УДК 004.358:681.518

В.А. Филатов, А.А. Хайрова

ПРЕДСТАВЛЕНИЕ ОНТОЛОГИЙ СРЕДСТВАМИ РЕЛЯЦИОННОЙ МОДЕЛИ ДАННЫХ

Введение

В настоящее время корпоративные системы являются одной из наиболее эффективных структурных компонент построения информационно-ориентированного общества. Основными тенденциями развития систем такого класса являются: управление знаниями, реинжиниринг, организационное обучение и самообучение, интеллектуальные интерфейсы, Internet/Intranet-стратегии, технологии аутсорсинга, информационные порталы.

При разработке корпоративных систем на первое место выходит не программный аспект, а задача извлечения, формулирования, структурирования и представления информации. Рассматриваемая статья посвящена исследованиям, направленным на эффективное применение онтологического подхода в задачах моделирования и представления знаний в системах обработки информации.

Онтология − это структурная спецификация некоторой предметной области, ее формализованное представление, которое включает словарь или имена указателей на термины предметной области и логические выражения, которые описывают, как они соотносятся друг с другом. Таким образом, онтологии представляют собой некоторое отображение предметной области посредством связей на множестве концептов, используя отношения: часть-целое, класс-подкласс, родитель-потомок, и др. [1]

Различные группы пользователей, занимающиеся обработкой и анализом информации, используют специальную терминологию, которая применяется другими сообществами в ином контексте. В то же время в различных предметных областях часто встречаются общие обозначения для одних и тех же понятий. Компании, организации, учреждения постоянно увеличивают свой объем первичной информации, обладая огромными информационными ресурсами, недоступными для других пользователей. Проблема интеграции информации состоит в том, как и какими средствами организовать доступ к этим данным, придав им форму, удобную для конечного пользователя, как представить знания, которые обеспечивали бы автоматизированную обработку информации.

1. Анализ предметной области и постановка задачи исследований

Онтологии и онтологические системы могут быть использованы при решении различных задач автоматизированной обработки и коллективного доступа к данным. Приведем несколько примеров использования онтологий:

- процесс создания программных продуктов, документов в системе управления документооборотом, WEB-страниц, подготовки HTML-документов и т.д. Информационная система в этом случае описывается средствами единого языка представления данных для использования в многозадачных системах. Преимуществом такого подхода является многократное использование знаний и универсальные свойства информационной базы.
- общий доступ к информации. Информация запрашивается одним или несколькими пользователями или приложениями. Онтологический подход помогает избежать избыточности и информационной несовместимости за счет общего словаря терминов или преобразования одного набора терминов в другой. Преимущество подхода заключается в эффективном использовании информационных ресурсов.
- индексация. Онтология используется как механизм индексации информации. Основным преимуществом использования онтологий в этом случае является быстрый доступ к информационным ресурсам по сравнению с существующими методами поиска информации.

С учетом особенностей использования онтологий и существенных преимуществ, которые дает онтологический подход, может быть сформулирована цель проводимых исследований: разработка эффективных структур хранения данных в онтологических системах.

При решении поставленной задачи должны быть учтены следующие свойства онтологий [2]:

- в онтологиях знания формализуются в виде описаний предметной области с помощью иерархии классов;
- для каждого класса задается свой набор свойств и объектов;
- свойства в онтологиях имеют область определения – класс, для которого задается это свойство и область значений. В зависимости от областей значений свойства делятся на два типа: свойства, значениями которых являются константы заданного типа данных, и свойства, значениями которых являются объекты заданного класса.

Следует отметить, что онтологии могут быть информационно насыщенными – в некоторых из них помимо сложной иерархии с множеством классов и свойств, могут храниться миллионы объектов. Иерархическая структура отологий может быть эффективно реализована на основе реляционной модели,

широко используемой в технологии баз данных. Методы хранения должны позволять помещать в базу знаний информацию о предметной области в виде онтологий, а также обеспечить высокую скорость выполнения запросов.

