ПРИМЕНЕНИЕ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИДЕНТИФИКАЦИИ И МОНИТОРИНГА ПРЕДМЕТНЫХ ОБЛАСТЕЙ

С.В.Мальцева,
д.т.н., профессор Государственного университета – Высшей школы экономики
smaltseva@hse.ru

Рассматриваются принципы создания онтологических моделей предметных областей с учётом динамики их изменения. Предложен шаблон хранилища данных для хранения и модернизации динамической онтологии, чьё основное свойство – изменение во времени состава и структуры кластеров понятий. Приведены сведения об использовании таких онтологий в практических задачах.

Введение

Понятие предметной области – одно из фундаментальных понятий в современных методологиях анализа и проектирования.

Предметная область определяется как часть реального мира, рассматриваемая в пределах определённого контекста, который может задавать область знания, отрасль экономической деятельности – в широком смысле, а в более узком – область исследования, область деятельности предприятия, конкретного специалиста и т.д. Сам термин предполагает описание совокупности объектов, являющихся предметом некоторой деятельности.

В образовательной сфере и сфере трудовых отношений используется термин «область профессиональной деятельности». Им обозначают области науки и техники, объединяющие совокупность объектов, средств, приёмов, способов и методов человеческой деятельности. В образовательных стандартах область профессиональной деятельности идентифицируется через описание объектов, видов и задач профессиональной деятельности выпускников.

Идентификация предметной области связана с построением её адекватной модели, имитирующей её структуру или функционирование.

Один из существующих сегодня подходов к идентификации предметной области, основанных на идее концептуального моделирования, – онтологическое моделирование. Концептуальная, или понятийная *модель предметной области* (МПО) описывает её как совокупность понятий (концептов, терминов) и отношений между ними, которым соответствуют сущности из реального мира [1]. Этому соответствует классическое представление онтологической модели, в котором онтология задаётся тремя конечными подмножествами концептов, связей и функций интерпретации. При моделировании предметной области как сферы деятельности отношения между понятиями также являются понятиями, описывающими отношения. Понятия, отнесённые к классу отношений, используются для описания процессов и явлений реального мира. Поэтому более правильной представляется концепция моделирования предметной области на основе объединения понятийной и содержательной МПО, приведённая в работе [1]. Понятийная МПО определяется как совокупность понятий (концептов, терминов) и отношений между ними, которым соответствуют сущности из реального мира, реализованная в виде ориентированного помеченного графа. Содержательная МПО для понятийной модели задаётся ориентированным помеченным графом, вершины которого интерпретируются как информационные элементы, соответствующие реальным объектам предметной области. Соответственно, выделяются два типа отношений в объединении моделей: содержательные, определяющие отношения одного информационного элемента к другому, и понятийные, определяющие отношения элемента к концепту из понятийной МПО.

Рассматривается задача практического использования онтологического моделирования для идентификации предметных областей.

Приведённое выше определение МПО косвенно указывает на два важных аспекта использования онтологий для моделирования предметных областей.

Первый аспект касается рассмотрения сегодняшних проблем практического применения онтологий, которые связаны с использованием онтологий, в большинстве случаев, как словарей или тезаурусов; при этом связи между понятиями не используются (исключение составляют лингвистические онтологии [2]). Интерпретация связей как объектов онтологии, позволяющих описывать процессы и явления, тесно коррелируется с проблемами объединения систем управления контентом предприятия (Enterprise Content Management, ECM.) и системами моделирования и управления бизнес-процессами (Business Process Management, BPM). Такой подход позволяет сделать онтологии пригодными для моделирования динамики изменения предметных областей.

Второй аспект связан с выделением в МПО понятийной и содержательной моделей. Для интенсивно развивающихся предметных областей МПО – это постоянно изменяющаяся и развивающаяся во времени структура. Можно говорить о том, что содержательная модель – это средство накопления изменений, которые с течением времени приводят к изменению понятийной модели. Использование *динамических онтологий*, являющихся функциями от времени (или, как альтернатива, включающих множество временных периодов, связанное с множествами концептов и связей), позволит обеспечить актуальность и адекватность онтологических моделей и сделает их практически применимыми на широком спектре задач.

