УДК 004.822

ОНТОЛОГИЯ КАК СИСТЕМАТИЗАЦИЯ НАУЧНЫХ ЗНАНИЙ: СТРУКТУРА, СЕМАНТИКА, ЗАДАЧИ

Кузнецов О.П. Суховеров В.С. Шипилина Л.Б.

olkuznes@ipu.ru suhoverv@ipu.ru lubship@ipu.ru

Институт проблем управления РАН им. В.А. Трапезникова

Москва

Рассматриваются вопросы создания и использования онтологий, перечисляются основные задачи онтологии и подходы к их решению. Описаны формы задания запросов, в том числе построение стандартных и расширенных запросов на языках, предназначенных для работы с базами знаний, в которых для представления данных используются дескриптивные логики. Приводятся фрагменты онтологии научного учреждения, разрабатываемой в среде Protégé.

Введение

В связи с интенсивным ростом объемов информации в сети Интернет актуальной задачей является повышение эффективности использования информационных ресурсов большого объема. Очевидно, что структурированность информации, присущая онтологиям, обеспечивает дополнительные возможности для решения задач информационного поиска, и в первую очередь это относится к такой «хорошо» формализуемой области, как научные знания.

Онтология (греч.) – раздел философии, учение о бытии (в отличие от гносеологии - учения о познании), в котором исследуются всеобщие основы, принципы бытия, его структура и закономерности. В сфере искусственного интеллекта онтология – это дисциплина, связанная с построением специфической системы понятий, которая описывает определенную предметную область. Содержание понятий отражается с помощью концептов. Формально в онтологии концепт отождествляется с объектом (классом), имеющим связи с другими классами. Класс определяется как множество экземпляров с общими свойствами и содержит описания собственно экземпляров и их свойств.

Сначала для представления знаний использовались модели баз данных, затем фреймовые структуры и семантические сети. В настоящее время формируется комплексный подход к разработке информационных систем обработки и управления знаниями. Он реализуется с помощью тематических Интернет-порталов, предоставляющих пользователю различные интерактивные сервисы, главным из которых является поисковая система портала. На основе семантических Интернет-порталов, использующих онтологические модели и семантику при описании предметных областей, разрабатывается согласованный набор методов для решения базовых онтологических задач, таких, как классификация понятий, создание таксономий и отношений, интеграция, отображение и модификация онтологий, навигация – информационный поиск и запросы. При создании онтологии важно определить круг пользователей, для которых создается онтология, и предполагаемую среду машинной реализации [1 – 5]. На сегодняшний день существуют десятки сред для разработки онтологий [6] из которых можно выделить Cyc [2], предоставляющую средства разработки на коммерческой основе, и Protégé [7], распространяемую свободно. В таких средах создаются специализированные онтологические Интернет-порталы для самых разных областей знаний. В портале CYC реализована база знаний, которая объединяет формализованное представление обширного количества фактов с правилами просмотра и эвристиками для запросов об объектах и событиях [2]. В Интернет-порталах реализован электронный словарь английского языка WordNet [27], представлены стандартизованные классификации предметов потребления (commodity) и терминологии для продуктов и сервисов

обеспечения коммерческой деятельности [26]. В области медицины широко известны такие информационные системы на основе онтологий, как UMLS (Unified Medical Language System), разработки американской NLM (National Library of Medicine) и связанные с ней специализированные системы, способствующие поддержанию громоздкой UMLS в актуальном состоянии и ускоряющие поиск [13,22-24]. Из других областей знаний можно указать портал по компьютерной лингвистике[25], а в области техники – порталы поддержки процессов управления знаниями в организациях, например, «Управляемые электроприводы», «Пусконаладочные работы» [20].

Онтологии научных знаний наделены возможностями ввода новых данных и интеграции с другими онтологиями и обладают большими возможностями для организации информационного поиска, как пользователями, так и программными агентами [3]. В онтологиях эффективность информационного поиска, то есть построения запросов и получения ответов на них с оценкой и интерпретацией результата, определяется соответствием семантики словаря и структурных связей онтологии реальным описаниям и связям объектов в предметной области.

В представленной работе для задач поиска показано использование семантики при анализе предложений и при расширении запросов методом построения связующих термов. Основные положения работы проиллюстрированы примерами из разработки онтологии, описывающей научно-организационную деятельность учреждения и содержащей в качестве отдельных ветвей онтологические описания научных дисциплин. Рассмотрены формы задания запросов в онтологии. Разработка проводилась в среде Protégé.

1. Структура онтологии

Онтология предназначена для представления структурированных знаний и формально описывается кортежами типа $\mathbf{<L, C, P, A, F, G, H^C, I>}$, где

$\boldsymbol{C}$ – понятия (классы), $\boldsymbol{I}$ –экземпляры $\boldsymbol{c_i} \square \boldsymbol{C}$, $\boldsymbol{c_j} \square \boldsymbol{I}$, $\boldsymbol{L}$ – словарь: $\boldsymbol{L^C \square L^P \square L^A \square L^{VA}}$,

$\boldsymbol{P: C \times C}$ - отношения, $\boldsymbol{A: C \times L^{VA}}$ – атрибуты, $\boldsymbol{F: L^C \rightarrow C}$, $\boldsymbol{G: L^P \rightarrow P}$,

$\boldsymbol{H^C}$ - частичный порядок на C: $(c_i, c_j) \square H^C \rightarrow c_i$ – подпонятие c_j ($c_i \square c_j$),

$p \square P$, $p(c_i, c_j)$: c_i – домен (domain), c_j – диапазон(range)

Пример: для отношения p = *иметь публикацию*:

Dom (p) : *персона* (класс), Range (p): *публикация* (класс)

(т. е. экземпляр класса *персона* связан с экземпляром класса *публикация* отношением p) .