2. Математическое обоснование реляционной модели представления онтологий

Формально онтологию можно представить в следующем виде.

Пусть дана некоторая предметная область D. Предположим, что данную предметную область можно представить, иерархией классов с именами cn_i ... Любой класс является подмножеством предметной области D то есть $I(cn_i) \subseteq D$, где $I(cn_i)$ - множество объектов класса с именем cn_i.

Тогда, для описания данной предметной области достаточно иметь конечное множество классов, которое обозначим $CN = \{cn_1, cn_2,...,cn_p\}$, $p<\omega$. Наследование представляет собой частичный порядок, заданный на CN: то, что cn_i наследует класс cn_j (является его подклассом) будет обозначаться как $cn_i \succ_D cn_j$. Если характеризовать классы с информационной точки зрения, то об элементах наследующего класса мы имеем больше (точнее, не меньше) информации, чем об элементах наследуемого класса. Это означает, что $cn_i \succ_D cn_j$ влечет $I(cn_i) \subseteq I(cn_j)$, то есть, чем больше информации о классе, тем меньше сам класс [3].

Атрибуты, имена которых будем обозначать через p_i должны принадлежать к классам. Например, атрибут «фамилия» присущ классу людей, но не имеет смысла для класса автомобилей. С другой стороны, аргумент (значение атрибута) также должен быть либо элементом некоторого класса, либо элементом базовой модели данных $\check{R} = <M_1,\ldots,M_s; \Omega>$, то есть значения атрибутов также должны быть типизированы. Отсюда появляется отношение, связывающее атрибуты и классы.

Выражение $p_i \lhd_D <cn, m>$, где $cn \in CN$ и $m \in CN \bigcup DT$, означает, что в предметной области D атрибут с именем p_i характеризует элементы класса cn_j, а значениями атрибута, в зависимости от ситуации, могут быть элементы класса или основного множества m. Например, в предметной области общественных связей появляется отношение фамилия $\lhd D$ <человек, строка>. Здесь значением атрибута является строка, то есть элемент базовой модели. В случае супруг $\lhd_D$ <человек, человек> значение атрибута выбирается уже не из базовой модели, а из класса «человек». В практически значимых случаях система описаний каждой конкретной предметной области использует, как правило, лишь конечное число атрибутов. Множества атрибутов будем обозначать через $Attr = \{p_1,...,p_k\}$.

В соответствии с концепцией инфологического моделирования, необходимо именовать объекты предметной области. Однако не все объекты предметной области могут иметь имя. Возможны и безымянные объекты. Имеет смысл использовать только конечные множества имен для каждой отдельной предметной области. С одной стороны, имена моделируют практические ситуации, когда таких имен требуется конечное количество. С другой стороны, конечность множества имен позволяет их использовать в более выразительном и гибком контексте. Множество имен, ориентированных на предметную область D, обозначим через $ID = \{id_1,\ldots,id_q\}$.

Таким образом, общая понятийная структура предметной области D определяется онтологической формой (онтоформой) предметной области D.

$$K_d = <DT, CN, Attr, ID, \succ_D, \lhd_D>, \quad (1)$$

где D – предметная область;

CN, Attr, ID – конечные множества имен классов, объектов и атрибутов, определенных для описания предметной области D;

$\succ_D$ – частичный порядок, определенный на множестве CN;

$\lhd_D$ – отношение, для которого $p_i \in Attr$ и выполняется условие на одной паре $p_i \lhd_D <cn, m>$, где $cn \in CN$ и $m \in CN \bigcup DT$.

