Интеграция онтологий с использованием тезауруса для осуществления семантического поиска

П.А. Ломов, М.Г. Шишаев

Аннотация. Разработан метод автоматической интеграции онтологий предметный областей на уровне соответствия. Полученный в результате интеграции тезаурус используется для трансляции запросов в термины частных онтологий, что позволяет производить семантический поиск по разнородным информационным ресурсам, представленным различными онтологиями. Разработана процедура добавления понятий новой онтологии в тезаурус, включающая анализ их семантической близости с уже содержащимися в тезаурусе и формирующая в результате взвешенную связь между ними.

Ключевые слова: онтология, семантический поиск, интеграция онтологий, тезаурус, алгоритм.

Введение

В современном мире благодаря быстрому развитию информационно-коммуникационных технологий накоплено огромное количество информации, представленной в электронном виде. Такое представление создает широкие возможности для организации автоматизированной обработки данных. При этом технологии автоматизации, изначально создававшиеся и функционировавшие на уровне технологий хранения и структур данных, сегодня во все большей степени охватывают и уровень их семантики. Проблема семантической интеграции данных предполагает решение нескольких задач, среди которых можно выделить создание семантического описания информационного ресурса, отражающего машинопонимаемое выражения смысла, и последующую интеграцию данных представлений. На сегодняшний день довольно эффективным средством явного представления семантики информационных элементов являются онтологические описания предметных областей [8]. С ростом объемов интегрируемых данных, в особенности – если их источники разнородны, возникает проблема интеграции различных онтологий для получения обобщенного семантического представления информации. В данной статье представлен метод интеграции множественных онтологий предметных областей, позволяющий формировать разделяемый общий тезаурус, который в дальнейшем используется для формулировки и трансляции пользовательских поисковых запросов в термины конкретной онтологии.

Проблематика интеграции онтологий

В современной литературе проблема интеграции онтологий обсуждается довольно широко. Уже сформулировано множество условий, определений и методов, представлены различные уровни интеграции. Но даже сейчас нет четкого соглашения о том, что включает в себя интеграция онтологий [10]. В общем, интеграцию онтологий принято определять как процесс нахождения сходства двух онтологий A и B и,

как результат, создание новой онтологии C, объединяющей и согласующей семантические представления исходных онтологий. В результате, две системы, основанные на онтологиях A и B, получают возможность взаимодействовать между собой, используя онтологию C.

Современные методы интеграции онтологий можно разделить на два типа: с замещением новой онтологией исходных (вновь созданная онтология используется вместо интегрируемых) и с совместным использованием интегрированной и исходных онтологий. Методы второго типа обладают большей гибкостью, поскольку позволяют в большей степени сохранить и в дальнейшем использовать структуру уже имеющихся онтологий. В то же время, в случае использования одной монолитной онтологии, необходимо включать в нее все термины исходных онтологий, что влечет за собой трудности, связанные с перестройкой связей с уже имеющимися терминами и разрешением семантических конфликтов. По этой причине методы первого типа оказываются применимы лишь в том случае, когда набор интегрируемых онтологий известен заранее и его расширение не предполагается [18].

Выделяют различные уровни интеграции онтологий в зависимости от числа изменений, которые необходимо сделать, чтобы получить некую общую онтологию из частных [12].

- Соответствие (alignment). Соответствие есть отображение понятий и отношений одной онтологии на другую. Соответствие может быть определенно не полностью, так, может существовать несколько понятий в одной онтологии, не имеющих своих эквивалентов в другой. Иногда для приведения онтологий в соответствие в них добавляют новые подклассы и надклассы понятий. Никаких других изменений аксиом, определений, доказательств или вычислений не производится.
- Частичная совместимость (partial compatibility). Частичная совместимость есть соответствие онтологий, которое поддерживает также эквивалентные выводы и вычисления для всех эквивалентных понятий и отношений. Если две онтологии являются частично совместимыми, то любой вывод или вычисление, которые могут быть выражены в одной онтологии с использованием только соответствующих понятий и отношений, могут быть транслированы в эквивалентный вывод или вычисление в другой онтологии.
- Унификация (unification). Унификация есть взаимнооднозначное соответствие всех понятий и отношений в двух онтологиях, которое позволяет любой процесс вывода или вычислений, выраженных в одной онтологии, отображать в эквивалентный процесс вывода или вычислений в другой. Обычным способом унификации двух онтологий является усовершенствование каждой из них в более детальные онтологии, чьи категории взаимнооднозначно эквивалентны.

Самой слабой формой интеграции является соответствие, так как она требует минимальных изменений исходных онтологий, но может поддерживать глубокие выводы и вычисления. Частичная совместимость требует больших изменений, при этом она обеспечивает более широкую способность к взаимодействию. Унификация или полная совместимость требуют, как правило, значительных изменений или, в некоторых случаях, полной перестройки исходных онтологий, но ее результатом является наиболее полная способность к взаимодействию, то есть все, что может быть сделано в одной онтологии, может быть сделано полностью эквивалентным способом в другой.