Структура онтологии представляется семантической сетью – ориентированным графом, вершинами которого являются понятия (классы) и экземпляры, а ребра отражают семантические отношения между понятиями и/или экземплярами. Классы и экземпляры имеют свойства (атрибуты). Кроме того, в онтологии имеются аксиомы и правила вывода, которые используются при решении онтологических задач (например, при построении ответов на запросы). Эффективность решения задач существенно улучшается, если структура онтологии выстроена в соответствии со следующими требованиями [2,3]:

• На семантической сети выделен скелет онтологии – иерархия классов предметной области в виде графа. При этом ребра дерева отражают семантические отношения типа: IS-A (KIND-OF) – таксономические, PART-WHOLE, TOPIC–SUBTOPIC.

• Классы содержат экземпляры. Между классами и экземплярами различных классов наряду с иерархическими связями могут быть и другие «горизонтальные» семантические связи, представляемые объектными бинарными отношениями, характерными для описываемой предметной области (Рис.1).

• Онтология снабжена глоссарием. При построении/модификации онтологии имена понятий, экземпляров, отношений и т. д. выбираются из глоссария.

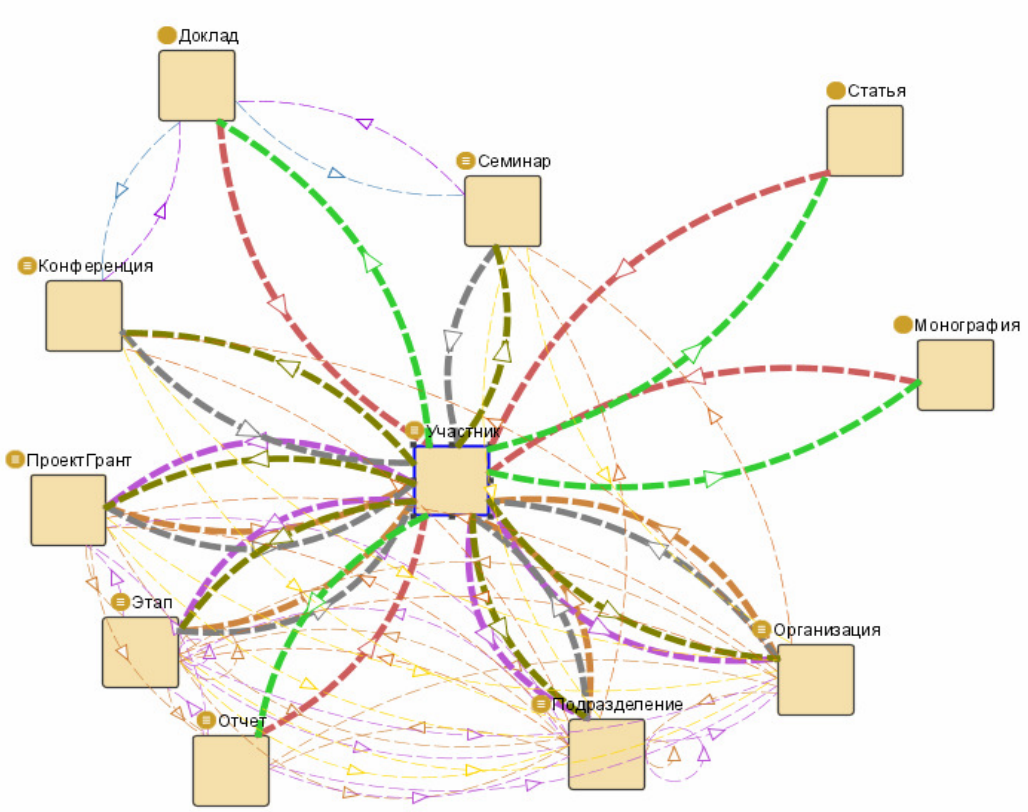

Рис.1. Структура бинарных отношений класса «Персона». Фрагмент онтологии «Научно-организационная деятельность»

• Онтологии научного знания разделяются на онтологии верхнего уровня, описывающие «крупные» разделы разных направлений, например, **предметно-независимые онтологии** определенного научного направления, то есть целого ряда научных дисциплин, и онтологии **предметных областей** (ПО) для конкретных научных дисциплин.

• При построении структуры онтологии в описании ПО выделяются содержательные сегменты, в которых определяются категории наиболее важных понятий и с их помощью составляется метаописание ПО. Например, для подразделов онтологии ПО «*Принятие решений»,* метаописание содержит категории-классы: *модель, метод, задача, интерпретация* и *объяснение результатов, инструментарий, область применения.*

2. Проблемы и задачи в онтологических информационных системах

Огромные объемы пополняемой информации требуют разумного компромисса в комбинированном применении онтологических и лингвистических методов, используемых для развития и поддержки Интернет-порталов. Например, тезаурус портала UMLS содержит около миллиона концептов. Им соответствует около двух миллионов строк на английском языке. При таком количестве концептов поддержка отношений на актуальном уровне в статической онтологии становится трудно выполнимой задачей. Это является препятствием для ввода и дальнейшего использования новых данных, содержащихся в корпусах текстов. Для того, чтобы поддерживать соответствие такой огромной коллекции концептов и новых терминов, необходимо применение методов NLP (*Natural Language Processing*), то есть лингвистических методов для выявления новых терминов из актуальных текстовых материалов, пополняющих информационную систему [1,28].

Для развивающихся областей, например биомедицины, не существует понимания области в целом. Поэтому статические онтологии не могут обеспечить эффективного взаимодействия с

пользователями. В этом смысле эффективным представляется моделирование предметной области через лексические описания для ввода новых терминов, чтобы дополнить знания о концепте и специфицировать его новые свойства или атрибуты. Новые открытия могут изменить понимание концепта, отражаемого термином, и, соответственно, изменить смысл концепта.

