УДК 004.652.942

О ВОЗМОЖНОСТИ ПРЕДСТАВЛЕНИЯ ОБРАЗЦОВ ПОНЯТИЙ ПОЛУСХЕМ МАРКИРОВАННЫМИ ДЕРЕВЬЯМИ

Г.Н. Жолткевич, Ахмад Юсеф Ибрахим Ибрахим
(Харьковский национальный университет им. В.Н. Каразина)

В работе предложена модель образца понятия полусхемы в терминах маркированного дерева. Предложенная модель проясняет структуру образцов понятий, возникающих при концептуальном моделировании предметных областей информационных систем средствами абстрактной алгебраической метамодели – теории полусхем.

качество информации, концептуальная модель, полусхема, дерево, образец

Введение. Проблема качества информации (QoI) является одной из острых проблем внедрения автоматизированных информационных систем в управление предприятиями. Обеспечение необходимого уровня QoI достигается за счет оптимизации различных факторов, рассмотренных, например, в работах [1 – 5]. Все авторы отмечают, что одним из подходов к решению проблемы повышения качества информации является использование методов интеллектуального анализа данных. Кроме того, обеспечение качества информации может осуществляться за счет формализации структуры документов и создания интеллектуальных программных средств поддержки оборота формализованных документов. Введение формализованных документов должно базироваться на строгих математических моделях, отражающих концептуальный (структурный) аспект предметной области. Авторами в настоящей работе в основу построения таких моделей положена общая алгебраическая метамодель (теория полусхем), впервые предложенная Г.Н. Жолткевичем и Т.В. Семеновой в работе [6]. Дальнейшее развитие этот аппарат получил благодаря работе [7], авторами которой удалось представить общую метамодель средствами реляционной алгебры и, тем самым, обеспечить ее технологическую направленность. В дальнейшем, в работе [8] были приведены алгоритмы, основанные на реляционном исчислении, позволяющие проверять корректность моделирования в терминах полусхем, что обеспечивает алгоритмический инструментарий концептуального моделирования предметных областей и обосновывает наш выбор в пользу предложенного аппарата. Следует отметить, что туманным местом теории полусхем является построение экстенсионалов понятий, которое достигается за

счет определения их образцов [6]. Однако математическая структура, которая моделирует образец понятия, используемая авторами теории полусхем, является достаточно формальной, а ее концепция – интуитивно не очевидной, что и побудило авторов настоящей работы прояснить эту концепцию.

Целью настоящей работы является построение моделей образцов понятий полусхем в терминах деревьев. Такое представление позволяет избежать излишнего формализма, которым отличается понятие образца, данное в оригинальной работе [6]. Кроме того, модель образца в терминах дерева близка к популярным концепциям DOM для XML-документов от W3C [9].

Обзор основных понятий теории полусхем. Также как и в работах [6 – 8] 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; $\mathrm{dom}(f)$ – область определения отображения $f \in M_+(X,Y)$; $\mathrm{im}(f)$ – область значения отображения $f \in M_+(X,Y)$; xR – подмножество $\{y \in Y \mid (x,y) \in R\}$ для бинарного отношения R между X и Y; X^* – множество конечных последовательностей элементов множества X, включающее пустую последовательность, которая всегда обозначается символом ε; $t(x_1x_2\ldots x_k) = x_2\ldots x_k$, $h(x_1\ldots x_k) = x_1$, – для любой непустой последовательности элементов 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 $r \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$, 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 структурного моделирования предметных областей информационных систем, в том числе и для использования с целью формализации структуры отчетной информации. Для того, чтобы от интенсиональной модели пред-

метной области, каковой является полусхема, перейти к экстенсиональной модели в работе [8] введено понятие образца понятия полусхемы. Понятие вводится в несколько этапов.

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

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) $\mathrm{im}(f) = h(p)$;

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 объектов, соответствующих понятию, поэтому желательно иметь для них более наглядную реализацию.

Функциональная модель графа. Поставленная цель работы достигается путем построения специальной, названной авторами функциональной, модели графа. Описанию этой модели посвящен настоящий раздел работы.