Сегодня существует множество средств, направленных на решение задачи отображения онтологических понятий, среди которых можно выделить [15]:

- PROMPT - представляет собой подключаемый модуль к редактору онтологий Protege. Процесс работы системы заключается в анализе онтологий и представлении эксперту списка предложений по объединению с последующей корректировкой списка в результате выбора того или иного действия. При генерации набора предложений система учитывает схожесть символических имен понятий, а также прямые таксономические связи между ними;
- Chimaera - компонент сервера Ontolingua, выполняющий задачи объединения онтологий и выявления последующих логических несогласованностей. При генерации списка своих предложений Chimaera основывается на сход-

стве имен понятий, их определений в онтологиях, а также акронимов, учитываются прямые таксономические связи;

• ONION в результате анализа онтологий предлагает пользователю набор правил объединения. Сам анализ включает две стадии: лингвистическую, в ходе которой определяются сходства названий терминов, а также их определений, полученных из стороннего словаря, и структурную, на которой производится оценка близости понятий на основе их отношений со схожими терминами. Эксперт в итоге определяет применение тех или иных правил, предлагающих различные разрешения неоднородностей в соответствии с их эвристической оценкой.

Данные системы требуют привлечения эксперта на каком-либо из этапов своей работы для разрешения спорных ситуаций. Это связанно с недостаточным отражением семантики термина в его формальном представлении в онтологии, что, в свою очередь, приводит к необходимости использования эвристических методов сравнения понятий, результаты которых могут быть не точны. Поскольку целью указанных систем является получение непротиворечивой результирующей онтологии, неточности обязательно должны быть устранены экспертом. Однако в случае решения задачи семантического поиска достаточно интеграции на уровне соответствия. При этом можно обойтись без привлечения эксперта, так как различные противоречия и неоднозначности могут быть разрешены пользователем в ходе процедуры задания поискового запроса или анализа результатов. Кроме этого, предлагаемый подход не требует замены отдельных онтологий единой, являющейся результатом объединения, и поэтому не требует внесения изменений в алгоритмы приложений, работающих с онтологиями отдельных информационных ресурсов.

Обобщенная архитектура системы семантического поиска

Рассмотрим систему семантического поиска, которая будет использовать полученный в результате интеграции онтологических представлений тезаурус.

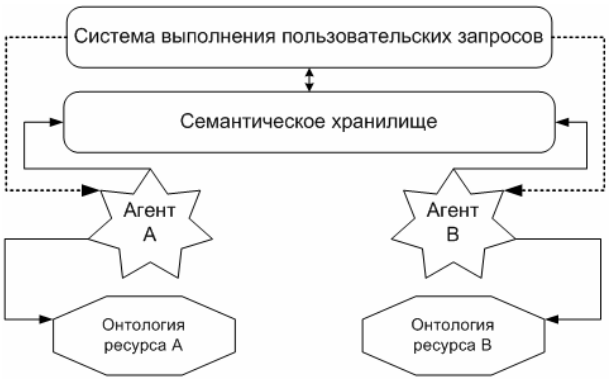

Рис.1. Обобщенная архитектура системы семантического поиска

Обобщенная архитектура системы (Рис. 1) включает следующие функциональные компоненты:

1. Семантическое представление ресурса (онтология ресурса) - хранит машинопонимаемую метаинформацию, отражающую семантику реального информационного ресурса;

2. Агент – программа, выполняющая представление определенной онтологии в семантическом хранилище. Агент выполняет поиск в соответствующей ему онтологии, передает результаты выполнения запроса в семантическое хранилище. В его задачи также входит предоставление понятий онтологии и их свойств для включения в тезаурус;

3. Система выполнения пользовательских запросов – осуществляет трансляцию запроса в понятия конкретной онтологии, производит отбор агентов в семантическом хранилище для выполнения запроса, генерирует задания для агентов и выполняет активизацию агентов, осуществляет представление результатов задания;

4. Семантическое хранилище - служит для взаимообмена информацией между агентами, осуществляет кэширование популярных запросов, хранит информацию об агентах и их задачах и результатах работы агентов, также включает тезаурус, представляющий обобщенное семантическое представление всех частных онтологий.

Важной частью семантического хранилища является тезаурус. В данном случае его можно определить как набор элементов типа «Объект», которые отражают понятие той или иной онтологии связаны между собой отношениями

синонимии, антонимии, гипонимии, гиперонимии, ассоциации. С каждым из этих элементов сопоставляются элементы тезауруса типа «Свойство», представляющие соответствующие атрибуты определенного понятия онтологии.

Таким образом, используя семантические связи между понятиями, можно использовать тезаурус для выполнения трансляции запроса с сохранением семантики в несколько вариантов, ориентированных на различные онтологии.

Разрешение конфликтов

Добавление информационного источника в пространство интеграции, как правило, сопровождается созданием для него нового онтологического описания, отражающего его семантику. Вследствие этого важной проблемой является расширение тезауруса, которое предполагает, как включение нового понятия, так и связывание его с уже имеющимися в тезаурусе. В ходе этого процесса могут возникнуть конфликты, связанные с различной точкой зрения на определенное понятие в различных онтологиях, степенью детализации определений и.т.п. Различные виды конфликтов и способов их разрешения описаны в [3, 4]. Следует отметить, что поскольку тезаурус будет использоваться для формулировки пользовательского запроса и его последующей трансляции в запросы к частным онтологиям для поиска конкретных экземпляров, можно ограничиться разрешением конфликтов на уровне концептов.