Для UMLS обычной является ситуация, когда вводимый термин позже обозначает несколько разных концептов, и возникает необходимость ввода новых терминов. С другой стороны, возможна ситуация, когда два различных термина, используемые разными сообществами пользователей, в итоге могут объединиться в один концепт и слиться в один термин. Часто концепты не имеют полного описания, поскольку они сами развиваются, то есть неполным является представление о них. Это отражается в разнообразии терминов, используемых для названия концепта в развивающихся предметных областях, и UMLS наглядно демонстрирует это.

В традиционных онтологиях схемы первичны по отношению к набору их логических следствий. В противовес этому, в научных онтологиях набор следствий (феноменов, которые нужно объяснить) появляется раньше, и исследователи пытаются выстроить онтологию, с помощью которой они могут вывести или объяснить эти феномены в наиболее непротиворечивой форме. Другими словами, создание «правильной» онтологии является научным исследованием, которое устанавливает непротиворечивые пути объяснения реальности.

Одна из проблем современных информационных систем – это создание средств для совместного использования и поддержки онтологий, разработанных независимо друг от друга. Решению этой проблемы посвящены задачи **отображения и интеграции** онтологий. Наиболее проблематичным пунктом в этом подходе является совместное использование словарей, соответствующих разным онтологиям, в которых сходная информация представлена разными терминами и понятиями. OBSERVER – один из Интернет-порталов, содержащих несколько онтологий и решающих задачу их совместного использования. В компьютерной сети, на базе которой реализован OBSERVER, можно выделить функциональные подсистемы. Одна из подсистем – Interontology Relationships Manager (IRM) – предназначена для решения словарной проблемы и используется при интеронтологичекой обработке запросов для выявления синонимической информации. В качестве достижений системы информационного поиска OBSERVER, следует отметить использование отношений «*подкласс-надкласс*», которые позволяют расширить семантику запроса [30].

В решении задач отображения и интеграции онтологий, информационного поиска и построения запросов, ввода новых документов важную роль играют оценки семантической близости (ОСБ) классов и экземпляров. Вначале ОСБ строились на утверждении: чем больше информации разделяют два понятия, тем они ближе (геометрический подход)**.** Но затем более объективный подход был предложен Тверски. Его идея состоит в том, что для ОСБ необходимо учитывать не только общие свойства объектов, но и их различные свойства. В настоящее время основу многих онтологических мер близости составляет контрастная модель (**contrast model**) Тверски, определяющая меру близости двух объектов через сопоставление свойств (feature matching), как одинаковых, так и различных. Далее представлена краткая классификация основных подходов к ОСБ.

2.1. Геометрический подход к ОСБ

Сравниваемым объектам ставятся в соответствие точки в некотором пространстве координат, базис которого составляют свойства объектов. Близость двух объектов представляется функцией от расстояния между соответствующими точками пространства. Сложности этого подхода связаны с выбором числа координат и метрик: не всегда возможно поставить в соответствие координате приведенную к численному значению качественную характеристику объекта. Кроме того, субъективная оценка близости объектов не всегда удовлетворяет аксиомам геометрического представления [31].

2.2. Теоретико-множественный подход к ОСБ

Близость двух объектов определяется соответствием их свойств – общих: $A \cap B$ и различных – свойств *A,* которых нет у *B:* $A–B$ и свойств *B,* которых нет у *A:* $B–A$. Свойства могут быть самыми разными по природе, в том числе, и не метрическими [31]. В простой, но очень емкой, контрастной модели близость между объектами $S(a,b)$ является функцией трех аргументов $A \cap B$,

A–B, B–A. Она удовлетворяет аксиомам монотонности, независимости, разрешимости и инвариантности и определяется формулой:

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

где *a* и *b* – два объекта, A и B – множества их свойств, *f* – интервально шкалируемая функция, □, □, и □□0 – веса для общих и различных свойств объектов. Веса позволяют определить ассиметричную меру близости. Развитием этой модели является **ratio model –** нормализованная близость *S*(*a*, *b*) со значениями в интервале {0, 1}:

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}$$

2.3. Использование структурных связей в онтологиях для ОСБ

2.3.1. Учет топологии

Близость двух объектов оценивается по взаимоположению вершин, соответствующих этим объектам в иерархических онтологических структурах отражающих, например, таксономические (IS-A) или меронимические (PART-OF, SUBTOPIC-OF) отношения между объектами [14]. Эта ОСБ использует факт наследования свойств подкласса/экземпляра от своего надкласса и учитывает топологические свойства графа, такие как: а) длина пути (по числу вершин) между исследуемыми вершинами; б) глубина поддеревьев, содержащих исследуемые вершины; в) длины путей от каждой исследуемой вершины до вершины *LCS* (Least common subsumer). Для двух исследуемых вершин *LCS* - наименьшая (ближайшая) вершина — родитель [1].

2.3.2. Учет «горизонтальных» отношений

Для ОСБ объектов используются, так называемые, «горизонтальные» отношения сравниваемых объектов с другими объектами. Эта модель опирается на предположение, что если два объекта имеют одно и то же отношение с третьим объектом, то они ближе, чем два объекта, которые имеют это отношение с различными объектами. А значит, близость двух объектов зависит от близости объектов, с которыми они имеют отношения [14].

2.4. Использование методов компьютерной лингвистики для ОСБ

Эти методы наиболее актуальны для задач ввода новых документов, информационного поиска и построения запросов. Для поддержки в больших онтологиях соответствия огромной коллекции концептов и новых терминов, содержащихся в исследуемых корпусах текстов (например, публикаций), используются методы обработки естественных языков NLP, то есть семантических лингвистических методов, обеспечивающих выявление новых терминов из актуальных текстовых материалов, пополняющих информационную систему[2,28]. Примером использования такого подхода служит Cyc.