Рассмотрим возможность создания некоторого типового шаблона реализации онтологии предметной области в виде концептуальной модели хранилища данных с учётом отображения динамики её изменений во времени, разделения понятийной и содержательной составляющих, интерпретации множества связей как подмножества понятий. Учёт этих требований позволит использовать предложенный шаблон не только для прикладных задач, но и для задач мониторинга предметной области и модернизации онтологии. При его создании необходимо учитывать общепринятый набор требований, предъявляемых к онтологическим моделям. Наиболее общие из них для большинства работ в этом направлении: ясность при передаче смысла терминов, обозначающих понятия; согласованность; возможность модернизации. При формировании информационных элементов важно учесть возможность введения мультиязычности и множественности толкований понятий.

Создавая шаблон хранилища данных, используем реляционную модель. Опишем её системой множеств и векторов.

Обозначим основные множества онтологии:

$$C = \{ c_i \mid i = 1,..., N \} -$$

множество понятий, обозначающих объекты, процессы или явления;

$$R = \{ r_{jl} \mid j = 1,..., M \} -$$

множество связей между понятиями.

Чтобы использовать преимущества интерпретации отношений между понятиями как некоторого класса понятий и множественность отношений между понятиями, что очень удобно при описании процессов, целесообразно рассматривать множество *R* как подмножество множества *C*. Это же справедливо и для всех множеств понятий, вводимых ниже.

Элементам множества *C* ставится в соответствие набор векторов, чьи значения компонент определяют их атрибуты. Минимальный набор атрибутов включает:

$$\beta_1 = \{\beta_{1i}\}, \;\; i = 1,...,N -$$

вектор идентификаторов понятий, где β_{1i} – идентификатор *i*-го понятия;

$$\beta_2 = \{\beta_{2i}\}, \;\; i = 1,...,N -$$

вектор названий понятий, где β_{2i} – название *i*-го понятия;

$$\beta_3 = \{\beta_{3i}\}, \;\; i = 1,...,N -$$

вектор описания смысла понятий, где β_{3i} – описание *i*-го понятия .

Элементам множества *R* можно поставить в соответствие набор векторов, значения компонент которых определяют их атрибуты. Минимальный набор атрибутов включает:

$$\gamma_1 = \{\gamma_{1j}\}, \;\; j = 1,...,M -$$

вектор идентификаторов связей между двумя связываемыми понятиями из множества *C*, где γ_{1j} – идентификатор *j*-ой связи;

$$\gamma_2 = \{\gamma_{2j}\},\ j = 1,...,M,$$
$$\gamma_3 = \{\gamma_{3j}\},\ j = 1,...,M\ -$$

векторы, компоненты которых γ_2 и γ_3, соответственно, задают идентификаторы первого и второго связываемых понятий c_i и c_l

$$(c_i, c_l \in C,\ i, l \in [1, N],\ \gamma_{2j} = \beta_{1i},\ \gamma_{3j} = \beta_{1l});$$

$$\gamma_4 = \{\gamma_{4j}\},\ j = 1,...,M -$$

вектор наименований связей между понятиями c_i и c_l

$$(c_l,\ c_l \in C,\ i, l \in [1, N],\ \gamma_{2j} = \beta_{1i},\ \gamma_{3j} = \beta_{1l}),$$

где γ_{4j} – наименование j-ой связи;

$$\gamma_5 = \{\gamma_{5j}\},\ j = 1,...,M -$$

вектор описаний связей между понятиями c_i и c_l

$$(c_i, c_l \in C,\ i, l \in [1, N],\ \gamma_{2j} = \beta_{1i},\ \gamma_{3j} = \beta_{1l}),$$

где γ_{5j} – описание j-ой связи.

Этот набор параметров для элементов множества C часто дополняется весовыми коэффициентами понятий. Вводится еще один вектор,

$$\beta_4 = \{\beta_{4i}\},\ i = 1,...,N -$$

вектор весов понятий, где β_{4i} – вес i-го понятия o_i, в интервале (0,1). Веса понятий характеризуют их важность для определения предметной области. Они определяются на основе экспертных оценок, на основе частотных характеристик появления в информационных ресурсах, а также контекста употребления.