Рассмотрим один из конфликтов более подробно и выделим условия, которыми необходимо руководствоваться при его разрешении. Рассмотрим части двух онтологий, описывающие две схожие предметные области (Рис. 2).

В онтологии А понятия «Организация» и «Человек» конкретизируют понятие «Хозяин», а в онтологии В понятия «Владелец» и «Арендатор» конкретизируют понятие «Персона». Если обратиться к понятиям «Владелец» и «Хозяин» в обеих онтологиях, то можно увидеть, что, несмотря на то, что сами слова - синонимы, их интерпретации концептов совершено разные. Такое отличие появляется в силу разного структурного положения данных терминов в онтологиях. Поэтому было бы неверно при добавлении в тезаурус понятий онтологий А и В одновременно связывать понятие «Персона» с понятием «Владелец» связью «имеетПодкласс», а затем понятие «Владелец» с понятием «Человек» связью «имеетПодкласс» (Рис 3.).

К тому же выводу можно прийти, если сопоставить множества необходимых и достаточных атрибутов понятий различных онтологий.

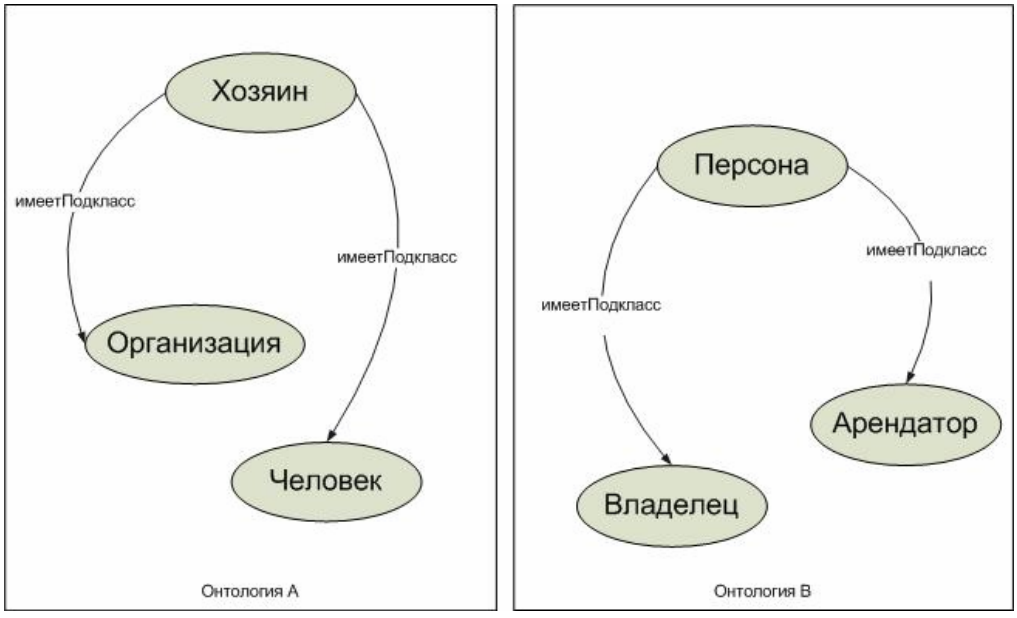

Рис.2. Пример разной трактовки смысла термина

Рис.3. Пример семантического конфликта

Пусть в онтологии А понятие «Человек» обладает следующим множеством необходимых (NES_A) и достаточных (SUF_A) условий:

$$NES_A = \{"имеетИмя", "имеетИндивидуальныйНалоговыйНомер"\},$$

$$SUF_A = \{"владеет"\},$$

а одноименное понятие в онтологии В соответственно:

$$NES_B = \{"имеетИмя", "имеетПол", "датаРождения"\}$$

$$SUF_B = \{"владеет"\}.$$

В данном примере множества необходимых свойств имеют лишь один одинаковый элемент, а остальные три отличаются, что свидетельствует об их достаточно малом семантическом сходстве.

Таким образом, в качестве семантической метрики, характеризующей степень сходства понятий, можно использовать следующие три оценки:

- сходство семантики символических имен терминов;
- структурное положение понятия в онтологии;
- степень сходства множеств необходимых и достаточных атрибутов.

Нетрудно видеть, что использование в качестве критерия идентичности понятий только одной из перечисленных оценок с высокой степенью вероятности приводит к семантическим конфликтам в процессе интеграции онтологий. По этой причине в данной работе предлагается в качестве метрики семантической близости использовать все три оценки в комплексе.

Общий вид процесса интеграции в тезаурус

Рассмотрим в общем виде процесс сопоставления терминов различных онтологических описаний. Онтологию предметной области можно представить в виде направленного графа, в котором отдельные вершины, соответствующие понятиям предметной области, связаны между собой направленными дугами, представляющими различные связи между понятиями. Основными связями онтологии являются: гипонимия (отношение между подмножеством и надмножеством) и классификация (отношения между множеством и его элементом). Если принять во внимание только данные виды связей, то онтологию можно представить в виде дерева, корневой вершиной которого представляется понятие «Сущность», являющееся наивысшей абстракцией для всех понятий. С корневой вершиной связаны вершины, представляющие наиболее общие понятия данной онтологии, причем каждая из этих вершин является корнем своего поддерева, на каждом уровне которого общее понятие конкретизируется. Разумеется, конкретизацию понятия можно проводить несколькими способами, с разной степенью детализации, уровень которой диктуется требованиями конкретной предметной области и потенциальными задачами, в которых будет использоваться онтология.

