УДК 681.3.06

Д.А. Руденко[1], В.В. Тулупов[2]

[1]*Харьковский национальный университет радиоэлектроники*
[2]*Харьковский национальный университет внутренних дел*

МОДЕЛЬ И СРЕДСТВА ПОДДЕРЖКИ ДАННЫХ В ЗАДАЧАХ ИНТЕГРАЦИИ НЕОДНОРОДНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Анализируются возможные помехи при совместной работе неоднородных информационных систем. Как источник порождения объектов предметной области рассматриваются задачи, которые решаются информационной системой. Основной акцент при построении модели делается на типы объектов, которые определяют возможность выводить скрытые объекты. Для этой цели используются правила, которые выражают семантику предметной области. В выводе предложена модель, содержащая выходные объекты, правила и семантику предметной области.

задачи интеграции неоднородных информационных систем

Введение

Постановка проблемы. Введение стандартов на организацию неоднородных сред управления распределенной информацией связано с преодолением проблем вызванных различным представлением данных. С одной стороны это связано с несоответствиями между представлениями или форматами данных, поступающих из разных источников. С другой стороны при взаимодействии множества источников данных часто возникает проблема нарушения семантической целостности. Различается также и сама семантика данных в поддерживаемых локальных базах данных (БД), и нужно найти способы не только для разрешения таких семантических конфликтов, но и для передачи семантической информации между БД.

Для исследований в этой области необходимо иметь возможность абстрактно рассматривать организацию управления информационными ресурсами. Абстракция любой системы представляет собой модель этой системы, в которой намеренно опущены некоторые детали. Выбор тех деталей, которые следует опустить, делается на основе рассмотрения, как заданного приложения, так и задач использующих эти абстракции.

Относительно легко решается задача абстрактного представления данных, если с самого начала информационная система проектируется и разрабатывается как открытая система, когда все компоненты являются мобильными и интероперабельными.

При интеграции неоднородных систем такая реализация является трудно достижимой, так как по разным причинам возникают потребности в обобщенном представлении независимых и по-разному организованных информационно - вычислительных ресурсов.

Практическое построение интегрированных систем неоднородных БД появилось в связи с необходимостью совместного использования систем БД, основанных на различных моделях данных и управляемых различными системами управления базами данных (СУБД).

Один из вариантов решения проблемы интеграции неоднородных БД является предоставление пользователям возможность видеть глобальную схему БД. Представление глобальной схемы обычно реализуется в некоторой модели данных, и поддерживает автоматическое преобразование операторов манипулирования данными глобального уровня к операторам, понятным соответствующим локальным СУБД. При такой интеграции неоднородных БД локальные системы утрачивают свою автономность.

Так как функционирование системы часто требует сохранения локальной автономности, при этом, обеспечивая возможность работы в интегрированной среде, то *актуальным* является развивающееся направление мультисистем БД или систем мультиБД для которых необходимо иметь возможность формально описывать область данных, в которых функционирует система, что является основным направлением статьи.

Анализ литературы. В системах мультиБД не поддерживается глобальная схема интегрированных БД, а применяются специальные способы именования для доступа к объектам локальных БД. Как правило, в таких системах на глобальном уровне допускается только выборка данных, что позволяет сохранить автономность локальных БД при совместном использовании данных.

Современные подходы к использованию неоднородных информационных ресурсов развивают идею их представления в виде набора типизированных объектов, сочетающих возможности сохранения информативности своего состояния и возможности обработки этой информации за счет наличия определенных методов, применимых к объектам.

Исследования в этой области ведутся с момента практического использования БД в распределенных и крупномасштабных системах. Оригинальные подходы были рассмотрены в [2], а также при построении испытательной распределенной базой данных [3]. Современные подходы к управлению распределенными ресурсами представлены в рабо-

тах Sheth A.P., Larson J.A. [4], Garcia-Solaso M., Saltor F., Castellanos M. [5], в которой рассматриваются задачи, возникающие при достижении локальной автономности. Среди переведенных источников можно выделить работу группы авторов [6], которая отличается широтой и глубиной охвата материала по вопросам проектирования и использования современных систем БД.

