И.В. АНТОНОВ

ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

В статье рассматриваются возможные подходы к расширению функциональных возможностей традиционных редакторов онтологии предметной области за счет дополнения их средствами автоматизированного синтеза онтологической информации. Рассмотрен метод автоматического определения места нового понятия в структуре онтологии.

В качестве основы для построения систем семантически-ориентированного доступа к информационным ресурсам могут быть использованы онтологии предметной области. С применением онтологий предметных областей в качестве компонента информационных систем в настоящее время связывают перспективы развития систем машинного представления и обработки знаний, обучающих систем и систем информационного поиска. Прикладное использование онтологий чаще всего предполагает взаимодействие с онтологической системой программных агентов, обеспечивающих обработку запросов, поступающих от пользователей и реализующих взаимодействие компонентов распределенных приложений. Однако онтологии могут использоваться и в качестве инструмента согласования и документирования точки зрения специалистов на состав, отношения и интерпретацию совокупности понятий, используемых в той или иной предметной области. В этом случае использование онтологии предполагает непосредственную работу пользователей с онтологической моделью. По мере расширения и эволюции представлений о предметной области в процессе практики должна обновляться и онтология. Соответственно, важным требованием к системам ведения онтологий является обеспечение ими поддержки операций редактирования и расширения состава существующих онтологий.

Стандартным инструментом построения онтологий являются редакторы онтологий – Protégé, OntoEdit, Ontolingua, OilEd и другие [1]. Традиционная методология построения онтологии в среде указанных редакторов предполагает итеративное формирование онтологии «сверху-вниз» на основе постепенно детализируемой модели системы понятий предметной области, создаваемой экспертом. Типичными операциями, поддерживаемыми редакторами онтологий являются ввод в модель атрибута понятия, ввод понятия и состава атрибутов понятия, ввод связи между понятиями, задание ограничений на значения атрибутов, создание экземпляра понятия и т.д. После первоначального формирования онтологии возможны операции редактирования – переназначение связей между понятиями, изменение состава атрибутов и некоторые другие операции. Уровень автоматизации операций с онтологиями, реализованный в современных редакторах, сводится к контролю корректности используемых идентификаторов и типов данных и контролю отсутствия логических противоречий в создаваемой модели. Состав и структура онтологии при этом полностью определяются экспертом.

В ряде работ рассматриваются возможности внедрения в процесс разработки онтологии элементов автоматизации. Например, в работе [2], где авторы предлагают осуществлять построение онтологии на основе извлечения знаний из терминологических словарей с помощью разрабатываемой системы продукций. Однако лексическая многозначность и контекстная зависимость семантики выражений естественного языка создает труднопреодолимые проблемы для разработчиков систем автоматического извлечения онтологической информации из текстов предметной области. В то же время, отсутствие формализованных и автоматизированных процедур синтеза структуры онтологий является фактором, сдерживающим дальнейшее распространение технологий, использующих онтологические модели предметной области. Представляется перспективным направление исследований, связанное с разработкой средств автоматизации операций онтологического инжиниринга на этапе введения в онтологическую модель новых понятий и выполнения операций по корректировке структуры онтологии. Как правило, мало формализуемой является предшествующая

указанному этапу деятельность экспертов по содержательному анализу предметной области, выделению в ней объектов и понятий, и определению их структурного содержания. То есть, объектная декомпозиция предметной области и структурная декомпозиция характеризующих ее понятий, это задачи, которые не поддаются жесткой формализации и требуют для своего решения непосредственного участия эксперта предметной области. Однако в рамках работы по формированию онтологии предметной области можно выделить такие задачи, где формализация не только возможна, но может послужить инструментом, повышающим эффективность и качество работы эксперта. К таким задачам можно отнести анализ связей и состава понятий, существующих в системе и вводимых в нее, на целостность, непротиворечивость и полноту, а также автоматизированное формирование и использование логических следствий, порождаемых вносимой в онтологию новой информацией.