Для связей вводятся весовые коэффициенты, указание направления связи и типизация связей в соответствии с классификацией, принятой в методологии объектно-ориентированного анализа. Однако это справедливо, если мы выстраиваем онтологию предметной области, подразумевая, что за понятиями стоят объекты, процессы и явления. Для лингвистической онтологии требуется другая типизация связей. Между двумя понятиями могут существовать интегрированные множественные связи, объединяющие связи нескольких типов. В различных задачах можно учитывать разные компоненты таких интегрированных связей.

Введём типы связей, объединяющих понятия онтологии:

$$A = \{a_q \mid q = 1,..., N_A\} -$$

множество типов связей между понятиями онтологии. Элементам множества A ставится в соответствие набор векторов, значения компонент которых определяют их атрибуты:

$$\alpha_1 = \{\alpha_{1q}\},\ q = 1,..., N_A -$$

вектор идентификаторов типов связей между понятиями онтологии, где α_{1q} – идентификатор q-го типа связи, a_q;

$$\alpha_2 = \{\alpha_{2q}\},\ q = 1,..., N_A -$$

вектор наименований типов связей между понятиями, где α_{2q} – наименование q-ого типа связи, a_q;

$$\alpha_3 = \{\alpha_{3q}\},\ q = 1,..., N_A -$$

вектор описаний типов связей между понятиями, где α_{3q} – описание q-ого типа связи, a_q.

С учётом введённых обозначений элементам множества R можно поставить в соответствие дополнительный набор векторов:

$$\gamma_6 = \{\gamma_{6j}\},\ j = 1,...,M -$$

вектор, компоненты которого задают направленную ($\gamma_{6j} = 1$) или ненаправленную ($\gamma_{6j} = 0$) связь между понятиями c_i и c_l

$$(c_i, c_l \in C, i, l \in [1, N],\ \gamma_{2j} = \beta_{1i},\ \gamma_{3j} = \beta_{1l}),$$

при этом связь направлена от понятия c_i к понятию c_l;

$$\gamma_8 = \{\gamma_{8j}\},\ j = 1,...,M -$$

вектор идентификаторов типов связей между понятиями c_i и c_l, где γ_{8j} – идентификатор типа j-ой связи, значение γ_{8j} выбирается из множества значений, заданных компонентами вектора α_1.

При создании онтологии введение весовых коэффициентов для понятий и связей, а также типизация понятий и связей требует, как правило, привлечения экспертов, даже при использовании автоматизированных методов, позволяющих извлекать термины из наборов документов и текстов, определять их веса и некоторые связи. Процедуры организации работы экспертов представляются достаточно трудоёмкими. Однако, результаты такой работы имеют большую

ценность при решении практических задач, так как позволяют активно использовать веса и связи в наиболее важных задачах выделения кластеров понятий, сравнения и объединения онтологий. Точность решения таких задач резко возрастает.

Приведённый шаблон описания онтологии в значительной степени превышает возможности тезауруса и может использоваться для достаточно широкого спектра прикладных задач, связанных с использованием локальных онтологий. Однако он описывает статичную во времени систему и нуждается в дальнейшем расширении.

Первое направление такого расширения – добавление к предметной онтологии возможностей лингвистической онтологии. Это делает необходимым введение лингвистических атрибутов в описание объектов и в описание связей.

Обозначим

$$L = \{L_k \mid k = 1,...,K\} -$$

множество языков, на которых определена онтология. Каждому языку L_k может быть поставлена в соответствие лингвистическая онтология O_k, задающая алфавит, словарь и правила языка.

$$\lambda_1 = \{\lambda_{1i}\},\ i = 1,...,N_L -$$

вектор идентификаторов языков.

Идентификатор языка L_k выступает как дополнительная координата для ряда атрибутов объектов и связей.

$$\beta_2 = \{\beta_{2i}\},\ i = 1,...,N -$$

вектор названий понятий, где β_{2i} – название i-го понятия c_i;

$$\beta_3 = \{\beta_{3i}\},\ i = 1,...,N -$$

вектор описания смысла понятий, где β_{3i} – описание i-го понятия c_i.