2.4.1. Учет частоты встречаемости терминов в корпусах текстов

Используется в статистическом подходе, где близость между понятиями апроксимируется частотой встречаемости термина в исследуемом корпусе текстов (информационном контенте – IC) [1].

2.4.2. Учет лексикографической близости терминов

Исследование контекстной близости основано на автоматическом распознавании и извлечении образцов словосочетаний, содержащих исследуемые термины. Для этой цели используется автоматическая процедура, сочетающая лингвистический и статистический подходы[18]. В выражениях из терминов конструируются контекстные образцы *Context pattern* (CP) – лексические регулярные выражения, соответствующее левому или правому контексту термина. Чтобы сконструировать CP, сначала для всех терминов составляются конкордансы (concordance -алфавитный список всех терминов с указанием контекста использования), затем выполняется нормализация контекста - сортировка составляющих по синтаксическим категориям и удаление незначимых составляющих (наречия, связующие слова). Таким образом получаем нормализованный контекстный образец CP, представляющий собой некое лексическое выражение *p*. После нормализации CP для них вычисляются значения меры, называемые CP-value, чтобы оценить их значимость. CP-value обеспечивает ранжирование контекстных образцов CP по трем

характеристикам: частота появления СР - *f(p)*, его длина |*p*|, как число составляющих, и частота появления СР внутри других СР (|T_p|, где T_p – набор всех СР, которые содержат *p*).

3. Языки и среды для работы с онтологиями

В настоящее время существует ряд машинных языков и сред [6] для создания онтологий и работы с ними. Так, консорциум W3C рекомендует использование языка OWL(Ontology Web Language), который является языком среды Protégé, ориентированной на создание семантических Интернет-порталов [7]. Экранная среда Protégé содержит набор вкладок, обеспечивающих доступ пользователя к инструментарию разработки онтологии, наполнения ее классами, экземплярами классов, свойствами. Свойства типа Object описывают двухместные отношения между экземплярами разных классов, а свойства типа Datatype - атрибуты конкретного класса. Для тестирования онтологии в процессе разработки используются специальные программные средства, называемые reasoner'ами (R) [7]. Имеется возможность создания дополнительных встраиваемых программ, plugin-in'ов, и внешних приложений на языке программирования JAWA.

Формальная семантика семейства языков OWL основана на применении дескриптивной логики (ДЛ) [8]. Изначально ДЛ зародилась как расширение фреймовых структур и семантических сетей механизмами формальной логики. ДЛ оперируют понятиями концепт и роль (бинарное отношение между классами). В терминах ДЛ набор утверждений общего вида называется TBox, набор утверждений частного вида — ABox, а вместе они составляют базу знаний или онтологию. ДЛ, как научное направление, занимается вопросами извлечения неявных знаний из содержащихся в онтологии данных и утверждений и созданием соответствующих программ и алгоритмов. Наиболее часто для разработки онтологий используется диалект OWL DL, который обеспечивает максимальную выразительность без потери полноты вычислений и разрешимости за конечный интервал времени. Ограничения в языковых конструкциях OWL DL (класс не может быть частным свойством; свойство не может быть индивидом или классом) позволяют сделать язык разрешимым. На практике получение новых (inferred) фактов, как результата анализа логических утверждений относительно классов, экземпляров и свойств онтологии, а также проверка синтаксиса документов, например, текста запроса, реализуется специальными программными средствами R (**semantic reasoner, reasoning engine, rules engine**). Средства R обеспечивают различные операции с онтологиями, в том числе, проверку непротиворечивости утверждений, проверку и создание полной иерархии классов, проверку на непротиворечивость экземпляров классов, определение типа данных для экземпляров класса (за счет использования иерархии классов после выполнения классификации). Наиболее часто совместно с Protégé используются R: RacerPro, Fact++ и Pellet.

4. Информационный поиск и запросы в онтологии

Ранее в поисковых системах использовалась индексация ресурсов сети (введение в текст гиперссылок, поиск по ключевым словам) при полном отсутствии средств анализа хранимой информации. Онтологический подход, используемый при создании порталов, позволяет учитывать семантику запросов. Онтология портала знаний включает как описание предметной области, так и описание релевантных ей ресурсов. Описание ресурса сети включает ссылки на ресурсы, описание страниц, сайтов и связей между ними [2,5]. Головная страница портала содержит справочник, имеющий онтологическую структуру, который позволяет управлять поиском информации на связанных страницах портала.

Семантические онтологические порталы обладают довольно мощными способностями к рассуждению, например, за счет семантических описаний терминов в глоссарии средствами клаузальной семантики [5]. В онтологических средах [2,7], использующих языки ДЛ, начальная обработка запроса предполагает переформулировку текста запроса, связанную с особенностями языков таких языков [10,11], и использование методов компьютерной лингвистики при представлении запроса на языке, близком к естественному [2]. Собственно вывод основан на исчислении первого порядка и поддерживается необходимым набором правил: аксиом

включения, эквивалентности и непересекаемости (T-Box) и утверждений об экземплярах (A-Box), например, утверждение о принадлежности экземпляра определенному классу [2].

На сайтах некоторых порталов структура онтологии представлена в графическом виде, и пользователь обладает возможностью *интерактивного информационного поиска*, просматривая фрагменты онтологии и отсекая ненужные ветви, как, например, в PubMed [13].

Но, чаще всего, от пользователя при построении запросов не требуется знания структуры онтологии. Это основной принцип языков запросов (логический уровень отделен от физического). Подходы, берущие начало в базах данных и основанные на синтаксисе, себя исчерпали. В отличие от них, системы поиска, работающие в онтологиии, используют семантические связи понятий для определения их сходства по определенным критериям, называемым мерами близости, что обеспечивает получение более полного и точного результата. Оценки сходства, основанные на топологических графа онтологии, свойствах и отношениях объектов и экземпляров [14], а также ряд методов оценки близости компьютерной лингвистики [15], обеспечивают получение более полных ответов на запросы и их ранжирование по оценке меры близости ответов к запросу [16].