Цель статьи. Разработка БД в терминах традиционных моделей сводится к сложному процессу построения структуры данных, так как они не содержат достаточных средств для независимого представления семантики предметной области (ПрО) от модели данных. По этой причине необходимо найти такие средства построения семантической модели, которыми можно описать данные на уровне представлений в пределах ПрО, а не на уровне структуры данных.

В настоящее время используется метод семантического моделирования БД, который заключается в выделении двух уровней [7]: концептуальное моделирование ПрО; моделирование БД.

В первом случае осуществляется переход от неформального описания ПрО и информационных потребностей необходимых для решаемых задач к их формальному представлению. Во втором случае реализуется преобразование концептуальной модели в схему БД.

Таким образом, основной проблемой при решении задач моделирования БД является обобщенное представление ПрО. В статье вводится определение ПрО на основе множества задач решаемых в информационной системе и рассмотрен подход к концептуальному построению модели ПрО.

Основной *целью статьи* является рассмотреть средства описания ПрО и формально сформулировать определение ПрО. Понятие “предметной области” является базисным понятием в теории БД и поэтому не имеет строго определения. Но, тем не менее, для дальнейших выкладок понадобится определить смысл “предметной области”.

Средства описания предметной области

Под объектом предметной области будем понимать материальное и идеальное явления реального мира. Объекты потенциально обладают огромным количеством свойств и находятся в потенциально бесконечном числе взаимосвязей между собой. Однако среди всего множество свойств и взаимосвязей имеет смысл выделять только необходимые с точки зрения потребителя информации.

Объекты необходимые для функционирования информационной системы определяются набором задач.

Задачу будем рассматривать как набор из трех элементов: входные параметры X, выходные параметры Y и алгоритмы α_i ($i = 1 \div n$), приводящие состояние X в состояние Y. Формально задачу z будем определять как

$$z = <X, \alpha_i, Y> \quad (1)$$

при этом совокупность задач Z определим как

$$Z = \{z_1, z_2, \ldots, z_p\}. \quad (2)$$

Множество (2) описывает “границы ПрО”, так как Z определяет набор объектов содержащихся во входных параметрах X, а также необходимые условия α для представления абстрактной картины реального мира.

Для достижения универсальности представления ПрО необходима высокая абстрактность базисных объектов и правил порождения новых объектов, которые интерпретируются в любой ПрО.

Под моделью ПрО будем понимать средство, позволяющее интерпретировать объекты в соответствии с указанными требованиями. Понятие модели тесным образом связано с понятием абстракции. Абстракция какой-либо системы представляет собой модель этой системы, в которой намеренно опущены некоторые детали.

Для моделирования информационных объектов наиболее приемлемой абстракцией являются алгебраическая система, объединяющая в себе помимо множества объектов и отношений между ними, также множество операций, задаваемых на множестве объектов определяющих допустимые состояния ПрО, то есть состояние, не противоречащее задачам. При этом модель задается тремя элементами

$$M = <O, R, \Omega>, \quad (3)$$

где O – множество объектов; R – отношение между объектами; Ω – операционная спецификация.

Определение (3) представляет статическую модель, не отражая при этом динамические свойства ПрО, которые выражаются возможными изменениями множества O. Исходя из требований задач Z, объекты ПрО могут быть заданные или статические и порожденные или динамические.

К статическим объектам относятся объекты входящие в множество входных параметров X задачи Z и инвариантны во времени. Объекты, порожденные некоторыми правилами из множества статических объектов O, представляют динамический набор объектов ПрО. Любая модель ПрО должна каким-либо образом представлять эти два вида объектов.

Введем множество правил $L = \{l_i\}$. Правила будем задавать в виде импликации объектов

$$o_1, o_2, \ldots, o_n \leftarrow o_1, o_2, \ldots, o_m, \quad (4)$$

где o_i – объекты предметной области; символ “$\leftarrow$” читается как “если-то”.

Выражение (4) определяет то факт, что если набор объектов $o_1, o_2, \ldots, o_m$ входят в множество объектов ПрО, то и объекты $o_1, o_2, \ldots, o_n$ также могут входить в это множество.