Чтобы обеспечить удобное хранение и использование при решении различных задач атрибутов, привязанных к конкретному языку, примем допущение, что идентификатор понятия однозначно определяет его вне зависимости от наименования на том или ином языке и текста, описывающего смысл понятия. Здесь нужно учитывать, что для одного понятия, обозначающего объект, процесс или явление, может быть (в общем случае) несколько определений и несколько различных толкований (при этом слова, представляющие собой омонимы или омоформы, обозначаются разными идентификаторами). Восприятие определений одного понятия на разных языках, учитывая различие в структуре языков, различны. Поэтому целесообразно при хранении множества определений в мультиязычных онтологиях хранить их как ещё одну версию определения. Такой подход не противоречит активно разрабатываемой идее создания некоторого универсального языка для представления онтологий. Таким образом, можно выделить наименования понятий и их определения, определить как отдельные множества наименований понятий (множество Z) и определений понятий (множество V).

Для решения конкретных задач, особенно в целях обеспечения интероперабельности, необходимо однозначное понимание терминов, обозначающих понятия. В определённые периоды времени в каждом языке существуют наиболее употребимые названия понятий и их определения. Целесообразно выделять такие термины в онтологии.

Введение временных параметров обусловлено возможными изменениями онтологии, так как с течением времени не только появляются новые, но претерпевают изменение существующие понятия, их толкование, веса, характеризующие их важность для предметной области, структура и веса связей между ними. Это приводит к новой структуре кластеров понятий и категорий, описывающих предметную область.

Важные параметры для многих предметных областей – указание источников определений и толкований понятий.

Рассмотрим атрибуты, задающие множество понятий, как информационных элементов.

Элементам множества

$$Z = \{z_{i_z} \mid i_z = 1,...,N_z\}$$

можно поставить в соответствие следующий набор векторов:

$$\delta_1 = \{\delta_{1i_Z}\},\ i_Z = 1,...,N_Z -$$

вектор идентификаторов названий понятий из множества C;

$$\delta_2 = \{\delta_{2i_Z}\},\ i_Z = 1,...,N_Z -$$

вектор кодов понятий из множества C, где δ_{2i_z} принимает значения из множества значений, которые принимают компоненты вектора β_1;

$$\delta_3 = \{\delta_{3i_Z}\},\ i_Z = 1,...,N_Z -$$

вектор кодов языков из множества L, где δ_{3i_Z} принимает значения из множества значений, которые принимают компоненты вектора λ_1;

$$\delta_4 = \{\delta_{4i_Z}\},\ i_Z = 1,...,N_Z -$$

вектор названий понятий из множества C, где δ_{4i_z} – название понятия с идентификатором δ_{2i_z} на языке с кодом δ_{3i_z};

$$\delta_5 = \{\delta_{5i_Z}\},\ i_Z = 1,...,N_Z -$$

вектор весов названий понятий из множества C, где δ_{5i_z} – вес названия понятия с идентификатором δ_{2i_z} на языке с кодом δ_{3i_z}.

Весовой коэффициент определяется на основе экспертных оценок и частоты употребления термина. Термин, имеющий самый высокий вес, можно использовать как основной термин для обозначения понятия, остальные названия – как синонимы.

Элементам множества

$$V = \{v_{i_V} | i_V = 1,...,N_V\}$$

можно поставить в соответствие следующий набор векторов:

$$\omega_1 = \{\omega_{1i_V}\},\ i_V = 1,...,N_V -$$

вектор кодов определений понятий из множества C;

$$\omega_2 = \{\omega_{2i_V}\},\ i_V = 1,...,N_V -$$

вектор идентификаторов названий понятий из множества C, где ω_{1i_1} принимает значения из множества значений, которые принимают компоненты вектора δ_1;

$$\omega_2 = \{\omega_{2i_V}\},\ i_V = 1,...,N_V -$$

вектор определений понятий из множества C, где ω_{1i_1} – текст определения понятия, название которого задано j-ой компонентой вектора δ_4, на языке, код которого задан j-ой компонентой вектора δ_3, такими, для которых j-е значение компоненты вектора δ_1, $\delta_{1j} = \omega_{2i_1}$.