Интересы и квалификация пользователя, а также инструментарий рабочей среды обусловливают многообразие форм запросов. Из них можно выделить стандартные и создаваемые разработчиками:

1. Стандартные:

- Запросы на языках SparQL и SparQL-DL
- Запросы с использованием R

2. Запросы, создаваемые разработчиками

- Запросы по шаблонам
- Запросы с возможностью извлечения неявных знаний.

4.1. Стандартные запросы

Некоторые простые запросы в версиях Protégé от 3.2 и выше можно программировать на языке запросов SparQL, который используется для запросов в OWL DL, например, для выделения (фильтрации) экземпляров определенного класса с заданными свойствами (data properties). Запрос производится по меню открытием панели "Open SPARQL Query" и задается в синтаксисе SPARQL при помощи операторов PREFIX, SELECT FROM, WHERE и т.д. Подробно использование SPARQL запросного механизма в Protégé OWL описывается в документации к Protégé [7]. SPARQL-DL позволяет создавать «правильные» выражения на свойствах OWL и смешивание запросов к TBox/RBox/ABox (к классам, свойствам, экземплярам) [10].

Средствами R Pellet реализована оптимизированная процедура выполнения запросов к ABox. Система выполнения запросов к ABox связана с онтологией через модуль интерфейса базы знаний. Если запрос написан на языке SPARQL, то он считается ABox-запросом при выполнении следующих условий:

- В выражении предиката не используются переменные.
- Каждое свойство в выражении предиката является свойством object или datatype, определенным в онтологии или одним из следующих типов свойств: rdf:type, owl:sameIndividualAs, owl:differentFrom.
- Если в выражении предиката есть свойство типа rdf:type, то его значением является URI.

4.2. Запросы по шаблонам, создаваемым разработчиками

Из-за отсутствия свойств в ДЛ, в запросах к базам знаний на OWL-DL нельзя задавать запросы типа «какое свойство имеется у такого-то понятия», или «какое значение имеет такое-то свойство у такого-то понятия». Все подобные запросы надо переформулировать в виде « найти X, где X описывается аксиомами A1, A2, ...», что бывает не просто и непривычно для пользователя. Запросы, похожие на этот, можно задавать лишь к экземплярам онтологии (запросы типа возвращения экземпляров либо нахождения понятия, наиболее точно описывающего экземпляр).

Для удобства пользователя и учетом возможных часто задаваемых (типовых) запросов мы предлагаем строить для таких запросов шаблоны вкладок-панелей с интерактивными окнами, предполагающие их многократное использование. Нечто подобное представлено в [16]. Использование онтологии с обширными семантическими связями предполагает возможность получать результативные ответы на весьма сложные запросы по шаблонам. Примеры типовых запросов к онтологиям научных дисциплин были перечислены в [3], для некоторых из них нами построены или строятся шаблоны.

Как указывалось в [10,11], особенности представления языков ДЛ требуют переформулировки текста запроса, например при запросе «*Найти всех сотрудников института, имеющих статьи в журнале «Проблемы управления»*», являющихся примером запроса к базе экземпляров (A-Box), будет выбрана соответствующая вкладка — шаблон с окнами для ввода названия класса *Персона* и экземпляра класса *журнал - «Проблемы управления»* и запрос будет переформулирован так: Найти экземпляры X, где X – экземпляры класса *Персона)*, персона (сотрудник) имеет публикацию в журнале (X, журнал «Проблемы управления»). Ответ покажет всех сотрудников института (поименно), которые имеют публикации в журнале «Проблемы управления».

Модель шаблона - вкладки служит конкретным целям запроса; интерактивные окна ввода определяют минимум запросной вводимой информации, а в окнах вывода появляется информация, предлагаемая пользователю как ответ на запрос. Ответ на запрос по шаблону строится на основании информации по предметно-независимой онтологии, онтологии ПО, бинарных отношений объектов и экземпляров, их атрибутов и т.д., фрагменты которых показаны на Рис. 1,3,4.

4.3. Запросы с возможностью извлечения неявных знаний

4.3.1. Исследование структуры онтологии

Анализ структуры онтологии (классов, их надклассов и подклассов, свойств и отношений) позволяет расширять ответы на запросы типа *«Подобрать в подразделениях института сотрудников, которые занимаются разными разделами науки, но имеют общие научные интересы»*. Это осуществляется за счет использования в программах поиска ОСБ, основанных на исследования свойств экземпляров, их классов и надклассов, что позволяет расширить и ранжировать по ОСБ список рекомендуемых специалистов, аналогично тому, как это предлагается в [16].

4.3.2. Построение связующих термов

Для получения более полных ответов на запросы, задаваемые на языке, близком к естественному, в онтологии Cyc [17] из текста запроса лексическими методами выделяются запросные термины, query terms (QT); которые дополняются связующими термами – connecting terms (CT), извлекаемыми из корпуса текстов с использованием семантических лингвистических оценок мер близости. Метод определения CT основан на том, что релевантность пар терминов (inter-term relevance) характеризуется статистикой их совместной встречаемости. Идентификация CT в корпусе текстов является ключевым моментом предлагаемого подхода. Отбор CT состоит в извлечении из предложений корпуса текста контекстных образцов, содержащие QT и их окружение. Затем из образцов отбираются CT с помощью меры контекстной близости [18], сочетающей лексический и статистический (частотный) подходы. Строится граф G (Рис.4), в котором вершинами являются QT и CT, а вес ребер отражает близость между термами и определяется лингвистическими оценками меры близости [15], причем на графе оставляют ребра, имеющие меру, статистически дающую наибольший результат. Затем в G выделяется минимальный связующий подграф (spanning tree), который содержит все QT и множество QT расширяется термами-вершинами связующего дерева, а в корпусе обследуемых документов выделяются предложения, наиболее представительные по частоте встречаемости CT. Вершины на пути в связующем дереве образуют вариант ответа на «расширенный» запрос. Заметим, что в зависимости от выбора вершины-корня в связующем дереве, ответы могут быть различными.

