СЕМАНТИЧЕСКАЯ СРАВНИМОСТЬ АТРИБУТОВ РЕЛЯЦИОННОЙ БАЗЫ ДАННЫХ ПРИ ПОСТРОЕНИИ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В КОРПОРАТИВНЫХ СИСТЕМАХ

УДК 519.863.5

КОЧУЕВА Зоя Анатольевна
к.т.н., доцент кафедры интеллектуальных компьютерных систем
Национального технического университета «Харьковский политехнический институт».
Научные интересы: автоматизированная обработка естественного языка,
информационно-поисковые системы, автоматизированные библиотечные системы.

БОРИСОВА Наталья Владимировна
к.т.н., доцент кафедры интеллектуальных компьютерных систем
Национального технического университета «Харьковский политехнический институт».
Научные интересы: автоматизированная обработка естественного языка, извлечение предметных знаний.

ПОСТАНОВКА ПРОБЛЕМЫ

С конца 70-х годов в теории баз данных (БД) получило развитие семантическое или концептуальное направление моделирования [1, 6, 7]. Причиной этому послужила необходимость интеллектуализации логической структуры БД, используемых в качестве основы информационных и экспертных (стохастических) систем. Целью интеллектуализации является выработка единых способов представления знаний и разработка на этой основе различных задач их использования. Частью этих задач является составление правил выводов и заключений на основе единого понимания пользователями, как семантики данных, так и семантики их сравнимости в заданной предметной области (ПО).

Ставится задача разработки распределенной системы поддержки принятия решений (СППР) на основе моделирования информационных объектов ПО [2], обеспечивающей интеграцию данных и механизмы устранения семантических аномалий [1], возникающих при обработке данных, хранимых в БД, согласно логической схеме [3-5] в больших ПО, с использованием алгоритмов сравнимости атрибутов.

АНАЛИЗ ПОСЛЕДНИХ ИССЛЕДОВАНИЙ И ПУБЛИКАЦИЙ

Существующие информационные модели [1, 3, 6, 7] представления знаний о ПО не предусматривают возможности использования условной синонимии – лексической неоднозначности в именах элементов структур данных в случае их семантической эквивалентности (и наоборот). Составленные на их основе методы вывода знаний существенно снижают возможности СППР. Следует отметить, что неоднозначность лексики и семантики всегда присутствует при использовании БД в качестве хранилища данных, с кото-

рым работают пользователи различных специальностей.

ФОРМУЛИРОВКА ЦЕЛЕЙ СТАТЬИ

Статья посвящена вопросам разработки методов и алгоритмов определения сравнимости и эквивалентности множеств значений атрибутов при условии многозначности лексических и семантических значений имен атрибутов на основе моделирования ПО средствами теории типов [2], что позволит не только осуществлять интеграцию данных в БД, но и избежать семантических аномалий в операциях реляционной алгебры [3-5].

ИЗЛОЖЕНИЕ ОСНОВНОГО МАТЕРИАЛА ИССЛЕДОВАНИЯ

Предлагаемая методика [2] представления знаний с помощью абстрактных типов данных дает возможность использовать теорию типов применительно к объектам ПО для определения сравнимости соответствующих структур данных в случае лексической и семантической неоднозначности атрибутов реляционной БД. Под неоднозначностью лексики и семантики атрибутов понимаем тот факт, что два атрибута могут иметь семантически различное значение в случае лексической однозначности их имен и в то же время два атрибута могут быть семантически эквивалентными вне зависимости от лексической однозначности их имен. Семантическая эквивалентность строго определяется далее в соответствии с критериями сравнимости. На содержательном уровне будем считать, что два атрибута, согласованные с одним доменом, семантически эквивалентны, если определены на одном объекте и их роли на нем совпадают [2].

Решение задачи устранения неоднозначности между лексическим значением имени и семантикой атрибута требует предварительного решения задачи сравнимости атрибутов и увязки имени атрибута со структурой ПО, а также способов формализованного описания информационных объектов, их характеристик и связей [2].