Продолжая рассмотрение лингвистических аспектов онтологии можно добавить к указанным атрибутам параметры источников названий понятий и определений. Это важно для интенсивно развивающихся областей знания, где возникает большое количество новых понятий и их интерпретаций, а также областей, для которых принципиальны корректные определения, например, в частных онтологиях, поддерживающих исполнение внутренних регламентов, при ведении электронных переговоров и т.д.

Для этих задач онтологию целесообразно дополнить разделами источников информации, введя предварительно некоторую их классификацию. Обозначим:

$$D = \{d_{i_D} | i_D = 1,...,K_D\} -$$

множество типов источников информации о понятиях онтологии;

$$H = \{h_{i_H} | i = 1,...,K_H\} -$$

множество источников информации о понятиях онтологии.

Опустим описание набора атрибутов указанных множеств, так как в зависимости от предметной области и решаемых в ней задач он может быть очень лаконичным или развернутым. Описание источников в зависимости от задач онтологии складывается из:

- ✧ описаний литературных источников, принятых в библиографических базах данных;
- ✧ описаний электронных источников информации, включая базы данных, электронные архивы, Интернет-источники;
- ✧ данных экспертов, знания и высказывания которых использовались при формировании онтологии.

Для каждой из этих категорий источников существуют стандарты или регламенты, задающие форму и атрибуты библиографического описания.

Первоисточник названия и его интерпретации не всегда можно точно указать. В этом случае в онтологии целесообразно указывать наиболее часто упоминаемый источник, хотя это и может приводить к некоторым искажениям с точки зрения временных параметров существования понятий. Исключение составляют понятия, которые вводятся в рамках законов, стандартов и различных регламентов. Многие из таких понятий возникают в практике задолго до появления соответствующих документов, их определяющих, например, такие понятия, как «информация», «информационный поиск» и многие другие. Для большинства задач, где используются онтологические модели, достаточно использования понятий

в интерпретации, задаваемой в соответствующих документах, однако хранение полного набора версий наименований и интерпретаций понятий повышает семантическую адекватность онтологической модели.

Временные изменения онтологии могут касаться любой из её частей, однако, наиболее частые следующие:

- ✧ добавление новых понятий;
- ✧ изменение весов понятий;
- ✧ изменение толкования понятий;
- ✧ изменение структуры и весов связей.

Самый существенный результат этих изменений – изменение структуры категорий, выделяемых в онтологии, и, как частный случай, выделение новых предметных областей.

Введение временных параметров при формировании хранилища данных (ХД) может производиться на основе введения идентификаторов временных периодов в описание соответствующих разделов ХД.

Можно выделять временные периоды на основе заданного интервала (например, год или полгода), некоторой последовательности разных по величине интервалов или по событийному принципу (отмечать точную дату изменения какого-либо атрибута объектов онтологии). Выбор варианта определяется интенсивностью развития понятийного аппарата предметной области, но первые два варианта представляются предпочтительными, так как для изменения многих параметров необходимо использовать статистические данные за некоторый период и привлекать экспертов. Вариант с разными выделенными интервалами возможно употребить при использовании понятий, появившимися в отдалённых временных периодах. При появлении новых понятий или их толкований, которые определены, например, новым стандартом или законом, начинающих действовать с определённой даты, необходимо определять точные временные параметры. Учесть эти соображения можно за счёт введения дополнительных временных атрибутов в описание указанных выше множеств.

Обозначим

$$T = \{t_{jT} \mid j_T = 1,...,M_T\} -$$

множество временных периодов, рассматриваемых при создании онтологии.

Для элементов множества T вводится стандартный набор атрибутов, описывающих координату времени в хранилищах данных и позволяющих ввести идентификатор временного периода, определить его начало и окончание, задать его описание.