Правила L позволяют выводить новые объекты, и тем самым задают расширенное множество S включающее как изначально заданные, так и выведенные объекты, то есть $O \subset S$.

Исходя из рассмотренных понятий и определений, возникает задача вывода множества S, причем объекты S должны выводиться при каждом измене-

нии в O. Для вывода элементов множества S воспользуемся алгоритмом, описанном в [8].

Общую схему вывода множества S можно представить в виде

$$\frac{\forall (o_1, o_2, ..., o_l \in O) \;\; \exists (o \leftarrow o_1, o_2, ..., o_l)}{o \in S}. \quad (5)$$

Схема (5) показывает, что если для множества объектов $o_1, o_2, \ldots, o_l \in O$ существует правило $o \leftarrow o_1, o_2, \ldots, o_l$, то S дополняется объектом o, то есть формируется расширенное множество S.

Будем предполагать, что каждый объект ПрО имеет некоторый функциональный тип τ (в дальнейшем для сокращения будем использовать термин "тип"). Функциональный тип не имеет пространственно – временной локализации и обеспечивает локальные точки зрения на объект различных потребителей информации или задач, то есть тип, и объект данного типа находятся в отношении "абстрактное - конкретное". Необходимость введения функционального типа определено представлением одного объекта в разных задачах по-разному.

Определив свойства типов и операции над ними можно говорить о возможности сведения различных ПрО к единому семантическому представлению, позволяющему сделать вывод о степени их похожести.

В каждом состоянии ПрО множеству объектов S сопоставим множество типов T_j для каждого объекта o_j $(i \neq j)$. Пусть $\wp = \{S_1, S_2, \ldots, S_n\}$ семейство множеств всех объектов соответствующее каждому состоянию ПрО и пусть $\Im = \{\tau_1, \tau_2, \ldots, \tau_n\}$ – семейство множеств типов соответствующих множеству $\wp$, причем $\tau_i = \{T_1, T_2, \ldots, T_m\}$ – множество типов каждого объекта $o_i \in S_i$, тогда между элементами $\Im$ существуют определенные теоретико-множественные отношения, например $\tau_i \subset \tau_j$ или $\tau_i \cap \tau_j = \varnothing$ и т.п. Объекты ПрО, как правило, делятся на семантически простые и семантически составные. Если в некотором правиле объекты стоящие слева то символа "$\leftarrow$" удовлетворяют правилу, то эти объекты могут являться логическим подмножеством семантически составного объекта стоящего справа от символа "$\leftarrow$".

Объект o будем называть семантически простым, если для некоторого множества объектов $\{o_1, o_2, \ldots, o_n\}$ выполняется условие (6):

$$\frac{\{o_i, o_2, \ldots, o_n\} := o}{\varnothing}. \quad (6)$$

Свойство объекта быть семантически простым не зависит от времени, но зависит от требований задач ПрО, так как один и тот же объект при различной интерпретации может иметь различные типы.

Объект o будем называть семантически составным, если в некоторый момент времени найдется такое множество объектов $\{o_1, o_2, \ldots, o_n\}$, для которого выполняться условие (7):

$$\frac{\{o_i, o_2, \ldots, o_n\} := o}{\text{Ttue}}. \quad (7)$$

Определим отношение принадлежности на множестве типов следующим образом: "Если $\tau_i := \tau_j$, то объект o типа τ_i может состоять из объектов типа τ_j" В общем случае отношение принадлежности может быть установлено между несколькими типами. Используя определение (8) введем отношение принадлежности для последовательности типов:

$$\frac{\tau_1 := \tau_2 := \ldots := \tau_n := \tau}{\text{Ttue}}. \quad (8)$$

Сформулируем основные свойства отношений частичного порядка и принадлежности.

A1: Рефлексивность. Если τ – тип объекта, то выполняется $\tau \le \tau$.

A2. Антисимметричность. Если выполняется $\tau_1 \le \tau_2$ и $\tau_2 \le \tau_1$, то выполняется $\tau_1 = \tau_2$.

