Разработка метода семантической интеграции информации в сфере государственного и муниципального управления♣

©Ломов П. А. Шишаев М. Г.

Институт Учреждение Российской академии наук Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН
lomov@iimm.kolasc.net.ru, shishaev@iimm.kolasc.net.ru

Аннотация

В данной работе предлагается подход к семантической интеграции данных в сфере государственного и муниципального управления с использованием разделяемого тезауруса, который позволяет устранить критичные для данной предметной области недостатки, присущие существующим подходам к интеграции. Представлена концептуальная модель тезауруса, механизм отображения в него онтологий, а также методика задания и сопоставления онтологических контекстов на основе набора общих атрибутов.

1 Введение

Проблема интеграции информации, заключающаяся в предоставлении единой точки доступа к распределенным и гетерогенным информационным ресурсам, характерна для многих предметных областей и сфер деятельности человека, в том числе и для сферы государственного или муниципального управления. Особенную актуальность данная проблема приобрела в настоящее время вследствие формирования так называемого электронного государства, которое предполагает создание разветвленной коммуникационной инфраструктуры, позволяющей государственным органам и гражданам взаимодействовать с использованием новых информационных технологий[1].

Целью работы является разработка метода интеграции данных, а также моделей информационных систем и программных средств, позволяющих производить интеграцию информации в сфере государственного и муниципального управления на основе ее семантики.

Труды 11ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

2 Проблематика семантической интеграции информации

2.1 Применение онтологий для формального отражения семантики

Традиционно для представления знаний об определенной предметной области использовались такие формализмы, как семантические сети, фреймы. Однако, при всей своей наглядности такое представление не позволяло формально отразить значение того или иного термина или отношения, делая невозможным обработку таких знаний с помощью ЭВМ.

Оперирование семантикой стало возможным благодаря появлению и развитию моделей представления знаний, позволяющих в формальном виде отразить смысл некоторого информационного элемента. Это, в свою очередь, позволяло в определенной степени производить машинную обработку информации, подобно эксперту. Среди таких моделей можно выделить онтологии - формальные спецификации разделяемой концептуализации.

Способность определения формального смысла появилась, благодаря использованию дескриптивных логик, которые лежат в основе многих распространенных языков описания онтологий, таких как, OIL, DAML+OIL, OWL-DL, OWL-LITE.

Онтология – спецификация концептуализации [5], или явное, формальное описание предметной области. Онтологию можно представить в виде упорядоченной тройки конечных множеств:

$$O=\langle T, R, F\rangle, \quad (1)$$

где:

T — термины предметной области, которую описывает онтология O;

R — отношения между терминами заданной предметной области;

F — функции интерпретации, определенные на терминах и отношениях онтологии O, имеющие следующий вид:

$$I(t) \subset dom, \quad (2)$$

$$I(R) \subset dom \times dom, \quad (3)$$

где t – термин предметной области, dom – множество объектов реального мира, $I \in F$.

Следует заметить, что функции интерпретации, явно в онтологии не присутствуют, эксперт, добавляя в онтологию аксиомы, определяет ограничения на интерпретации терминов и отношений, в соответствии со своим пониманием их смысла. Полученная в итоге совокупность аксиом и определяет формальную семантику элементов онтологии.

Для описания онтологий существует несколько языков, отличающихся выразительностью, наличием возможности полного логического вывода. Однако при разработке современных информационных систем является более предпочтительным использовать языки или технологии, прошедшие стандартизацию и рекомендованные к применению в промышленных проектах. Примерами таких технологий могут служить технологии Semantic Web, такие как RDF(Resource Definition Framework)[9], DAML+OIL, OWL(Ontology Web Language)[6], OWL2[7]. Основным назначением данных языков является формальное описание семантики данных в виде совокупности объектов и отношений между ними, что, в свою очередь, позволяет производить интеграцию информации, руководствуясь ее смыслом, а не форматом представления.