Процесс построения онтологии предметной области является итерационным и предполагает последовательное включение новых понятий в онтологическую иерархию. В случае формирования онтологии предметной области, используемой для аннотирования ресурсов, источником информации для введения в онтологию новых понятий будут являться объекты предметной области, характеризуемые определенным набором свойств. Набор атрибутов, характеризующих объект, представляет собой онтологическую моделью понятия, обозначающего совокупность объектов данного типа. Соответствующую модель должен формировать эксперт предметной области при включении в онтологию нового понятия.

Опишем процедуру итерационного наращивания онтологии, обеспечивающую автоматическое определение места нового понятия в структуре онтологии. Традиционно на вершине онтологической иерархии располагается единственное базовое понятие, включающее в себя универсальные атрибуты, характеризующие любой из объектов предметной области. Будем рассматривать это как общее условие для формируемых онтологий предметной области. В случае отсутствия у объектов предметной области универсальных атрибутов список атрибутов корневого понятия иерархии может быть пустым. Обозначим корневое понятие онтологии как c_0.

Для формирования модели вводимого в систему понятия эксперту предоставляется полный набор атрибутов входящих в онтологию понятий M:

$$M = \bigcup_{\forall i} M_i = \{m_1, m_2, ..., m_n\}$$

Эксперт может расширить указанный набор дополнительными атрибутами, характеризующими новое понятие. Затем эксперт формирует на основе множества M множество атрибутов M_x, характеризующих включаемое в онтологию понятие.

$$M_x = \{m_{1_x}, ..., m_{d_x}\}$$

Пусть эксперт присвоил вводимому понятию имя c_x, а также сформировал множество его атрибутов M_x. После введения этой информации дальнейшие операции по включению в онтологию понятия c_x выполняются автоматически.

Выполнение процедуры поиска начинается с инициализации вспомогательных множеств. Корневое понятие включается в качестве первого элемента в множество RP, которое на промежуточных шагах процедуры содержит родительские понятия c_x, включая опосредованных и прямых его предков. По завершении процедуры поиска множество RP содержит полный набор прямых предков понятия c_x, что означает решение задачи по определению места включения c_x в иерархию понятий. Также для вводимого понятия формируется множество потенциальных родительских понятий $P = \{c_{j_1}, ..., c_{j_f}\}$ таких, что $\forall j (j \in \{j_1, .., j_f) : r_{0j} = 1$, где r_{ij} - определенное на

множестве всех понятий отношение, равное 1 при условии прямого наследования понятием j от понятия и равное 0 в остальных случаях.

Опишем последующие шаги предлагаемой процедуры.

В процессе поиска выполняется просмотр множества P потенциальных родителей вводимого понятия.

1. Из P выбирают очередной элемент. Пусть из P выбрано понятие $c_j \mid j \in \{j_1,..,j_f\}$. Тогда возможны четыре варианта:

2.1. $M_j \equiv M_x$. В данном случае обнаруживается тождественность вводимого экспертом понятия c_x понятию c_j онтологии. Процедура завершается уведомлением пользователя об этом факте.

2.2. $M_j \not\subset M_x \wedge M_x \not\subset M_j$. В этом случае понятие c_j и производные от него понятия не могут состоять в отношении наследования с понятием c_x, так как состав атрибутов этих понятий не соотносится с атрибутами понятия c_x как множество и подмножество. Ветвь дерева понятий, начинающаяся с понятия c_j, исключается из дальнейшего рассмотрения. Для выбора следующего кандидата переходят к п.1.

2.3. $M_j \not\subset M_x \wedge M_x \subset M_j$. Это тот случай, когда вводимое понятие должно быть вставлено между понятиями c_k и c_j, где $k : r_{kj} = 1$. В этом случае понятие c_j и производные от него понятия исключаются из дальнейшего поиска. Вместо существовавшего в онтологии отношения $(c_k R c_j)$ водятся новые два отношения $(c_k R c_x)$ и $(c_x R c_j)$, $r_{kx} = 1$ и $r_{xj} = 1$. Для дальнейших операций поиска родительских понятий переходят к п.1.

2.4. $M_j \subset M_x$. Это означает, что c_x наследует c_j. Вводим в множество RP элемент c_j и переходим к п.3.