Таким образом, можно рассматривать все интегрируемые онтологии как совокупность поддеревьев, в вершинах различных уровней которых, возможно, находятся сходные, с точки зрения семантики, понятия (Рис 4.).

Соответственно, проведение интеграции онтологий предполагает: создание новых ветвей

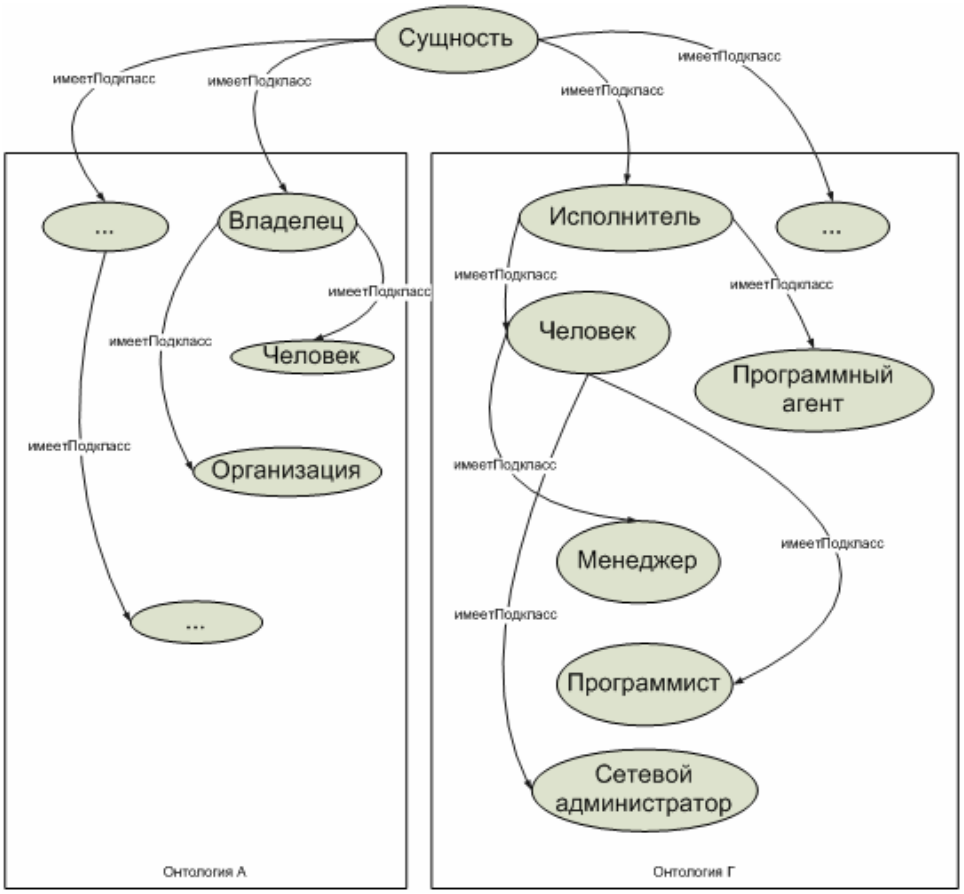

Рис. 4. Общее представление онтологий

для конкретизации нового общего понятия или добавление новой ветви на какой-либо уровень в уже существующем дереве для определения новой конкретизации. Такой подход позволяет избежать семантических конфликтов при объединении онтологий и создать интегрированную онтологию, универсальную в смысле требований к точности семантического соответствия терминов: одно приложение может расценивать одну и ту же пару понятий как идентичные, тогда как другое – считать их различными. Это зависит от заданных (в приложениях, но не в онтологии) требований к минимальной величине семантической близости понятий.

Формальные представления объектов

В данном разделе даются формальные определения терминов, используемых в процедуре расширения тезауруса.

Определим объект онтологии, представляющий некоторое понятие, в виде следующей совокупности [0]:

$$T =< N_t, P_t, S_t, D_t, I_t >, \quad (1)$$

где N_t – наименование объекта T, P_t – множество свойств объекта T, S_t – множество суперклассов объекта T, $S_t \subseteq T_U$, D_t – множество подклассов объекта T, $D_t \subseteq T_U$, I_t – множество экземпляров объекта T, T_U - множество всех объектов онтологии.

Дадим формальные определения элементов тезауруса, хранящегося в семантическом хранилище. Понятие некоторой предметной области будет представляться в тезаурусе соответствующим ему элементом тезауруса типа «Объект», который можно представить в виде следующей тройки:

$$O =< N_O, L_O, A_O >, \quad (2)$$

где N_O – символьное имя объекта O, соответствующее названию представляемого им понятия, L_O – множество связей, в которых состоит объект O, A_O – множество агентов, использующих данное понятие в представляемых ими онтологиях.