Введение множества периодов позволяет объединить идентификатор периода с идентификаторами элементов всех перечисленных выше множеств, получив модель развития онтологии во времени. Эта модель позволяет получать временные срезы онтологии, проследить траектории изменения трактовки понятий, изменение структуры классов понятий.

Решение задачи кластеризации понятий с учётом прогнозируемых изменений весов понятий позволяет прогнозировать появление новых областей профессиональной деятельности на основе глобальных онтологий. Для локальных онтологий, как онтологии корпоративных информационных систем, сетевых сообществ, можно решать задачи прогнозирования появления новых направлений деятельности.

Описанный шаблон позволяет вводить новые множества понятий онтологии, связывая их с уже существующими, а также с множеством языков и временных периодов. Так, во многих задачах, где используются локальные онтологии, в состав модели вводятся разделы, связанные с идентификацией пользователей онтологии.

Разработанный шаблон ХД использован при разработке концепции сервисной компоненты для формирования предметной области (домена) сетевого сообщества практики, реализующей функции «сервера отношений» [3]. Использование для реализации сервера онтологии позволяет формировать группы участников с учётом тематики их деятельности и возникающих задач, что повышает качество и интенсивность взаимодействия. Это обеспечивает систематизацию интегрального знания сети, идентификацию её домена, сохранение и планируемое изменение границ домена сети; направленное формирование её ресурсов, взаимодействие с внешними объектами.

При проектировании сервера определены необходимые сервисы сети по поддержке и развитию сетевого домена на основе динамической онтологии:

- ✧ сервисы формирования онтологий: создание и развитие онтологии домена сети, создание онтологий объектов, определение онтологий внешних объектов;
- ✧ определение сходства онтологий: для внутренних объектов; для внешних и внутренних объектов,
- ✧ кластеризация объектов сети на основе измерения сходства онтологий.

Динамика изменений доменов сетевых сообществ характеризуется высокой интенсивностью. Поэтому разработана методика мониторинга и модернизации домена сетевых сообществ.

Методика базируется на принципе объединения лингвистической и предметной онтологий и основывается на методах автоматического лингвистического анализа работ участников сети для выделения новых понятий и поиска возможных связей с понятиями домена и автоматизированном учёте изменения частотных и весовых характеристик существующих в домене и новых понятий.

Заключение

Введение системы координат, по которым происходит изменение онтологии (время, языковая группа) позволяет создавать онтологические модели предметных областей, которые не просто фиксируют появление новых понятий и их связь с уже существующими понятиями, но позволяют проследить изменение состава предметных областей и их границ по этим координатам. Это позволяет повысить качество онтологического моделирования предметных областей за счёт создания более адекватных моделей. Для практической реализации этого подхода необходимо объединение концепций построения онтологий с концепциями хранилищ данных и методами OLAP (On-line Analytical Processing). Разработанный шаблон концептуальной модели хранилища данных может быть применён для широкого спектра задач создания онтологических моделей предметных областей. ■

Работа выполнялась при поддержке Научного фонда Государственного университета – Высшей школы экономики (индивидуальный исследовательский проект №07-01-189 «Применение онтологических моделей для решения задач идентификации и мониторинга развития областей профессиональной деятельности».

Литература

1. Интегрированные информационно-телекоммуникационные системы и сети, телекоммуникационные и информационные ресурсы, информационные процессы в управляющих системах и сетях . Отчёт о НИР/ (2004–2006 г.г.). Программа 3.2. Раздел 3.2.2. Разработка фундаментальных основ создания распределённых информационно-вычислительных ресурсов ИВТ СО РАН. http://www.sbras.ru/Report2006/Report321.
2. Б. В. Добров, Н. В. Лукашевич. Лингвистическая онтология по естественным наукам и технологиям как ресурс для приложений информационного поиска. Web Journal of Formal, Computational & Cognitive Linguistic // http://fccl.ksu.ru/issue_spec/docs/oent-kgu.doc.
3. С.В. Мальцева, Д.С. Проценко. Серверы отношений сетевых сообществ практики на основе онтологических моделей. Автоматизация и современные технологии, №3 , 2008. – С. 26–29.