Пусть заданы два конечных непересекающихся множества $\mathbf{V}$ и $\mathbf{E}$. Ориентированным графом с множеством вершин $\mathbf{V}$ и множеством ребер $\mathbf{E}$ назовем такую четверку $G = (\mathbf{V}, \mathbf{E}, \mathrm{beg} : \mathbf{E} \to \mathbf{V}, \mathrm{end} : \mathbf{E} \to \mathbf{V})$, для которой

$$(\forall e_1, e_2 \in \mathbf{E} \mid \mathrm{beg}(e_1) = \mathrm{beg}(e_2) \wedge \mathrm{end}(e_1) = \mathrm{end}(e_2)) \Leftrightarrow (e_1 = e_2).$$

При этом для ребра $e \in \mathbf{E}$ вершину $\mathrm{beg}(e)$ будем называть началом ребра e, а вершину $\mathrm{end}(e)$ – концом этого ребра.

Очевидно, что с определенной нами структурой однозначно связано $\Gamma(G) \subset \mathbf{V} \times \mathbf{V}$ – бинарное отношение на множестве вершин графа, а именно

$$(v_1, v_2) \in \Gamma(G) \Leftrightarrow (\exists e \in \mathbf{E} \mid beg(e) = v_1 \wedge end(e) = v_2).$$

Обратно, если задано отношение $\Gamma \subset \mathbf{V} \times \mathbf{V}$, тогда мы можем рассмотреть множество вершин $\mathbf{V}$, множество Γ в качестве множества ребер и определить функции:

$$beg(v_1, v_2) = v_1; \quad end(v_1, v_2) = v_2.$$

Очевидно, что $(\mathbf{V}, \Gamma, beg : \Gamma \to \mathbf{V}, end : \Gamma \to \mathbf{V})$ является графом, а связанное с ним бинарное отношение совпадает с Γ.

Учитывая вышеизложенное, можно сделать вывод об взаимно однозначном соответствии между построенными нами структурами и ориентированными графами, определенными как бинарные отношения на множествах вершин.

Переформулируем в рамках введенного формализма основные определения теории графов.

Конечная последовательность $e_1, e_2, \ldots, e_{n-1}, e_n$ ребер графа $G = (\mathbf{V}, \mathbf{E}, beg : \mathbf{E} \to \mathbf{V}, end : \mathbf{E} \to \mathbf{V})$ называется путем в графе, если

$$(\forall i \mid 1 \le i \le n-1)\ end(e_i) = beg(e_{i+1}).$$

Для пути $e_1, e_2, \ldots, e_{n-1}, e_n$ вершина $beg(e_1)$ называется его началом, а $end(e_n)$ – концом.

Путь $e_1, e_2, \ldots, e_{n-1}, e_n$ в графе называется циклом, если его начало и конец совпадают, т. е. $beg(e_1) = end(e_n)$.

Для построения моделей образцов полусхем нам понадобиться ввести понятие дерева.

Под деревом мы будем понимать ориентированный граф $G = (\mathbf{V}, \mathbf{E}, beg : \mathbf{E} \to \mathbf{V}, end : \mathbf{E} \to \mathbf{V})$, обладающий следующими свойствами:

1) множество $\mathbf{V} - end(\mathbf{E})$ состоит из одного элемента, который мы обозначим через root и будем называть корнем дерева;

2) для каждого $v \in \mathbf{V}$, отличного от r, существует единственный путь, началом которого является r, а концом v.

Элементы множества $\mathbf{V} - beg(\mathbf{E})$ являются листьями дерева. Условия 1) и 2) определения гарантируют, что множество листьев дерева не пусто и, более того, каждый путь из корня дерева может быть продолжен до пути из корня в лист, по крайней мере, одним способом.

Введение функциональной модели дерева нам необходимо для того, чтобы формально определить понятия маркировки вершин и ребер дерева.