3. Из множества родительских понятий RP исключают, если таковое обнаруживается, родительское по отношению к c_j понятие. В результате этойопе-рации множество RP освобождается от опосредованных предков понятия c_x: $\exists k : (c_k, c_j) \in R \wedge c_k \in RP \rightarrow c_k \notin RP$, т.е. $rp_{kx} = 0$

4. Для c_j формируется множество непосредственно наследующих от него понятий $V = \{c_{v_1},...,c_{v_g}\}$ таких, что $\forall v (v \in \{v_1,..,v_g\} : r_{jv} = 1$. Множество V пополняет множество P' (множество понятий, подлежащих анализу в следующем цикле поиска родительских понятий) : $P' = P' \cup V$. Если множество P просмотрено не полностью, переходят к п.1, иначе - к п.5.

5. Если множество P' не пусто ($|P'| \neq 0$), оно замещает множество P: $P = P'$ и переходят к п.1. Иначе переходят к п.6.

6. Место для включения вводимого понятия в иерархию найдено ($|RP| \neq 0$) и оно включаться в состав онтологии.

Итоговое множество RP в общем случае может включать в себя более одного элемента. Это будет означать, что понятие c_x имеет в онтологии более одного прямого предка, то есть, формируется на основе отношений множественного наследования:

$$\forall j : (c_j, c_x) \in RP \rightarrow c_x \in C \wedge \forall (c_j, c_x) \in R, \text{ т.е.}$$

$$\forall j : (c_j, c_x) \in RP \rightarrow r_{jx} = 1$$

Информация, позволяющая на основе формально-логических методов восстановить всю иерархию обобщенных (абстрактных) понятий онтологии, полностью содержится в множестве понятий самого нижнего уровня онтологии, Обобщения верхних уровней могут быть выявлены на основе обработки состава понятий нижних уровней онтологии путем обнаружения повторяющихся в различных понятиях наборов атрибутов. Существуют формальные процедуры, позволяющие синтезировать иерархию обобщенных понятий по матрице, характеризующей распределение атрибутов по объектам предметной области или понятиям нижнего уровня, характеризующим эти объекты. Соответствующий метод был предложен Р.Вилле в 1982 году [3] и получил название метод анализа формальных понятий (FCA). Однако на основе данного метода синтезируются все формальные обобщения, включая те, которые могут не иметь семантической ценности для онтологии предметной области. Поэтому метод анализа формальных понятий не применим в качестве инструмента автоматизированного формирования прикладной онтологии. Как возможной вариант решения этой проблемы для автоматизации онтологического инжиниринга можно использовать метод, объединяющий итерационный формальный синтез обобщений с семантической разметкой обобщенных понятий, выполняемой экспертом. В рамках предлагаемого подхода после определения места нового понятия в иерархии могут быть автоматически определены новые понятия более высоких уровней на основе выявления общих для различных понятий наборов атрибутов:

$\forall k,x: M_n = M_x \bigcap M_k, M_n \neq M_x, M_n \neq M_k, |M_n| \neq 0, c_n \notin C \rightarrow c_n \in C$, где c_n - новое понятие верхнего уровня по отношению к понятиям c_x и c_k, имеющее набор атрибутов M_n.

Выявляемые таким образом на основе формальных признаков понятия верхних уровней могут, как представлять семантическую ценность для онтологии, так и не иметь ее. Поэтому на следующем шаге эксперт должен классифицировать полученное понятие тем или иным образом. В первом случае понятие пополняет состав онтологии, во втором – попадает в множество «ложных понятий», не имеющих семантической значимости. Значимые понятия именуются экспертом и включаются в состав онтологии предметной области. Их незначимых понятий формируется список, используемый в дальнейших операциях с онтологией таким образом, что вниманию эксперта на каждом следующем шаге будут предлагаться в качестве обобщенных понятий лишь не встречавшиеся ранее сочетания атрибутов. Итеративное формирование множества «ложных понятий» в составе онтологической модели может рассматриваться как обучение системы распознаванию повторяющихся сочетаний атрибутов, не имеющих семантической ценности.

