N) | (n,f) \in D \}$

$F_2 = \{ f : M(R,N) | (m,f) \in D \}$

$F = F_1 \cap F_2$

if ($F = F_1$)

if($F = F_2$) { return "n и m - структурные синонимы";}

else { return "m обобщает n";}

else

if($F = F_2$) { return "n обобщает m";}

else { return "понятия не находятся в отношении обобщения";}

Дальнейшее построение моделей данных для описания предметных областей может быть проведено достаточно стандартным способом.

Рассмотрим многоосновную алгебру $\mathbf{A}$ [8], сорта которой находятся во взаимно - однозначном соответствии с множеством N_0: $\mathbf{A} = \left(\{ \mathbf{A}_n | n \in N_0 \}, \Omega \right)$. Тогда для каждого образца $p = \{ t_1, t_2, ..., t_s \}$ можно построить экземпляр понятия как элемент множества $\prod_{i:1...s | \tau(n,w_i) \in N_0} \{ (w_i, v_i) \}$, где $v_i \in \mathbf{A}_{\tau(n,w_i)}$.

Таким образом, открывается возможность дальнейшего расслоения данных путем наложения условий на экземпляры в терминах многоосновной алгебры $\mathbf{A}$. Экземпляры, удовлетворяющие этим условиям, могут интерпретироваться как *примеры* понятия, а не удовлетворяющие – как *контрпримеры*.

Подводя итоги изложенному, отметим, что предложенная модель носит четырехуровневый характер (рис. 1), что позволяет достаточно хорошо отразить как структурные связи объектов предметной области, так и семантические ограничения целостности данных.

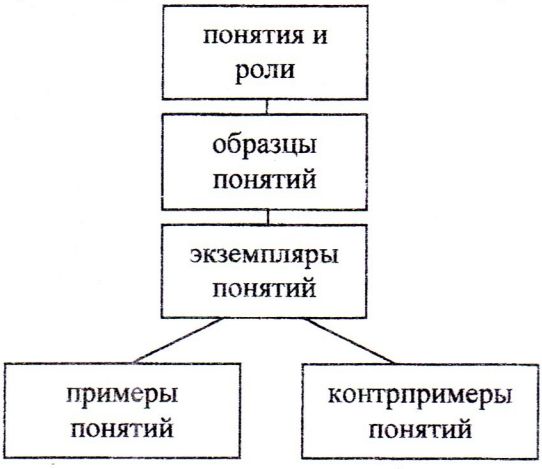

Рис. 1. Иерархия модели.

ЛИТЕРАТУРА

1. Мацяшек Л. А. Анализ требований и проектирование систем. Разработка информационных систем с использованием UML, - М.: Издательский дом «Вильямс», 2002. – 432 с.
2. Ларман К. Применение UML и шаблонов проектирования, - М.: Издательский дом «Вильямс», 2001. – 496 с.
3. Таундсен К., Фохт Д. Программирование и программная реализация экспертных систем на персональных ЭВМ. – М.: Финансы и статистика, 1990. – 320 с.
4. Майер Д. Теория реляционных баз данных, - М.: Мир, 1987. – 608 с.
5. Цикридзис Д., Лоховский Ф. Модели данных, - М.: Финансы и статистика, 1985. – 344 с.
6. Дейт К. Дж. Введение в системы баз данных, 6-е издание. – Москва – С. Петербург – Киев: Вильямс, 1999. – 846 с.
7. Жолткевич Г.Н., Семенова Т.В. Концептуальное моделирование данных в исследовательских информационных системах средствами реляционных СУБД. – Вестник Херсонского государственного технического университета. – 2002, №15. – С. 75-79.
8. Плоткин Б.И. Универсальная алгебра, алгебраическая логика и базы данных. – М.: Наука. Гл. ред. физ.-мат. лит, 1991. – 448 с.