В [2] объекты ПО рассматриваются как структурированные типы данных, которые могут взаимодействовать с другими структурами фрагмента посредством своих элементов. Описать такой объект можно при помощи признаков или смыслов $P=\langle RTO_1, RTO_2, ..., RTO_n\rangle$. Признак можно рассматривать с точки зрения его роли в классификации типов объектов таким образом, что признаком *P* можно назвать кортеж предикатов $P=\langle RTO_1, RTO_2, ..., RTO_n\rangle$, а значением признака, задающим тип объекта *X*, определенную последовательность из нулей и единиц вида $\langle |RTO_1(X)|, ..., |RTO_i(X)|\rangle$, где $|RTO(X)|=1$, если $RTO(X)$ истина и нуль, если $RTO(X)$ – ложь. Таким образом, каждому значению кортежа соответствует тип объекта с именем O_i на множестве типов объектов $O=\{O_i\}$, соответствующего множеству значений кортежа P. Здесь $O=\{O_i\}$ – множество типов объектов ПО, O_i имя структурированного типа данных, отвечающего некоторому значению признака *P*, на подмножестве объектов ПО. Каждый элемент $O_i \in O$, где *O* – множество типов объектов рассматриваемой ПО, определим в виде: $O_i(\{RTO_1, ..., RTO_n\}; \Omega_i)$, понимая при этом тип объекта как множества данных со смыслами элементов структурированного типа $RTO_1, RTO_2, ..., RTO_n$ соответственно определенных посредством операции Ω_i интерпретации типа. Авторами [2], предложено формализованное описание среды (ПО) для последующего анализа семантики данных БД. Под средой понимается тройка

$\langle O, RTO, \Omega \rangle$, где *O* – множество объектов, *RTO* – множество смыслов (свойств) на этих объектах, Ω – множество операций, которые можно выполнить с элементами множеств *O* и *RTO* при решении задач.

Итак, ПО состоит из множества *O* информационных объектов O_i: $O = \{O_1, O_2 ..., O_i, ..., O_r\}, i = 1, 2, ..., r$, каждый из которых является параметризованным (родовым) типом данных вида: $O_r\left(\left\{RTO_1^r, RTO_2^r, ..., RTO_s^r\right\}; \Omega_r\right)$, где Ω_r – множество операций на RTO_i. В свою очередь каждый из параметров RTO_k^r на объекте *r* может быть представлен множеством своих значений $\{A_1, A_2, ..., A_k, ..., A_q\}$, в частности множествами значений имен атрибутов БД. Последнее позволяет задавать критерии сравнимости (в частном случае эквивалентности атрибутов РБД).

В этом случае родовой тип данных будет иметь вид:

$$O_r\left(\left\{\begin{array}{l}\left\langle RTO_1^{(r)}, \{A_1, A_2, ..., A_{n1}\}\right\rangle, ..., \\ \left\langle RTO_s^{(r)}, \{A_1, ..., A_{ns}\}\right\rangle\end{array}\right\}; \Omega_r\right).$$

Использованию информации в СППР будет предшествовать представление знаний о ПО и составление на его основе критериев сравнимости атрибутов. Под атрибутом $[A_i]$ будем понимать пару $\langle A_i, \{a_i\}\rangle$ из имени атрибута и множества его значений соответственно. Под доменом $[D_i]$ будем понимать пару $\langle D_i, \{d_i\}\rangle$ из имени домена и множества его значений.

В статье предлагаются следующие критерии сравнимости атрибутов:

1. Если для атрибутов $[A_k], [A_l], A_k \in A, A_l \in A$ и для некоторого домена $[D_i]$ выполняется условие отображения множеств значений атрибутов в один и тот же домен,

$$([A_k] \to [D_k]) \wedge ([A_l] \to [D_k]) \quad (1)$$

и множество атрибутов отображается в множество имен доменов, а каждый атрибут с именем A_k в один и только один домен с именем D_i то будем говорить **о сравнимости атрибутов с точностью до связи с доменом**. В этом случае аргументы смысловой функции *P* могут отличаться типами объектов и именами идентификаторов, но домены значений области определения для атрибутов с именами A_k и A_l совпадают.

На содержательном уровне это означает, что атрибуты вне зависимости от лексического значения его имени имеют один шаблон значений, свойственный домену.

2. Если тип объекта задан перечислением имен идентификаторов набором операций, то будем говорить о его задании **с точностью до имен идентификаторов**. Под возможностью замены имени $[A_i]$ именем $[RTO_i]$ будем понимать, что A_i принадлежит множеству значений идентификатора, то есть факт отображения $[A_l]$ в $[RTO_i]$ означает, что: $(A_l \in \{A_i\}) \wedge [RTO_i] = \langle RTO_i, \{A_i\}\rangle$.