Можно показать, что в виде (1) формальное описание онтологии полностью соответствует описанию кортежа в реляционной базе данных [4]:

$$R= <U,D,dom,R,,d,E, O>, \quad (2)$$

где U – универсум, включающий множество атрибутов предметной области;

D – множество доменов, имеющих смысл словарей однородных атрибутивных значений (столбцы таблиц БД);

dom – полное отображение из U в D, указание способа разнесения атрибутов универсума по доменам из D;

R – множество различных схем отношений, причем $R_i \subseteq U$;

d – множество наборов отношений из схемы R;
E – совокупность бинарных отношений над доменами из D;
O – операторы реляционной алгебры: объединения, пересечения, разности, и т.д.
Тогда справедлива эквивалентность следующих понятий:

$$\begin{cases} DT \to U, \\ \{CN, Attr, ID\} -> D, \\ \succ_D \to R, \\ \lhd_D \to \{d, E, O\} \end{cases} \quad (3)$$

В результате полного описания объектов и их свойств, предметная область может быть представлена в виде сложной иерархической базы знаний, спроецированной на реляционную структуру базы данных, над которой можно осуществлять «интеллектуальные» операции, такие как семантический поиск, определение целостности и достоверности данных.

3. Методы реализации онтологических моделей средствами реляционных баз данных

Поисковые системы, основанные на реляционных базах данных, могут реализовать различные технологии доступа к данным. Используя только кортеж для ввода данных, пользователь сможет комбинировать возможности онтологического представление данных с возможностью поисковой системы самой предоставлять отдельные классы данных похожего содержания с целью сужения области поиска. Наиболее перспективным с точки зрения технологии открытых систем является использование Oracle Database.

Система Oracle Database 10g для управления данными, которые могут быть размечены с помощью языка семантической разметки, позволяет использовать ряд преимуществ по сравнению с подходами, основанными на технологии файловых систем или на специализированных базах данных. Прежде всего, это высокая надежность, открытая архитектура, производительность и безопасность. Хранилище онтологий в Oracle можно реализовать с использованием модели данных, в которой отношения между объектами поддерживается помощью связей [5].

В качестве логической структуры представления онтологии может быть использован триплет: ресурс, именованное свойство и его значение. Будем называть эти три составные части: субъект, предикат и объект. Ресурсом называют все, что описывается средствами данного триплета. Под свойством следует понимать некий аспект, характеристику, атрибут или отношение, используемое для описания ресурса. Каждое свойство имеет свой специфический смысл, допустимые значения, тип ресурсов, к которым оно может быть применено, а также отношения с другими свойствами.

На рис. 1 показана структура информационной системы, основанной на реляционной базе данных и использующей онтологический подход к доступу и поиску информации.