2.2 Рассмотрение существующих подходов к интеграции и их применения в области государсвенного и муниципального управления

При выборе того или иного интеграционного подхода необходимо учитывать специфические особенности предметной области, что может в определенной степени облегчить проведение интеграционных процессов.

В данном случае одной из таких особенностей является то, что большинство определений как субъектов и объектов, а также процедур, ситуаций, находит свое отражение в различных документах, которым должно следовать то или иное государственное образование. Данное обстоятельство облегчает задание формальных моделей как различных сущностей, так и взаимодействий между ними. Однако в тоже время существует тенденция внесения различных изменений в правовые документы, что должно находить свое отражение в изменениях моделях определенных сущностей.

Также в рассматриваемой предметной области данные об определенной сущности не хранятся централизованно, а разбросаны по различным источникам. Причем добавление новой информации или изменение существующей, должно происходить с учетом значений в других источниках, в противном случае может быть нарушена семантическая целостность информации об объекте.

Наряду с этим, немаловажной является обеспечение возможности гибкого регулирования доступа к атрибутам объекта в зависимости от задачи и решающего ее ведомства.

Исходя из данных особенностей, можно использовать централизованный подход к интеграции, который заключается в создании единой онтологии, постулирующей формальную семантику терминов предметной области. Однако разработка и поддержка достаточно объемной онтологии с большим количеством аксиом, является очень трудной задачей, и любая модификация будет требовать привлечения как экспертов предметной области, так и инженеров по знаниям. С ростом онтологии также появиться проблема, связанная с ее практическим использованием для определения семантики добавляемых информационных ресурсов, чей набор терминов не будет находить точного отображения в концептах онтологии. Данные недостатки отсутствуют у децентрализованного подхода, который не накладывает ограничений на создание онтологий отдельных информационных ресурсов, что позволяет точно отразить формальную семантику конкретных терминов. Его применение для рассматриваемой предметной области можно увидеть в работе[3]. Однако он требует определения способов отображения между концептами и отношениями различных онтологий, что является нетривиальной задачей. Также довольно сложным становиться централизованное установление прав доступа и получение совокупной информации об объекте из разных источников.

Неким компромиссом между рассмотренными подходами может являться гибридный подход. В сравнении с централизованным, он позволяет обеспечить гораздо большую выразительность при создании частных онтологий информационных ресурсов и, как следствие, более точное отражение семантики терминов. Наряду с этим, в отличие от децентрализованного, существенно облегчается задача установления различных отношений с терминами отдельных онтологий.

Перечисленные положительные свойства обеспечиваются благодаря использованию общего словаря, на основании которого строятся частные онтологические описания. Однако само построение словаря может производиться разными способами. Так, например, в работе[11] в разделяемом словаре содержатся термины-примитивы, которые, комбинируясь друг с другом, формируют лейблы описывающие отдельные концепты. Вследствие этого появляется возможность производить автоматизированное сравнение концептов, исходя из описывающих их лейблов. Недостаток этого подхода заключается в том, что выразительность описания того или иного информационного

элемента ограничивается выразительной мощностью общего словаря, и в ряде случаев приводит к усреднению семантических описаний. При расширении же словаря появляется проблема задания одного и того же лейбла, с помощью различных комбинаций терминов-примитивов, что приводит к проблеме неоднозначности формально заданной семантики термина.

В других подходах[2,10] общий словарь реализован в виде онтологии верхнего уровня, постулирующей общие концепты, которые уточняются частными онтологиями. Это позволяет устанавливать различные семантические отношения между концептами частных онтологий, относящихся, например, к одному классу верхнего уровня. Данный подход ориентирован на применение в рамках одной предметной области, где можно явно установить набор базовых классов. Однако представленные методы, оставляют без внимания проблему установления соответствия между экземплярами онтологий – моделями, отражающими основные свойства конкретных объектов реального мира. Данное обстоятельство является очень важным для рассматриваемой предметной области, и его учет требует применения особого подхода к семантической интеграции данных.