A3: Транзитивность. Если выполняется $\tau_1 \le \tau_2$ и $\tau_2 \le \tau_3$, то выполняется $\tau_1 \le \tau_3$.

A4. Поглощение. Если выполняется $\tau_1 \le \tau_2$ и $\tau_2 := \tau_3$, то выполняется $\tau_1 := \tau_3$.

A5. Замкнутость. Если выполняется $\tau_1 := \tau_2$, $\tau_2 := \tau_3$, …, $\tau_{n-1} := \tau_n$, то выполняется $\tau_1 := \tau_n$ при условии $\tau_1 \cap \tau_n \neq \varnothing$.

Для конкретной ПрО могут быть сформулированы дополнительные свойства принадлежности, включающие все или только некоторые типы.

Исходя из свойства A5, можно говорить о композиции объектов, которые представляются конечными последовательностями других объектов. В качестве примера композиционного объекта можно рассмотреть дату, состоящую из года, месяца и числа, причем объект "дата" не является множеством из трех элементов год, месяц и число, и связан с элементами отношением, отличным от отношения принадлежности.

Для формального определения композиции введем функцию (9), отражающую в каждый момент времени *t* состояние композиционного объекта:

$$f_i^t : T(o)^t \rightarrow \{T_i(o_i)\}^t, \quad (9)$$

где T(o) – тип объекта o.

Таким образом, для каждого объекта справедливо выражение (10):

$$\forall\, o_i, o_j \in T \text{ и } f_1^t, \ldots, f_k^t,$$
$$\exists\, \{f_1^t(o_i), \ldots, f_k^t(o_i)\} \text{ и } \{f_1^t(o_j), \ldots, f_k^t(o_j)\}, \quad (10)$$

где $f_1^t(o_j), \ldots, f_k^t(o_j)$ – кортеж значений функций $f_1^t, \ldots, f_k^t$.

Конечное множество троек $V^T = \langle T, f_i, T_i \rangle$ при $i = 1 \div k$, $f_i \neq f_j$, $i \neq j$, будем называть представлением типа T, если выполняется функция (9), а условие (10) представляет собой условие различимости объектов.

Следует отметить, что множество V^T не может быть установлено задачами, и должно быть задано в дополнительной информации о ПрО.

Представление $V^T = \langle T, f \rangle$ будем называть тривиальной, то есть результатом применения функции f к объекту o типа T будет кортеж $\{f^t(o)\}$ состоящий из одного объекта o. Тип T называется композиционным, если для него существует нетривиальное представление V^T при этом типы $T_1, \ldots, T_k$ входящие в представление называются компонентами типа T. Представление $\{\langle T, f_1, T_1 \rangle, \ldots, \langle T, f_k, T_k \rangle\}$ называется минималь-

ным, если удаление из него любой тройки приводит к нарушению отображения (9). В любом представлении содержится, по крайней мере, одно минимальное представление, однако такое представление должно либо указываться явно, либо выводиться из заданных условий, выражающих закономерности ПрО.

Семантика предметной области

Для полного представления структуры ПрО, необходимо задать инвариантные свойства состояний и их последовательностей, то есть выразить семантику ПрО. Инвариантные свойства состояний ПрО будем называть семантикой ПрО, поскольку нарушение этих свойств приводит к неадекватности представления модели ПрО.

Частично такие свойства определяются исходными задачами, которые указывают соотношения между объектами и их множествами выполняемые в каждый момент времени. Кроме этого модель данных M, используемая для хранения и обработки исходных данных X, расширяют эти возможности, так как отношения между данными R отражают часть семантики ПрО.

Если определить множество L как выражения отражающие семантику объектов ПрО, то можно говорить о наборе правил отражающих адекватность состояния ПрО к требованиям задач. Это является тем фактом, что множество L определяет зависимости существования одних объектов от существования других, т.е. взаимосвязи между объектами, которые представляют состояние ПрО r^t.