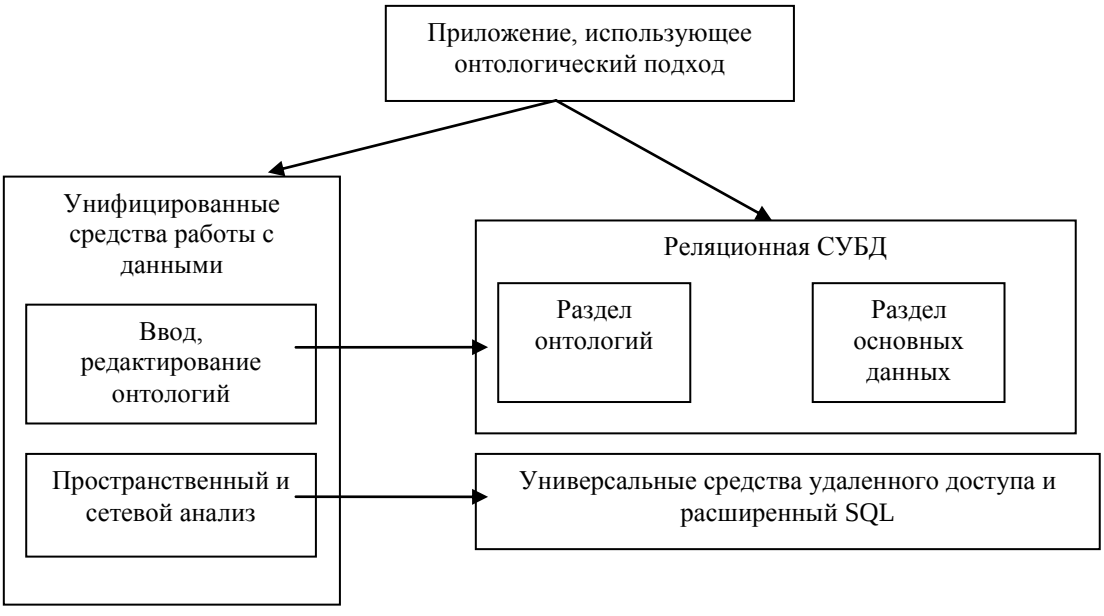

Рис. 1. Структурная схема поддержки онтологии средствами реляционной СУБД

Реляционная модель онтологических данных позволяет хранить как атрибутивные, так и пространственные (онтологические) данные в едином формате СУБД. Хранение триплета в базе данных осуществляется структурой объектного типа, например TRIPLE_TYPE. Такой объект представляет собой кортеж вида:

$$K_d = < id\_model, id\_triple, id\_subject, id\_predicate, id\_object >, \quad (4)$$

Где id_model – уникальный идентификатор предметной области, к которой относится данная запись;

Id_triple – уникальный идентификатор триплета;

Id_subject – уникальный идентификатор субъекта отношения;

Id_predicate – уникальный идентификатор предиката;

Id_object – уникальный идентификатор объекта.

Логическая структура записи в базе данных, содержащая в себе триплет представлена на рис.2.

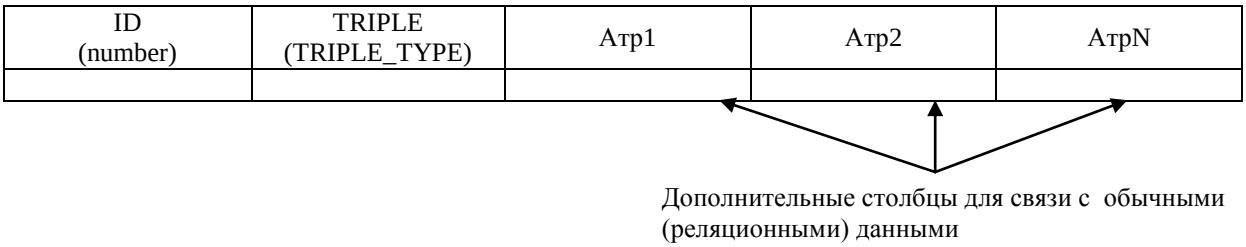

Рис. 2. Структура кортежа реляционной модели представления онтологии

Где ID – идентификатор кортежа в отношении; TRIPLE – триплет; Атр 1, Атр 2, Атр N – имена атрибутов данных.

Для эффективного хранения онтологий в реляционной базе данных может быть рассмотрена унифицированная модель, представленная на рис.3.