3 Подход с использованием общего разделяемого тезауруса

3.1 Определение тезауруса

Исходя из описанных характеристик различных подходов к семантической интеграции, было решено выбрать разновидность гибридного метода, предполагающую использование расширяемого тезауруса вместо общего словаря. Основными задачами тезауруса являются: централизованное хранение элементов отдельных онтологий с сохранением их семантики, установления различного рода связей между терминами. Сохранение семантики является одной из ключевых особенностей данного подхода, и невозможно при использовании общей онтологии, определяющей некоторый общий смысл для различных объектов и отношений в отдельных информационных ресурсах, к которому приводиться смысл каждого термина. При этом часть его семантики теряется или искажается.

Тезаурус можно определить как четверку множеств – объектов, связей, атрибутов и агентов:

$$TRS =< O_U, L_U, P_U, A_U >, \qquad (4)$$

Дадим формальные определения элементов тезауруса. Понятие некоторой предметной области представляется в тезаурусе соответствующим ему элементом тезауруса типа «Объект»:

$$O =< N_O, L_O, P_O, A_O >, \qquad (5)$$

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

Связь между объектами тезауруса представим в виде:

$$L =< TP_l, O_1, O_2, W_1 >, \qquad (6)$$

где TP_l – тип связи L, $TP_l \in TP_U$, 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_1 – 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:

$$P =< N_P, O, A_P >, \qquad (7)$$

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.

3.2 Отображение онтологий в тезаурус

Данный подход предполагает задание отдельных информационных моделей - онтологий для каждого информационного ресурса, это позволяет учесть и отразить различные особенности семантики элементов данных. Далее производится процесс отображения концептов и отношений отдельных онтологий в разделяемый тезаурус. В ходе этого процесса каждому концепту, свойству и отношению онтологии, ставиться в соответствие элемент тезауруса, которому также приписывается идентификатор агента – приложения выполняющего различные задачи по обработке информации в отдельном информационном ресурсе. При этом между терминами, уже находящимися в тезаурусе, которые являются семантически близкими добавляемым, формируются взвешенные связи.

В тезаурусе межклассовые отношения представляются связями гипонимии:

$$L_0 =< hyponymOf, O_i, O_k, 100 >, \qquad (8)$$

Связи между терминами различных онтологий, такие как синонимия и ассоциация формируются на основании трех оценок:

- сходства семантики символических имен терминов;
- структурного положение понятия и термина в онтологии и тезаурусе;
- степени сходства множеств необходимых и достаточных атрибутов.

Данные оценки определяются следующими функциями:

$Syneq(O,T) = x$, (9)

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

Функция принимает объект онтологии и элемент тезауруса в качестве аргументов и возвращать степень сходства семантики символических имен. Она включает такие методы, как сравнения токенов имен терминов, определения расстояния между ними, вычисление близости определений терминов, сравнения синонимов терминов. Введем также предельные значения функции (9). Если значение функции превышает предельное значение, то два ее аргумента считаются эквивалентными:

$1 \le UPSYN \le 100$, (10)

если $Syneq(O,T) \ge UPSYN$, то $N_O = N_T$, N_O и N_T – символические имена объекта тезауруса или понятия онтологии.

Оценка сходства положений в иерархии терминов будет осуществляться функцией:

$Poseq(O,T) = x$ (11)

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

Функция (11) содержит такие методы выявления подобия, основанного на таксономическом положении терминов, как сравнение связанных путей, правило над/под термина, определение числа схожих надтерминов.

Следует отметить, что в данном случае представленные оценки являются эвристическими и поэтому использование связей на их основе возможно только для задач не требовательных к точности результата. К таким задачам можно, к примеру, отнести семантический поиск, результаты которого будут так или иначе обрабатываться экспертом, способным выявить различные неточности.