Связь между объектами тезауруса представим в виде четверки:

$$L =< TP_l, O_1, O_2, W_1 >, \quad (3)$$

где TP_l – тип связи L, $TP_o \in TP_U$, O_1- первый объект, входящий в связь, $O_1 \in O_U$, O_2- второй объект, $O_2 \in O_U$, входящий в связь, W – вес связи $(W \in \mathrm{N}) \& (0 \le W \le 100)$, O_U- множество всех объектов тезауруса.

Множество типов связей между объектами, представляющими термины, в тезаурусе:
$TP_U = \{synonymOf, hyponymOf, associateWith\}$.

Атрибут объекта онтологии предметной области будет представлен в тезаурусе соответствующим элементом типа «Свойство», которое представим в виде тройки:

$$P =< N_P, O, A_P >, \quad (4)$$

где N_P– символьное имя свойства P, соответствующее наименованию атрибута объекта онтологии предметной области, O – объект тезауруса, который характеризует данное свойство, $O \in O_U$, A_P – множество агентов, использующих соответствующий атрибут в представляемых ими онтологиях.

Синонимы понятия предметной области представляются множеством объектов, вступающим в соответствующую связь с объектом, отражающем данное понятие в тезаурусе. Так, если O является объектом тезауруса, соответствующим некоторому понятию, то множество его синонимов определим следующим образом:

$$SYN_O = \{O_i \mid i \in \mathrm{N}\}, \quad (5)$$

где для каждого O_i существует

$L_1 =< synonymOf, O_k, O_i, W >$

и $L_2 =< synonym, O_i, O_k, W >$.

Определим также множество непосредственных гипонимов (6) объекта тезауруса O, которое соответствует множеству прямых потомков (7) соответствующего объекта онтологии - T:

$$HYPT_O = \{O_i \mid i \in \mathrm{N}\}, \quad (6)$$

где для каждого O_i существует

$$L_0 =< hyponymOf, O_i, O_k, W >,$$

$$HYPO_T = \{T_i \mid i \in \mathrm{N}\}, \quad (7)$$

где для каждого $T_i : S_{Ti} = \{G\}, T_i \in D_G$.

Определим функцию семантического сопоставления имен объектов, которая будет принимать объект онтологии и элемент тезауруса в качестве аргументов и возвращать степень сходства семантики символических имен:

$$Syneq(O,T) = x \quad (8)$$

где $O, T \in O_U \cup T_U$, N_O и N_T – символические имена соответствующих объектов или понятия, $0 \le x \le 100$.

Необходимо заметить, что данная функция включает такие методы, как сравнения токенов имен терминов, определения расстояния между ними, близости определений терминов, сравнения синонимов терминов. Ее эффективность и точность можно повысить, подключив сторонний тезаурус предметной области, к которой относится онтология.

Введем предельные значения функции (8). Если значение функции превышает предельное значение, то два ее термина-аргумента считаются синтаксически эквивалентными:

$$1 \le UPSYN \le 100, \quad (9)$$

если $Syneq(O,T) \ge UPSYN$, то $N_O = N_T$

Оценка сходства положений в иерархии терминов и множеств необходимых и достаточных свойств будет осуществляться функциями (10) и (11) соответственно:

$$Poseq(O,T) = x \quad (10)$$

$$Atreq(O,T) = x \quad (11)$$

где $O, T \in O_U \cup T_U$, $0 \le x \le 100$.

Процедура расширения общего тезауруса

Рассмотрим процедуру добавления нового термина в тезаурус по шагам:

Шаг 1. Зададим начальные значения переменных-счетчиков: $n=1$, $k=1$, $l=1$, $u_l=1$. Пусть начальными для рассмотрения объектом тезауруса - TRT и понятием онтологии - TRO будут соответственно объект и понятие «Сущность»:

$$TRO =<"Сущность",\varnothing,\varnothing,D_U,L_U>,$$

$$TRT =<"Сущность",L_U,A_U>,$$

Переходим к шагу 2.

Шаг 2. С помощью функции семантического сопоставления имен (8) производим сравнения имени каждого элемента множества гипонимов $HYPO$ объекта TRO с именем каждого элемента множества гипонимов $HYPT$ объекта TRT. Понятия онтологии, для которых функция сопоставления возвращала значение, превышающее пороговое (9), формируют множество EQ_l, остальные понятия попадают в множество NEQ_k:

$$EQ_l = \{T_i \mid i \in N\}, \quad (12)$$

где для каждого

$$T_i : (T_i \in HYPO_{TRO}) \,\&\, (\exists TA_j : (TA_j \in HYPT_{TRT}) \,\&$$
$$\&\, (Syneq(T_i, TA_j) > UPSYN)), i \in \mathrm{N},$$

$$NEQ_k = \{T_i \mid i \in N\}, \quad (13)$$

где для каждого

$$T_i : (T_i \in HYPO_{TRO}) \,\&\, (\exists TA_j : (TA_j \in HYPT_{TRT}) \,\&$$
$$\&\, (Syneq(T_i, TA_j) \le UPSYN)), i \in \mathrm{N}$$

Переходим к шагу 3.

Шаг 3. Если $n > |NEQ_k|$, тогда переходим к шагу 3.3, иначе создаем в тезаурусе элемент типа «Объект» - P_n, соответствующий понятию T_n: $T_n \in NEQ_k$ и переходим к шагу 3.1.