Предлагаемый подход позволяет строить системы семантически-ориентированного доступа к информационным ресурсам на базе онтологий предметной области. При этом подходе онтология формируется на основе последовательного ввода пользователем в систему моделей объектов предметной области. Расширение состава учитываемых системой объектов сопровождается формированием «вертикальной» иерархии, включающей абстрактные понятия, характеризующие предметную область. В отличие от традиционных редакторов онтологий предлагаемая система предоставляет автоматизированную поддержку операций пополнения онтологии новыми понятиями, включающую в себя автоматическое определение места нового понятия в онтологической иерархии и автоматический синтез новых обобщений. Сформированная таким образом онтология может использоваться в качестве средства семантического аннотирования информационных ресурсов, к которым относятся, в частности, любые файлы и web-ресурсы. Аннотированные ресурсы включаются в онтологическую систему в качестве объектов (экземпляров) онтологии.

ЛИТЕРАТУРА

1. Овдей О.М., Проскудина Г.Ю. Обзор инструментов инженерии онтологий // Электронные библиотеки – Москва: Институт развития информационного общества. – Т. 7, вып. 4, 2004.

2. Найханова Л.В., Хаптахаева Р.Б., Янсанова Е.Н. Создание декларативного метода извлечения знаний из терминологических словарей // Информационные технологии. – 2008. – №12. – С. 2–8.
3. Wille R. Concept lattices and conceptual knowledge systems. // Computers and Mathematics with Applications. – №23, 1992.

И.В. АНТОНОВ

МОДЕЛЬ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ СИСТЕМ СЕМАНТИЧЕСКИ-ОРИЕНТИРОВАННОГО ДОСТУПА

В статье рассматривается модель онтологии предметной области и структура базы данных для хранения онтологии, поддерживающие возможности итерационного автоматизированного формирования структуры онтологии.

Онтологии предметной области в настоящее время находят основное применение в области построения поисковых систем, систем представления знаний, инженерии знаний и при решении задач семантической интеграции информационных ресурсов. Под онтологией понимается «формальная спецификация концептуализации, которая имеет место в некотором контексте предметной области» (Т. Грубер, 1993 [1]). В свою очередь, концептуализация - представление предметной области через описание множества понятий (концептов) предметной области и связей (отношений) между ними. Основным отношением, учитываемым при построении онтологии, является родовидовое отношение между понятиями (отношение гипоним-гипероним), на основе которого формируется таксономия понятий. Представление совокупности понятий предметной области и их отношений в основном реализуется в современных онтологических системах на основе модели семантической сети фреймов. Узлы сети представляют отдельные понятия предметной области, дуги – отношения между понятиями. Отдельное понятие в этой модели представляется фреймом, слоты которого содержат атрибуты понятия. Производные (дочерние) понятия наследуют атрибуты базовых (родительских) понятий. На этапе определения понятий онтологии для их атрибутов обычно задается имя и тип атрибута. Конкретные значения эти атрибуты получают при создании на основе понятий онтологии экземпляров (объектов). Операции по созданию экземпляров понятий поддерживает большинство онтологических систем. При этом экземпляры чаще соответствуют понятиям нижних уровней онтологической иерархии. Таким образом, онтология представляет собой иерархию понятий, характеризующих предметный мир, объекты которого соответствуют преимущественно понятиям нижних уровней онтологии, а промежуточный и верхний ее уровни представляют, как правило, абстракции различной степени обобщения.

Существующие системы, построенные на основе онтологий, рассчитаны, как правило, на работу с онтологией программных агентов, обрабатывающих те или иные информационные запросы. Одним из перспективных направлений развития онтологических систем является построение систем, использующих онтологическую систематизацию как инструмент классификации объектов предметной области, с которыми работают пользователи, и как средство для организации семантически-ориентированного доступа пользователей к этим объектам. К числу потенциальных областей продуктивного применения указанного подхода относится работа пользователей персональных компьютеров с файлами и документами. Традиционные средства доступа к файлам основаны на выборе пользователем папок и файлов в иерархической структуре файловой системы. Инструментом доступа в таком случае является программа, реализующая функции файлового менеджера. С ростом числа файлов и усложнением структуры файловой системы поиск нужного документа и файла становится все более затруднительным для пользователей. Решением проблемы может быть организация доступа посредством семантически-ориентированных интерфейсов,