Пусть $T = (\mathbf{V}, \mathbf{E}, beg : \mathbf{E} \to \mathbf{V}, end : \mathbf{E} \to \mathbf{V})$ – дерево. Маркировкой вершин дерева T со значениями во множестве $M_{\mathbf{V}}$ назовем отображение из $\mathbf{V}$ в $M_{\mathbf{V}}$. Аналогично, маркировкой вершин дерева T со значениями во множестве $M_{\mathbf{E}}$ назовем отображение из $\mathbf{E}$ в $M_{\mathbf{E}}$. Таким образом, мы приходим к следующему определению.

Маркированным деревом называется восьмерка

$$T = (\mathbf{V}, \mathbf{E}, M_{\mathbf{V}}, M_{\mathbf{E}}, beg : \mathbf{E} \to \mathbf{V}, end : \mathbf{E} \to \mathbf{V},$$
$$m_{\mathbf{v}} : \mathbf{V} \to M_{\mathbf{V}}, m_{\mathbf{E}} : \mathbf{E} \to M_{\mathbf{E}}),$$

для которой $(\mathbf{V}, \mathbf{E}, beg : \mathbf{E} \to \mathbf{V}, end : \mathbf{E} \to \mathbf{V})$ является деревом.

Моделями образцов понятий полусхем, как будет видно из дальнейшего, являются специальным образом маркированные деревья.

Графовые модели образцов понятий полусхемы. Пусть задана полусхема $S = (\mathbf{N}, \mathbf{R}, \mathbf{D})$ и $\mathbf{N}_0$ – множество ее базовых понятий.

Маркированное дерево

$$T = (\mathbf{V}, \mathbf{E}, \mathbf{D} \cup \mathbf{N}_0, \mathbf{R}, beg : \mathbf{E} \to \mathbf{V}, end : \mathbf{E} \to \mathbf{V},$$
$$m_{\mathbf{V}} : \mathbf{V} \to \mathbf{D} \cup \mathbf{N}_0, m_{\mathbf{E}} : \mathbf{E} \to \mathbf{R}),$$

удовлетворяющее следующим условиям:

1) $m_{\mathbf{V}}(v) \in \mathbf{N}_0$ тогда и только тогда, когда v – лист дерева;

2) если $m_{\mathbf{V}}(v) = (n, f) \in \mathbf{D}$, то $m_{\mathbf{E}}$ устанавливает взаимно однозначное соответствие между $\{e \in \mathbf{E} \mid beg(e) = v\}$ и $dom(f)$;

3) если $v, v' \in \mathbf{V}, e \in \mathbf{E}, beg(e) = v, end(e) = v'$; $m_{\mathbf{V}}(v) = (n, f)$, а $m_{\mathbf{V}}(v') = (n', g)$ (для некоторого g) или $m_{\mathbf{V}}(v') = n'$ и $m_{\mathbf{E}}(e) = r$, то $f(r) = n'$, будем называть граф-образцом понятия n_t, где $m_{\mathbf{V}}(root) = (n_t, f_t)$ для некоторого f_t.

Для иллюстрации введенного определения приведем простой пример.

Пусть $\mathbf{N} = \{'list', 'data', 'null'\}$; $\mathbf{R} = \{'car', 'cdr', 'end'\}$. Определим пару частичных отображений из $\mathbf{R}$ в $\mathbf{N}$ следующим образом:

$$default \left| \frac{'car' \quad 'cdr'}{'data' \quad 'list'} \right.; \; empty \left| \frac{'end'}{'null'} \right..$$

И, наконец, зададим

$$\mathbf{D} = \{('\text{list}', \text{default}), ('\text{list}', \text{empty})\}.$$

Тройка $S = (\mathbf{N}, \mathbf{R}, \mathbf{D})$ является полусхемой, моделирующей понятие списка. Пример образца для этой полусхемы в виде маркированного дерева приведен на рис. 1.

Эквивалентность образцов и граф-образцов. Основной результат: образцы понятий полусхемы [6] и введенные в настоящей работе граф-образцы эквивалентны.

Легко видеть, что понятие в метке корня и метки дуг, прочитанные вдоль всякого пути в граф-образце от корня к листу, является терминальной именующей нитью, а рекуррентные условия определения образца соответствуют правилам разметки дерева граф-образца.