Шаг 3.1. С помощью функции (8) производим сопоставление T_n со всеми объектами тезауруса. Если функция возвратила значение, превышающее пороговое (9), для каких-либо двух аргументов, то производиться сравнительная оценка их положения в иерархии и множеств их атрибутов с помощью функций (10) и (11). В итоге в тезаурусе между созданным элементом P_n и элементом, отобранным с помощью функций, создается ассоциативная связь с весом, равным среднему арифметическому трех оценок:

$$L =< associateWith,\ P_n,\ F,\ W >,$$

где $P_n, F \in O_U$ и для

$$P_n, F : Syneq(P_n, F) \ge UPSYN,$$

$$W = (Syneq(P_n, F) + Poseq(P_n, F) + Atreq(P_n, F))/3.$$

Переходим к шагу 3.2.

Шаг 3.2. Создаем связи гипонимии L объекта P_n с объектами в тезаурусе, соответствующим его суперклассам в онтологии:

$$L =< hyponymOf,\ P_n,\ S,\ 100 >,$$

где S представляет понятие онтологии, являющееся суперклассом для понятия T_n.

Далее инкрементируем счетчик n, переходим к шагу 3.

Шаг 3.3 Формируем новое множество NEQ_{k+1}, состоящее из понятий, являющихся непосредственными подклассами понятий из множества NEQ_k:

$$NEQ_{k+1} = \{T_i \mid i \in N\},$$

где $T_i \in HYPO_H$, $H \in NEQ_k$

Далее инкрементируем счетчик k, счетчик n сбрасываем в единицу. Переходим к шагу 3.4.

Шаг 3.4. Если $NEQ_k \neq \varnothing$, то переходим к шагу 3, иначе переходим к шагу 4.

Шаг 4. Если $l = 0$, то завершаем процедуру, иначе переходим к шагу 5.

Шаг 5.Если $u_l > |EQ_l|$, то декрементируем счетчик l, инкрементируем счетчик u_l, переходим к шагу 4, иначе добавляем агента, представляющего интегрируемую онтологию, к множеству агентов-представителей элемента тезауруса H, который признан синтаксически эквивалентным понятию этой онтологии T_{ul}, $T_{ul} \in EQ_l$:

$$H : Syneq(T_{\mathrm{ul}}, H) \ge UPSYN$$

Переходим к шагу 6.

Шаг 6. Устанавливаем в качестве новых объектов для рассмотрения элемент тезауруса H и соответствующие ему понятия онтологии T_{ul}:

$TRO = T_{ul}$, где $T_{ul} \in EQ_l$

$TRT = H$, где $H : Syneq(T_u, H) \geq SYNEQ$.

Переходим к шагу 7.

Шаг 7. Инкрементируем счетчик l, сбрасываем u_l в единицу, переходим к шагу 2.

Основная идея алгоритма состоит в формировании новой ветви дерева терминов тезауруса, исходящей из вершины-корня, обозначающей предельную абстракцию «Сущность», если в тезаурусе отсутствует термин, вершина которого непосредственно связанна с корневой и который сопоставим с понятием онтологии, также непосредственно связанной с понятием «Сущность». В противном случае вершины ветвей сливаются, и далее по такому же принципу рассматриваются вершины следующего уровня. В случае разного положения в иерархии терминов на каком-либо уровне ветви расходятся.

Пример работы алгоритма

Расширение тезауруса начальной онтологией является тривиальным, поэтому будем полагать, что в тезаурус уже включены входящие в нее термины (Рис. 5, а). Рассмотрим процедуру расширения тезауруса на примере добавления в тезаурус новой онтологии (Рис. 5, б).

В начале работы алгоритма гипонимами термина «Сущность» в тезаурусе будут: «персона», «документ», «ребенок», а гипонимами понятия «Сущность» в онтологии – «сотрудник» и «документ». В ходе сравнения семантики имен терминов с помощью функции (8) во множество схожих понятий онтологии попадет – «документ», а несхожих – «сотрудник».

Далее будут обработаны элементы множества несхожих понятий. В данном случае оно состоит из одного элемента – «сотрудник», который помещается в тезаурус, а далее с помощью функции (8) будет сравниваться с терминами тезауруса. При достижении термина «работник», являющегося синонимом термина «сотрудник», оба будут переданы в функции (10) и (11). В зависимости от результирующей оценки между ними будет создана или связь ассоциации с набранным весом, или связь синонимии. Далее аналогичным образом обрабатываются гипонимы термина «сотрудник» и остальные элементы множества несхожих терминов, если таковые имеются. После этого будет обработан единственный элемент множества схожих терминов – «документ». Будут вновь определены: множество гипонимов термина «документ» в тезаурусе, состоящее из элементов: «свидетельство», «запись акта», «заявление», и множество его гипонимов в онтологии: «штатное расписание», «служебная записка», «акт». В ходе сравнения с помощью функции (8) терминов данных множеств будет сформировано множество непохожих терминов, состоящее из элементов: «штатное расписание», «служебная записка», «акт», а множество похожих в данном случае будет пустым. По рассмотренному ранее принципу будет обработан каждый элемент множе-