Формальную оценку сходства понятий дает функция сравнения множеств необходимых и достаточных атрибутов терминов:

$Atreq(O,T) = x$ (12)

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

Предельное значение функции (12) будет иметь вид:

$1 \le UPATR \le 100$, (13)

если $Atreq(O,T) \ge UPATR$, то объект тезауруса O и понятие онтологии T имеют близкие интерпретации.

Определение и роль необходимых и достаточных атрибутов терминов описываются далее в данной работе.

Рассмотрим процедуру включения элементов онтологии 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»:

$TRO =< "Сущность", \varnothing, \varnothing, D_U, L_U >$, (14)

$TRT =< "Сущность", L_U, A_U >$, (15)

Переходим к шагу 2.

Шаг 2. С помощью функции семантического сопоставления имен (9) и сравнения необходимых и достаточных атрибутов (12) производим сравнение каждого элемента множества гипонимов $HYPO$ (17) объекта TRO с каждым элементом множества гипонимов $HYPT$ (16) объекта TRT:

$HYPT_{TRT} = \{O_i \mid i \in \mathrm{N}\}$, (16)

где для каждого O_i существует $L_0 =< hyponymOf, O_i, O_{TRT}, W >$,

$HYPO_{TRO} = \{T_i \mid i \in \mathrm{N}\}$, (17)

где каждый T_i является прямым подклассом TRO.

Понятия онтологии, для которых обе функции возвратили значения, превышающие пороговые (10) и (13), формируют множество схожих понятий онтологии - EQ_l, остальные понятия попадают во множество несхожих - NEQ_k:

$EQ_l = \{T_i \mid i \in N\}$, (18)

где для каждого $T_i : (T_i \in HYPO_{TRO}) \& (\exists TA_j : (TA_j \in HYPT_{TRT})$
$\&(Syneq(T_i, TA_j) \ge UPSYN))$
$\&(Poseq(T_i, TA) \ge UPPOS), i \in \mathrm{N}$

$NEQ_k = \{T_i \mid i \in N\}$, (19)

где для каждого $T_i : (T_i \in HYPO_{TRO}) \& (\exists TA_j : (TA_j \in HYPT_{TRT})$
$\&(Syneq(T_i, TA_j) < UPSYN))$
$\&(Poseq(T_i, TA) < UPPOS), i \in \mathrm{N}$

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 функций (9) и (12) производим сопоставление T_n со всеми объектами тезауруса. Если одна из функций возвратила значение, превышающее пороговое, для каких-либо двух аргументов, то производиться оценка их близости, исходя из положения в иерархии с помощью функции(11). В итоге в тезаурусе между созданным элементом P_n и элементом, отобранным с помощью функций, создается ассоциативная связь с весом - W, равным среднему арифметическому трех оценок:

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

где $P_n, F \in O_U$ и для

$P_n, F : (Syneq(P_n, F) \ge UPSYN)$
$\&(Poseq(P_n, F) \ge UPPOS)$,

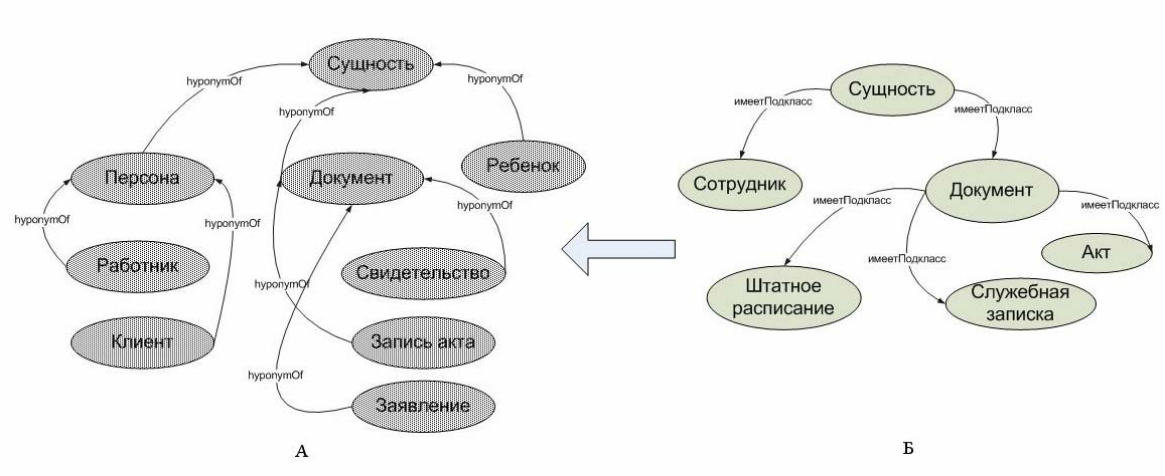

Рисунок 1. Расширение тезауруса терминами новой онтологии.

$$W = (s * Syneq(P_n, F) + p * Poseq(P_n, F) + a * Atreq(P_n, F)) / 3$$

где s, p, a - коэффициенты от 0 до 1.

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 объект тезауруса, представляющий понятие онтологии - надкласс для понятия T_n.

Далее инкрементируем счетчик n, переходим к шагу 3.

Шаг 3.3 Формируем новое множество NEQ_{k+1}, состоящее из понятий онтологии, 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 онтологии T_{ul}, $T_{ul} \in EQ_l$:

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

$$H : Atreq(T_{ul}, H) \geq UPATR$$

Переходим к шагу 6.

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

$$TRO = T_{ul}, \text{ где } T_{ul} \in EQ_l$$

$$TRT = H, \text{ где}$$

$$H : (Syneq(T_u, H) \geq SYNEQ)$$

$$\&(Poseq(T_u, H) \geq UPPOS)$$

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, то есть когда термин в тезаурусе признан эквивалентным понятию онтологии, их вершины сливаются. Далее по такому же принципу сравниваются их прямые потомки.

3.3 Пример работы алгоритма

Включение в тезаурус начальной онтологии является тривиальным, поэтому будем полагать, что в тезаурусе уже имеются термины какой-либо онтологии (рис. 1, А). Рассмотрим процедуру расширения тезауруса на упрощенном примере добавления в тезаурус новой онтологии (рис. 1, Б).

В начале работы алгоритма гипонимами термина «Сущность» в тезаурусе будут: «персона», «документ», «ребенок», а подклассами понятия «Сущность» в онтологии – «сотрудник» и «документ». В ходе сравнения семантики имен терминов и множеств необходимых и достаточных атрибутов с помощью функции (9) и (12), во множество схожих понятий онтологии (18) попадет – «документ», а несхожих (19) – «сотрудник».

Далее будут обработаны элементы множества несхожих понятий (19). В данном случае оно состоит из одного элемента – «сотрудник», который помещается в тезаурус, а далее с помощью функций (9) и (12) будет сравниваться с терминами