Если имена атрибутов A_l и A_k могут быть заменены одним и тем же значением RTO_i из $\{RTO_i\}$, то считаем их сравнимыми с точностью до имени идентификатора. В этом случае, аргументы смысловой функции, определяющие подмножество значений атрибутов A_l и A_k ,содержат одно и то же имя идентификатора, однако атрибуты могут быть согласованы с различными типами объектов и являться подмножествами доменов D_i и D_j, где $i \neq j$.

На содержательном уровне это означа-

ет, что атрибуты имеют одинаковый смысл, возможно, на различных объектах ПО и принадлежат различным доменам.

3. Будем считать, что атрибуты $[A_k]$ и $[A_l]$ **сравнимы с точностью до идентификатора,** если они сравнимы с точностью до имени идентификатора и отображаются в один домен из множества имен доменов, в которые отображаются атрибуты логических схем БД:

$$([A_k] \to RTO_i, [A_i] \to RTO_i) \wedge (A \to D) \wedge \exists D_s \in D \wedge [A_k] \to [D_s] \wedge [A_l] \to [D_s] \quad (2)$$

Это означает, что атрибуты $[A_l]$ и $[A_k]$ согласованы с одним доменом и, возможно с различными типами объекта, но со смыслом RTO_i.

4. Если для атрибутов $[A_k]$ и $[A_l]$ выполняется условие:

$$([A_k] \to [RTO_k], [A_l] \to [RTO_l]) \wedge (RTO_k \in O_i) \wedge (RTO_l \in O_i), \quad (3)$$

то атрибуты $[A_k]$ и $[A_l]$ сравнимы с точностью до типов объектов и аргументами функции смысла для определения подмножеств доменов значений атрибутов являются основы одного типа объекта.

5. Если для атрибутов $[A_k]$ и $[A_l]$ выполняются условия (1-3), то будем говорить, что атрибуты $[A_k][A_l]$ эквивалентны. Это означает, что аргументами функции смысла для определения значений атрибутов является одна и та же основа и наборы ограничений совпадают. Таким образом, подмножества значений атрибутов допускают интеграцию по данным и семантические ограничения на операции реляционной алгебры для атрибутов с именами A_k и A_l отсутствуют. В случае, если атрибуты являются результатом отображений в логическую схему БД свойств объектов связей инфологической модели ПО [2], то в случае эквивалентности значений таких атрибутов типы объектов и значения идентификаторов на них должны совпадать.

ВЫВОДЫ

С использованием методов моделирования ПО средствами теории типов [2] разработаны критерии определения сравнимости и эквивалентности атрибутов реляционной БД. Их использование позволяет эффективно решать задачи интеграции данных в больших ПО корпоративных СППР и устранения семантических аномалий при операциях над данными.

ЛИТЕРАТУРА

1. Calenko M.Sh. Modelirovanie semantiki v bazakh dannykh /M.Sh. Calenko. – M.: Nauka, 1989. – 285 s.
2. Kochueva Z.A. Metody modelirovanija informacionnykh ob'ektov predmetnoj oblasti pri postroenii sistem podderzhki prinjatija reshenij v korporativnykh sistemakh /Z.A. Kochueva, N.V. Borisova //Problemy informacionnykh tekhnologij. – 2015. – №02 (018). – S. 117-120.
3. Buslik M.M. Global'nye shemy reljacionnyh baz dannyh /M.M. Buslik. – H.: HNURJe, 2002. – 67 s.
4. Dejt K. Vvedenie v sistemy baz dannyh: per. s angl. /K. Dejt. – M.: Nauka, 2001. – 1071 s.
5. Mejer D. Teorija reljacionnyh baz dannyh: per. s angl. /D. Mejer. – M.: Mir, 1987. – 608 s.
6. Chen P.P.-S., The entity Relationship Model: Towards a Unified View of Data. – ACM Trans. Database Syst., 1976, #1, P. 9-36.
7. Codd E.F., Extending the Database Relational Model to Capture More Meaning. – ACM Trans. Database Syst., 1976, #4, P. 397-434.

Рецензент: д.т.н., проф. Гамаюн І.П.
Національний технічний університет «Харківський політехнічний інститут»