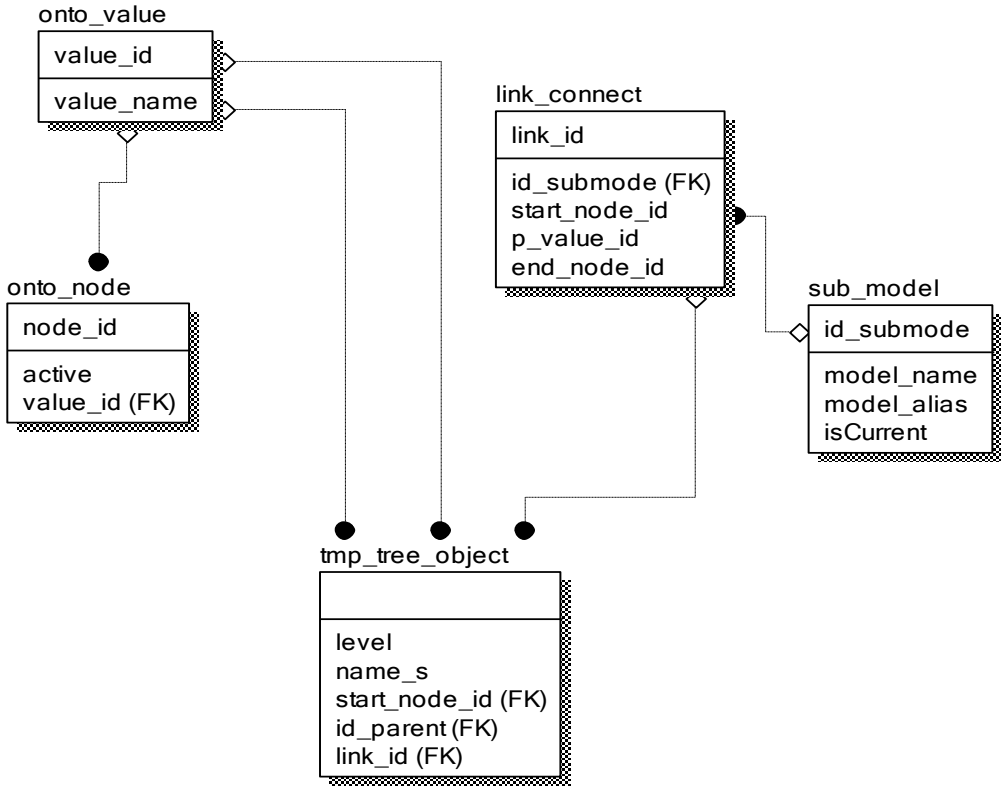

Рис. 3. Реляционная модель данных представления онтологии

На рис.3 представлены: sub_model – отношение, предназначенное для хранения имен онтологий; onto_node – отношение для текстовых значений субъектов и объектов; lnk_connect – отношение для

хранение триплетов всех моделей онтологий базы данных; onto_value – отношение для хранения текстовых значений, URI или литералов для каждой из частей триплета; tmp_tree_object – отношение, предназначенное для хранения иерархической зависимости объектов онтологий.

Выводы

В результате исследования основных свойств онтологической модели разработан универсальный метод, который позволяет использовать реляционную модель данных для хранения данных онтологии.

Основными преимуществами такого подхода являются: представление онтологических данных в открытом, стандартизированном формате; представление онтологических и семантических данных средствами СУБД; стандартизованный, унифицированный доступ, как к онтологическим, так и к семантическим данным через SQL.

ЛИТЕРАТУРА:

1. Филатов В.А., Хайрова А.А. Технология организации образовательных web-сервисов на основе XMLDB, HTMLDB, ORACLE SPATIAL // Міжнародний науково-технічний журнал «Інформаційні технології та комп'ютерна інженерія» – Вінниця: ВНТУ. – 2007. – Вип. 1(18) – с. 240 – 247.

2. Филатов В.А., Щербак С.С., Хайрова А.А. Разработка высокоэффективных средств создания и обработки онтологических баз знаний // Системи обробки інформації. – Х.:ХВУ. – 2007. – Вип.8 (66) – с.120 – 125.

3. Филатов В.А. Модель поведения автономного агента на основе теории автоматов // Вестник Херсонского государственного технического университета. – Херсон: ХГТУ. – 2004. – №1(19). – С. 108 – 111.

4. Арсеньев Б.П., Яковлев С.А. Интеграция распределенных баз данных – СПб.: Издательство «Лань», 2001.

5. Xavier Lopez, Susie Stephens, Jeam Ihm, Jayant Sharma, Melliyal Annamalai, Omar Olonso. Semantic Data integration for the Enterprise, March 2006.

ФИЛАТОВ Валентин Александрович – доктор технических наук, профессор кафедры Искусственного интеллекта Харьковского национального университета радиоэлектроники

Научные интересы: базы данных, агентные технологии, мультиагентные системы, извлечение знаний из данных.

ХАЙРОВА Антонина Алиевна - магистр кафедры Искусственного интеллекта Харьковского национального университета радиоэлектроники

Научные интересы: онтологии, базы данных и знаний, распределенные информационные системы.