Использование СТ позволяет получить «неочевидные» ответы на запросы, в которых QT совместно не упоминаются ни в одном документе поиска. Эффективность метода построения связующего дерева для расширения запроса показана на примере из [17]:

What does Soviet Cosmonaut Valentina Tereshkova and U.S. Astronaut Sally Ride have in common?. При этом ни в каком документе онтологии они совместно не упоминаются.

Термы запроса QT: *Soviet Cosmonaut, Valentina Tereshkova, U.S. Astronaut, Sally Ride*

Связующие термы СТ: *first, space, woman*

На Рис 4 изображен фрагмент графа G. Вершины, подсвеченные серым цветом представляют минимальный связующий подграф - связующее дерево, в котором *Q1* и *Q2* представляют QT (*Tereshkova, Ride),* а *C1*, *C2* и *C3* представляют СТ.

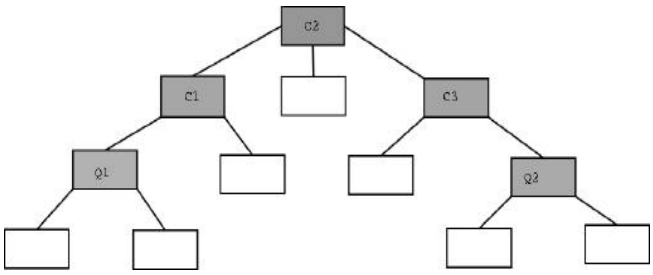

Рис. 2. Пример выбора связующего дерева

Различный порядок выбора связующих деревьев и путей в них позволил получить 6 релевантных ответов, приведем три из них:

1. In 1963, Soviet cosmonaut Valentina Tereshkova returned to Earth after spending nearly three days as the **first woman** in **space**.

1. In 1978, NASA named 35 candidates to fly on the **space shuttle**, including Sally K. Ride, who became America's **first woman** in **space**, and Guion S. Bluford Jr., who became America's **first** black astronaut in **space**

2. Anger in **space,** by astronauts and cosmonauts, has been common since early in the manned **space** program.

Заметим, что часть 2-го и 3-й ответы не содержат ответа на запрос и такой «мусор» в ответах при использовании описанного подхода неизбежен и не подавляется применением более хитроумных мер близости. Для похожей схемы алгоритма расширения запросов в среде Protégé прорабатываются запросные программы (plug-in) для онтологии научных знаний, показанной на Рис.3.

Потребность в расширении ответа на запросы часто возникает в онтологиях, представляющих геометрическую и географическую информацию [19]. Примером такого запроса является запрос типа «*Найти учебное заведение* (университет или институт), *имеющее специализацию по заданной научной дисциплине* (или близкой к ней) *и расположенный в заданном регионе* (или вблизи него)».

5. Разработка онтологии научной деятельности организации

Эта онтология предназначена для описания аспектов научной и организационной деятельности учреждений научной направленности, например, ИПУ РАН. При разработке использовалась среда PROTÉGÉ. На всех этапах разработки, начиная с составления таксономии и кончая наполнением экземплярами, для проверок на непротиворечивость применялся R Pellet. Особое внимание уделялось работе с запросами. Разработан подход к созданию с помощью plugin'ов интерактивных окон, содержащих списки запросов, характерных для определенной предметной области. Исследуются отношения классов предметно-независимой онтологии и соответствующих классов онтологий ПО. Полная онтология включает в себя предметно-независимую часть и совокупность ветвей, описывающих ПО. Ветви ПО строились как отдельные онтологии и импортировались в онтологию предметно-независимой части.