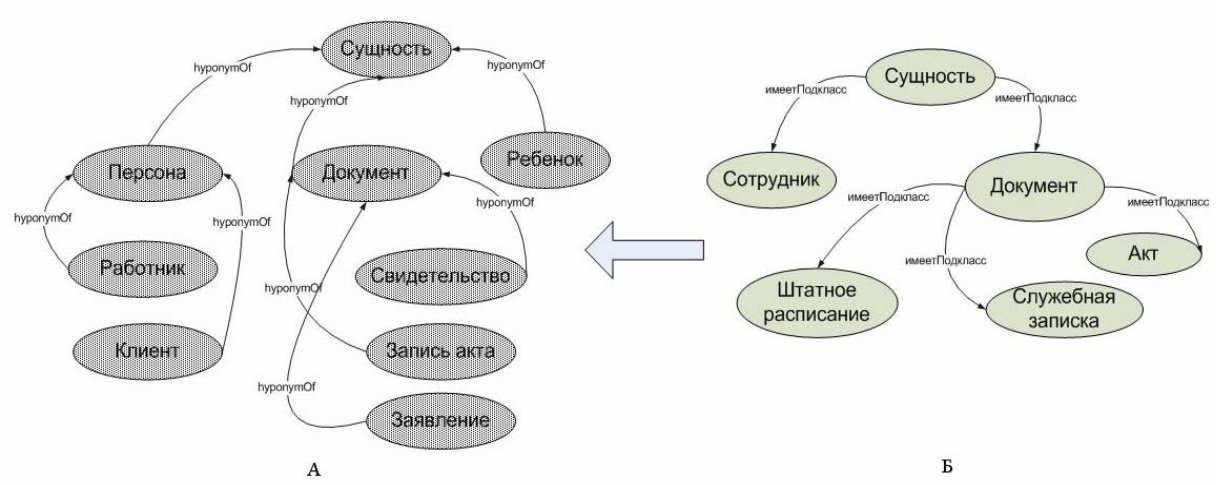

Рис. 5. Расширение тезауруса терминами новой онтологии

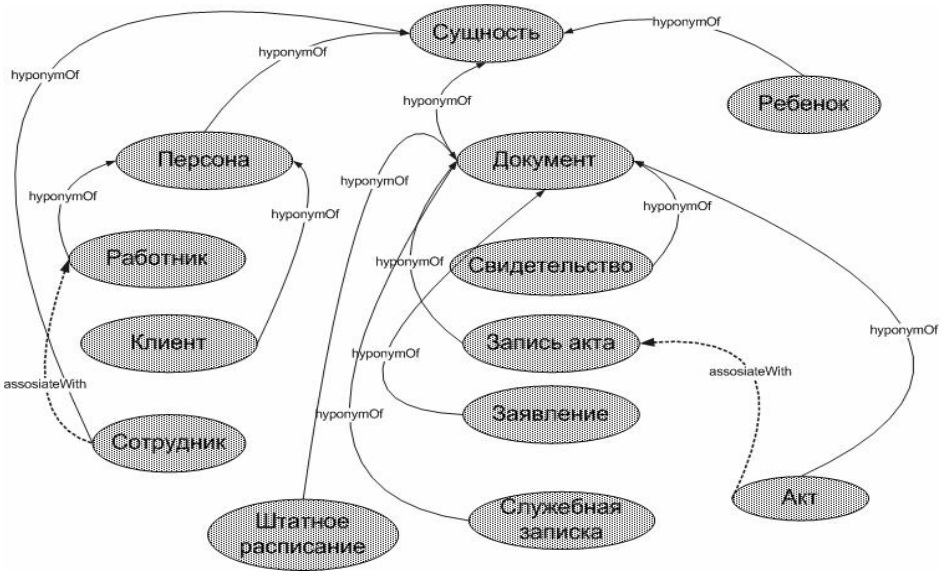

Рис. 6. Тезаурус в результате работы алгоритма

ства непохожих терминов, в результате чего все они будут включены в тезаурус, а между новым термином «акт» и «запись акта» будет создана связь ассоциации. Результирующий вид тезауруса представлен на Рис. 6.

Вероятно, при практическом применении данный алгоритм может привести к порождению ветвей, по-разному уточняющих некоторый общий термин или имеющих в качестве своих начальных вершин семантически схожие наименования терминов. Однако это позволяет избежать появления кардинальных противоречий в семантической трактовке схожих с какой-либо точки зрения терминов.

Необходимо также отметить, что соединение некоторых вершин ветвей путем установления связи «ассоциация» с определенным весом позволяет представлять пользователю для поиска понятия, ассоциированные с тем, которое он выбрал, но употребляемое в несколько другом смысле. Каждый факт выбора пользователем ассоциированных друг с другом терминов для поиска будет увеличивать вес определенной связи, вес же связей с другими терминами, которые не были выбраны, будет уменьшаться. Впоследствии, при достижении весом связи верхнего порогового значения, данный термин будет выбираться системой для поиска автоматически, тогда как достижение нижнего порога будет вызывать удаление данной связи. Конкретные значения, на которые будет уменьшаться или увеличиваться вес связи, а также пороговые предполагается определить экспериментально.