Приведение ПрО к адекватному состоянию представим схемой вывода (11), при этом процедура определения нового объекта описывается выражением (4):

$$O \xrightarrow{L} S. \qquad (11)$$

Следует отметить, что при наличии множества объектов O и правил L расширенное множество S потенциально присутствует всегда, хотя явно может быть не задано (например, некоторое состояние r_i не требует наличия выводимых объектов) при описании ПрО. Главное предположение при таком подходе состоит в том, что множество объектов S и множество правил L должны быть совместны. Множество S совместно, если оно не содержит объектов, не выводимых по правилам L из множества O по схеме (11).

Выводы

Рассмотрев традиционное представление информационной модели, а так же проанализировав набор дополнительных требований к построению модели ПрО можно сделать вывод, что алгебраическая модель, состоящая из объектов, отношений и набором операций недостаточна для отражения семантики ПрО. Детализировать представление ПрО можно за счет описания не только объектов, но и их типов, что позволяет определить семантические свойства объект и установить смысловую принадлежность одних объектов к другим.

С другой стороны описание состояний ПрО как набор задач Z позволяет установить возможность появление одних объектов в зависимости от существования других объектов. Более подробное рассмотрение структуры правил (не представленное в данной статье) говорит о том, что определяемый набор объектов является семантически целостным и отсутствие хотя бы одного элемента не дает возможность получить новый объект.

Таким образом, *научной новизной* статьи является формальное описание модели ПрО, которая должна включать кроме элементов отражающих физические свойства ПрО элементы, отражающие семантические свойства, то есть модель представляет собой четверку вида (15):

$$M_{ПрО} = <Z, O(\tau), L, S>, \qquad (15)$$

где Z – множество задач; $O(\tau)$ – множество объектов с соответствующими типами; L – множество правил; S – расширенное множество объектов ПрО, допускается $S = \varnothing$.

Практическая значимость полученных результатов заключается в возможности интегрировать данные различной структуры и описанные различными моделями данных. При этом нет необходимости проектировать единую схему данных и поддерживать ограничения на глобальном уровне.

В заключение следует отметить, что в данной статье не рассматривались такие важные вопросы, как построение расширенного множества объектов *S*, которое является конечным представлением содержимого ПрО. Эта зада является *предметом дальнейшего исследования* при построении структуры ПрО. Также отдельной задачей являются вопросы целостности представления ПрО. Очевидно, что при сравнении предметных областей их семантика может быть различной, следовательно, необходимо рассмотреть вопрос определения степени похожести двух ПрО.

Таким образом, модель ПрО представленная в статье определяет дополнительные направления в разработке методов управления информационными системами различной структуры и детализации.

Список литературы

1. Интероперабельные информационные системы: архитектуры и технологии / Д.О. Брюхов, В.И. Задорожный, Л.А. Калиниченко, и др. // Системы управления базами данных. – 1995. – № 4. – С. 62-74.

2. Карденас А.Ф. Управление неоднородными распределенными базами данных // ТИИЭР. – 1987. – Т. 75, № 5. – С. 72-86.

3. Дуайер П.А., Ларсон Дж.А. Опыт работы с испытательной распределенной базой данных // ТИИЭР. – 1987. – Т. 75, № 5. – С. 126-138.

4. Sheth A.P., Larson J.A. Federated database for managing distributed, heterogeneous, and autonomous databases // Computing Surveys. – 22:3 (1990). – P. 183-236.

5. Garcia-Solaso M., Saltor F., Castellanos M. Semantic heterogeneity in maltidatabase system. // In Bukhres and Elmagarmid. – 1996. – P. 129-195.

6. Гарсиа-Молина Г., Ульман Дж., Уидом Дж. Системы баз данных. – М.: Вильямс, 2003. – 1088 с.

7. Львов В. Создание систем поддержки принятия решений на основе хранилищ данных // Системы управления базами данных. – 1996. – № 2. – С. 6-36.

8. Танянский С.С., Козырь О.Ф. Об одном алгоритме определения семантически неоднородных баз данных // Образование, наука, производство и управление в XXI веке: Сб. тр. Международной конференции. В 4-х т. – Старый Оскол: ООО «ТНТ». – 2004. – Т. 1. – С. 320-322.

Поступила в редколлегию 19.03.2007

Рецензент: д-р техн. наук, проф. И.В. Гребенник, Харьковский национальный университет радиоэлектроники, Харьков