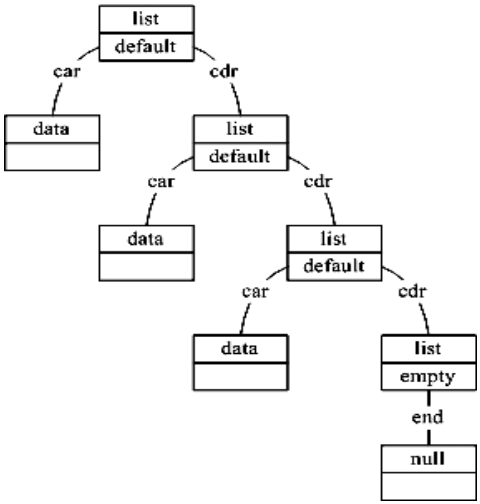

Рис. 1. Дерево одного из граф-образцов полусхемы примера

Выводы. Таким образом, в статье предложен способ моделирования образцов понятий полусхемы в терминах маркированных деревьев. Этот результат для теории полусхем получен впервые. Важность его заключается в том, что образцы понятий представляют собой структурные модели объектов, формирующих экстенсионалы понятий. Фактически модель подтверждает точку зрения, согласно которой задание структуры объекта производится за счет объективации отношений между объектом как целым и его составными частями. Интуитивно убедительным является моделирование такой ситуации иерархией, т. е. деревьями. Однако в теории полусхем образцы понятий моделировались при помощи громоздкой и интуитивно неубедительной формальной конструкцией, что порождало сомнения в адекватности теории. Основной результат этой статьи снимает возражения такого рода, демонстрируя, что в математическом плане формальная конструкция образца, определенная в [6], на самом деле эквивалентна иерархии. Этот результат получен впервые и имеет большое значение для дальнейшего развития теории полусхем как общей метамодели для концептуального моделирования предметных областей информационных систем.

Возникновение деревьев в качестве моделей образцов понятий полусхем позволяет надеяться на возможность построения для каждой полусхемы схемы XML-документов, объектные модели которых изоморфны образцам понятий полусхем. В этом случае причиной использования XML-подхода в информационных системах окажется не только его высокая технологичность, но и объективная необходимость описания структуры образцов информационных объектов.

ЛИТЕРАТУРА

1. *Rose F. The Economics, Concept, and Design of Information Intermediaries. – Berlin: Springer, 1999. – 266 p.*
2. *Аксенов Е. Качество информации: от очистки данных – к модели предприятия // PCWEEK. Корпоративные системы. – №36 (354). – 2002. – [Электр. ресурс]. – Режим доступа: http://www.pcweek.ru/Year2002/N36/CP1251/CorporationSystems/chapt1.htm.*
3. *Eppler M.J. Managing Information Quality. – Berlin: Springer, 2003. – 302 p.*
4. *Sy B.K., Gupta A.K. Information-Statistical Data Mining: Warehouse Integration with Examples of Oracle Basics. – The International Series in Engineering and Computer Science. – 2004. – Vol. 757. — 312 p.*
5. *Bacharach M., Board O. The quality of information in electronic groups. – Netnomics. – № 4. – 2002. – P. 73-97.*
6. *Жолткевич Г.Н., Семенова Т.В. К проблеме формализации концептуального моделирования информационных систем // Вісник Харк. нац. ун-та. Серія „Математичне моделювання. Інформаційні технології. Автоматизовані системи управління". – 2003. – № 605 (2).– С. 33-42.*
7. *Жолткевич Г.Н., Семенова Т.В., Федорченко К.А. Представление полусхем предметных областей информационных систем средствами реляционных баз данных // Вісник Харк. нац. ун-та. Серія „Математичне моделювання. Інформаційні технології. Автоматизовані системи управління". – 2004. – № 629 (3). –С. 11-24.*
8. *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.*
9. *Extensible Markup Language (XML) 1.0 (Second Edition) (W3C Recommendation). – [Электр. ресурс]. – Режим доступа: http:/ www.w3.org/TR/2000/REC-xml-20001006.*

Поступила 19.01.2006

Рецензент: доктор технических наук, профессор Л.Г. Раскин,
Национальный технический университет «ХПИ», Харьков.