Заключение

В статье рассматриваются некоторые аспекты довольно сложной и актуальной проблемы интеграции онтологических описаний предметных областей. Обозначены основные факторы, которые необходимо учитывать при сопоставлении терминов различных онтологий, такие как сходство символических имен, положение в иерархии терминов и набор необходимых и достаточных свойств.

Представлен показатель семантического сходства терминов двух различных онтологий, значение которого используется в процедуре добавлении новых терминов в тезаурус для начальной оценки веса ассоциативной связи. Впоследствии значение веса корректируется в соответствии со статистикой учета данной связи пользователем при выборе схожих понятий для поиска.

Рассматривается процедура расширения тезауруса. Ее можно интерпретировать как способ интеграции онтологий на уровне соответствия, что позволяет впоследствии использовать тезаурус для конструирования и трансляции пользовательского запроса при произведении поиска с учетом семантики.

Литература

1. Ломов П. А., Шишаев М. Г. Интеграция семантически связанных информационных ресурсов на основе онтологий для эффективного информационного обеспечения рационального природопользования / П. А. Ломов, М. Г. Шишаев // Глубокая переработка минеральных ресурсов: Сборник материалов IV школы молодых ученых и специалистов «Сбалансированное природопользование» (6-8 ноября 2007 г.) – Апатиты: Изд-во КНЦ РАН. 2008. – С.243-247
2. Е.А. Жыжырий, С.С. Щербак «Математическое обеспечение систем поиска, основанных на онтологиях» – Режим доступа: http://shcherbak.net/mat_obez/printpage
3. Т.В. Левашова, М.П. Пашкин, А.В. Смирнов, Н.Г. Шилов. Управление онтологиями. Часть II // Известия академии наук. Теория и системы управления. №5, 2003, С.89 -101
4. Смирнов А.В., Пашкин М.П., Шилов Н.Г., Левашова Т.В. Онтологии в системах искусственного интеллекта: способы построения и организации// Новости искусственного интеллекта.
5. D. Bianchini, V. De Antonellis "Ontology-based Integration for Sharing Knowledge over the Web" – Режим доступа: http://www.doc.ic.ac.uk/~pjm/diweb2004/DIWeb2004_Part8.pdf
6. C. Maria (Marijke) Keet «Aspects of Ontology Integration» – Режим доступа: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/7-pinto.pdf
7. M. Petrenko, H. Jamil "Query Ontologies for Autonomous Online Resource Integration Systems" – Режим доступа: http://www.cs.wayne.edu/~max/docs/QueryOntologies.pdf
8. А.Н. Бездушный, А.М. Меденников, В.А.Серебряков – «Подход к интеграции информационных коллекций в ИСИР РАН» – Режим доступа: http://dbserv.ihep.su/~pubs/aconf00/dconf00/ps/056.pdf
9. Клещев А.С., Артемьева И.Л. Математические модели онтологий предметных областей. Часть 3. Сравнение разных классов моделей онтологий. // Научно–техническая информация, Сер. 2. Информационные процессы и системы, № 4, 2001, С. 10–15.
10. Л.А.Калиниченко, «Методология организации решения задач над множественными распределенными неоднородными источниками информации» - Режим доступа: http://synthesis.ipi.ac.ru/synthesis/publications/itedu05/itedu05.pdf
11. Н. А. Скворцов, «Применение уточнения понятий в решении задач манипулирования онтологиями» - Режим доступа: http://rcdl2007.pereslavl.ru/papers/paper_70_v1.pdf
12. M. Ehrig, York Sure «FOAM – Framework for Ontology Alignment and Mapping Results of the Ontology Alignment Evaluation Initiative» - Режим доступа: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-156/paper11.pdf
13. М.Р. Когаловский «Тенденции развития технологий управления информационными ресурсами в электронных библиотеках» - Режим доступа: http://www.rcdl2006.uniyar.ac.ru/papers/paper_98_v1.pdf
14. Wache H., and other. Ontology-Based Integration of Information. A Survey of Existing Approaches. Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing, 2001. Режим доступа: http://www.cs.vu.nl/~heiner/public/ois-2001.pdf
15. Скворцов Н.А. Вопросы согласования неоднородных онтологических моделей и онтологических контекстов// Онтологическое моделирование/ Под ред. Л.А. Калиниченко. - М.: ИПИ РАН, 2008. -С. 149-166.

Ломов Павел Андреевич. Стажер-исследователь лаборатории региональных информационных систем Института информатики и математического моделирования технологических процессов Кольского научного центра РАН. Окончил Петрозаводский государственный университет (2006 г.). Автор 5 печатных работ. Область научных интересов: технологии Semantic Web, разработка и использование онтологий предметных областей. E-mail: lomov@iimm.kolasc.net.ru/.

Шишаев Максим Геннадьевич. Заведующий лабораторией региональных информационных систем Института информатики и математического моделирования технологических процессов Кольского научного центра РАН. Окончил в 1993 году Санкт-Петербургский государственный технический университет (Политехнический институт). Кандидат технических наук. Автор более 80 печатных работ. Область научных интересов: распределенные информационные системы, проблемно-ориентированные информационные системы, системы информационной поддержки инноваций. E-Mail: Shishaev@iimm.kolasc.net.ru.