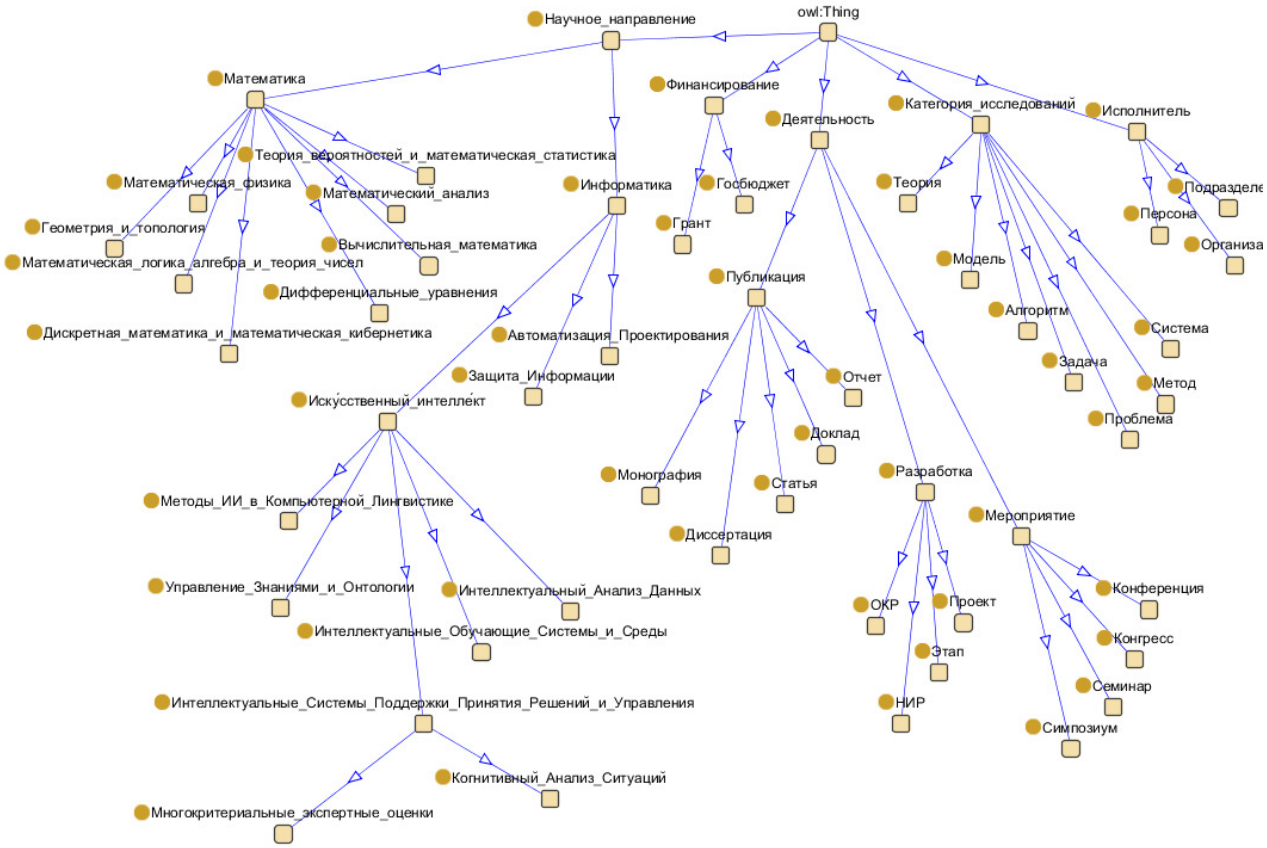

Рис.3. Фрагмент таксономии классов онтологии «Научно-организационная деятельность»

5.1. Предметно-независимая онтология

Эта онтология является предметно-независимой частью общей объединенной онтологии и включает классы: *Деятельность, Исполнитель, Категория исследований, Финансирование* и соответствующие подклассы (Рис. 3).

5.1.Онтология ПО

В качестве онтологии предметной области рассмотрена онтология когнитивного анализа ситуаций (КАС). КАС предназначен для поддержки принятия решений в слабоструктурированных динамических ситуациях. Анализ влияний позволяет формализовать представления знаний экспертов о законах развития и свойствах анализируемой ситуации в моделях когнитивных карт (КК), и получать прогнозы развития ситуации с последующими проверками корректности

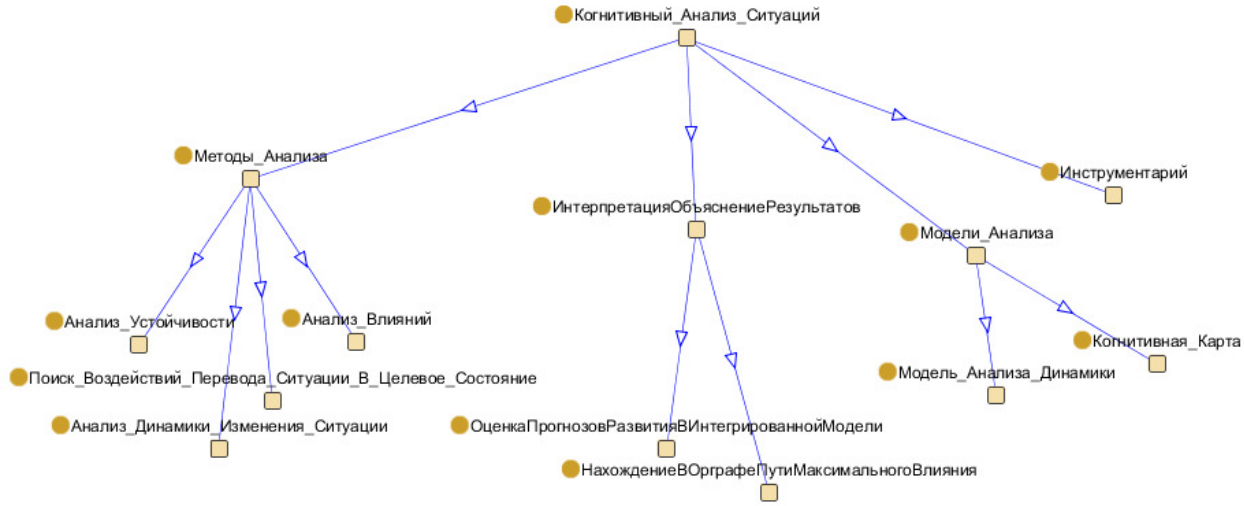

Рис. 4 Фрагмент структуры онтологии ПО КАС

используемых моделей. Разнообразие моделей когнитивных карт обусловлено различными способами задания причинно-следственных отношений и значений в вершинах-факторах [9]. ПО КАС строилась согласно метаописанию с помощью категорий, выделенных в подобласти наук об управлении *«Принятие решений»*. К метаописанию относятся классы: *модель, метод, задача, интерпретация и объяснение результатов, инструментарий, область применения.* Таксономия ПО КАС показана на Рис. 4.

Литература

1. *Hoa A. Nguyen, B.Eng.* New Semantic Similarity Techniques of Concepts Applied in theBiomedical Domain and WORDNET / Thesis Presented to the Faculty of The University of Houston- Clear Lake In Partial Fulfillment of the Requirements for the Degree Master Of Science, The University Of Houston-Clear Lake, December 2006

2. CYC - Semantic Web http://www.cyc.com

3. *Кузнецов О.П., Шипилина Л.Б.* Об онтологиях научного знания: цели, методологии построения, языки, инструментарий /"Технические и программные средства систем управления, контроля и измерения (УКИ'08): Конференция с международным участием (10-12 ноября 2008 г.,Москва, Россия) М: Учреждение Российской академии наук Институт проблем управления им. В.А. Трапезникова РАН, Электронная книга, 2008 С. 248 – 256

4. *Пронина В.А., Шипилина Л.Б.* Использование отношений между атрибутами для построения онтологий предметной области/ Проблемы управления, 2009, № 1, с. 27-32

5. *Загорулько Ю.А., Боровикова О.И., Загорулько Г.Б.* Организация содержательного доступа к гуманитарным информационным ресурсам на основе онтологий / Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

6. Michael Denny Ontology Tools Survey / http://www.xml.com/pub/a/2004/07/14/onto.html

7. Среда проектирования онтологий PROTÉGÉ http://protege.stanford.edu/doc ?

8. *F. Baader и др.* The Description Logic Handbook: Theory, Implementation, and Applications, Изд-во: Cambridge University Press, 2003, P. 555.

9. *Кулинич А.А.* Когнитивные Карты и Методы их Анализа /Одиннадцатая национальная конференция по исскуственному интеллекту с международным участием КИИ-2008 (28 сентября - 3 октября 2008г., г. Дубна, Россия): Труды конференции. Т. 3.-М.: Ленанд, 2008, С.292-299

10. Evren Sirin, Bijan Parsia SPARQL-DL: SPARQL Query for OWL-DL / http://pellet.owldl.com/papers/sirin07sparqldl.pdf

11. *Sergio Tessaris*. Questions and Answers: Reasoning and Querying in Description Logic / PhD thesis, University of Manchester, Department of Computer Science, April 2001

12. *V. Haarslev, R. Möller Racer:* An OWL Reasoning Agent for the Semantic Web, / Pro-ceedings of the International Workshop on Applications, Products and Services of Web-based Support Systems (Heldat 2003 IEEE/WIC Int'l Conf. on Web Intelligence, Halifax, Canada), 2003, P.91-95

13. Банк MedLine (PubMed) http://kodomo.fbb.msu.ru/~da_shal/term2/pubmed.html

14. *Крюков К.В., Панкова Л.А., Пронина В.А., Суховеров В.С., Шипилина Л.Б.* Меры семантической близости в онтологии. Обзор и классификация / Проблемы управления, № 5, 2010

15. *Gerd Stumme, Andreas Hotho, Bettina Berendt.* Semantic Web Mining, State of the art and future directions / Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2006, № 4, P. 124–143 или http://www.sciencedirect.com

16. *Stojanovic N., Madche A., Staab S., Studer R., Sure Y.* SEAL – A framework for developing semantic portals. / Proceedings of the first international ACM conferences on knowledge capture. – Victoria, 2001, http://www.aifb.uni-karlsruhe.de/WBS/Publ/2001/sealkcap2.pdf

17. Daniel Mahler Holistic Query Expansion using Graphical Models / New directions in Question Answering, Mark Maybury(ed.), Fall 2003, Chapter 24 http://www.cyc.com/cyc/technology/whitepapers_dir/HolisticQueryExpansion.pdf

18. *Irena Spasiгc, Goran Nenadiгc, Kostas Manios, and Sophia Ananiadou*. Supervised Learning of Term Similarities // H. Yin et al. (Eds.): IDEAL 2002, LNCS 2412, 2002, P. 429–434

19. *Maria Andrea* Rodríguez Assessing Semantic Similarity among Spatial Entity Classes / PhD thesis, The Graduate School, University of Maine, 2000

20. *Тузовский А. Ф.* Системы управления знаниями. Методы и технологии./ А. Ф. Тузовский, С. В. Чириков, В. З. Ямпольский – Томск: Изд-во НТЛ, 2005, 260 с., ISBN 5-89503-241-9.

21. Интернет-портал AIFB университета Карлсруэ (Германия) http://www.aifb.kit.edu/web/Wissensmanagement/Portal

22. Проект UMLS (Unified Medical Language System). U.S. National Library of Medicine http://www.nlm.nih.gov/mesh/meshhome.html

23. SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) - большой медицинский онтологический словарь http://www.ihtsdo.org/snomed-ct/

24. Портал NCI (американский интитут рака)

http://www.cancer.gov/cancertopics/terminologyresources

25. Портал знаний по компьютерной лингвистике http://uniserv.iis.nsk.su/cl/index.php?ent=133

26. UNSPSC ontology – www.unspsc.org

27. WordNet http://www.hum.uva.nl/~ewn/gwa.html

28. *Junichi Tsujii, Sophia Ananiadou* Thesaurus or logical ontology, which one do we need for text mining? In Language Resources and Evaluation, Springer Science and Business Media B.V., 2005, vol. 39, no 1, 77-90.

http://personalpages.manchester.ac.uk/staff/Sophia.Ananiadou/Tsujii_Ananiadou_2005.pdf

29. Eduardo Mena, Arantza Illarramendi, Vipul Kashyap, Amit P. Sheth OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies

30. Mena, E. et al. (1996), "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies". In: Proc. of the First IFCIS Int'l Conf. on Cooperative Information Systems, Brussels (Belgium), IEEE, pp. 14-25. Available at:

http://sid.cps.unizar.es/PUBLICATIONS/POSTSCRIPTS/coopis96.ps.gz

31. Goldstone R. L., Son J. Similarity / In Cambridge Handbook of Thinking and Reasoning, K. Holyrak & R. Morrison (Eds.), Cambridge: Cambridge University Press, 2005, p. 13-36

http://cognitivn.psych.indiana.edu/rgoldsto/pdfs/similarity2004.pdf