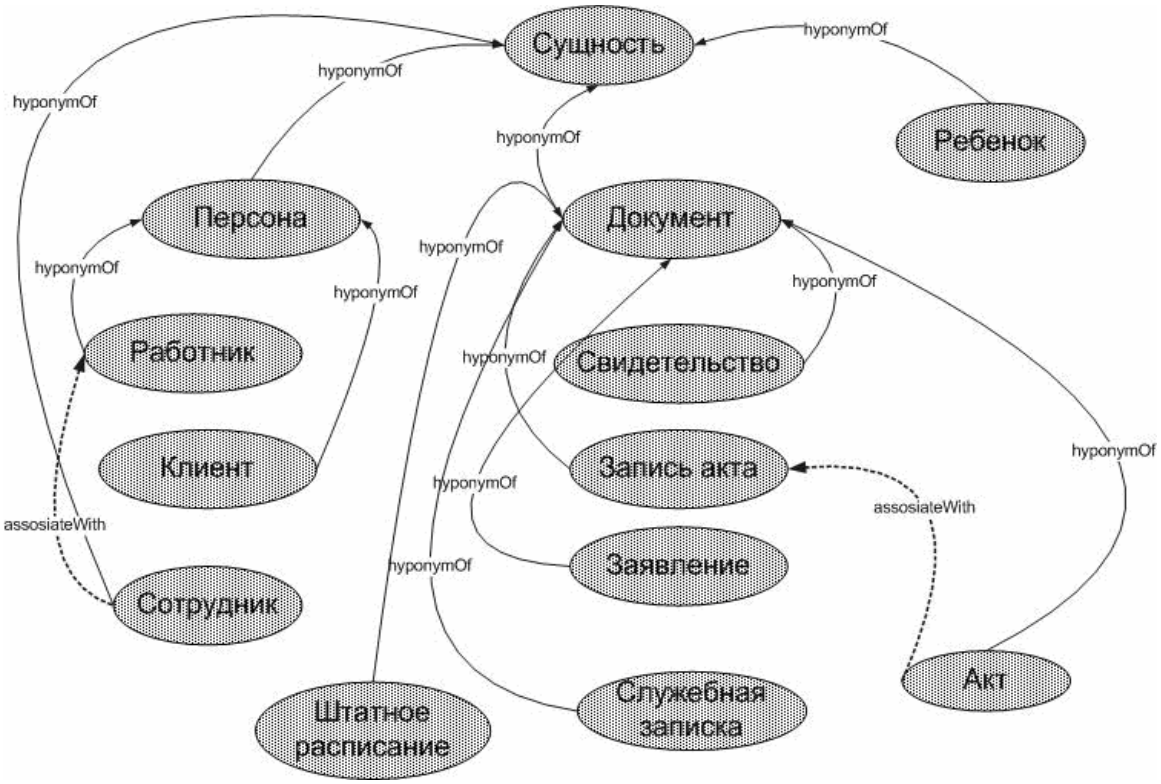

Рисунок 2. Тезаурус в результате работы алгоритма.

тезауруса. При достижении термина «работник», так как он является синонимом к «сотрудник», оба термина будут переданы в функцию (11). В зависимости от результирующей оценки между ними будет создана 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 определено 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 функций (9) и (11) терминов данных множеств будут сформировано множество непохожих терминов(19), состоящее из элементов: «штатное расписание», «служебная записка», «акт», а множество похожих(18) в данном случае будет пустым. По рассмотренному ранее принципу будет обработан каждый элемент множества 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 рисунке 2.

3.4 Определение и использование набора идентификационных атрибутов в тезаурусе

Одной из ключевых проблем семантической интеграции является сопоставление моделей представление данных. В данном случае она частично решается на этапе интеграции онтологий в тезаурус. Однако это не позволяет избавиться от ошибок и неточностей, что довольно критично для таких задач, как проверка семантической целостности и получение совокупной информации об объекте из различных источников. Это связано с тем, что интерпретации концептов и отношений явно в онтологиях не заданы, поэтому получить и сравнить их не представляется возможным. Также часто складывается такая ситуация, что один концепт может иметь интерпретацию концепта другой онтологии, не противоречащую системе аксиом первой, но по-сути иметь иной подразумеваемый смысл. Например, рассмотрим следующие наборы OWL аксиом, двух различных онтологий (для записи используется манчестерский синтаксис[8]):

Class: Person

SubClassOf: owl:Thing that hasFirstName only string[minLength 1]

Class: Cat

SubClassOf: owl:Thing that hasName only string[minLength 1]

Формально эти два концепта эквиваленты, однако, как видно из названий классов, они подразумевают разные интерпретации, которые нельзя отличить, используя лишь данные наборы аксиом.

Разумеется, степень формальной выразительности зависит от сложности онтологии с точки зрения количества заданных аксиом, в большей степени типа Abox (assertional box), используя которые, можно с помощью машины

вывода получить формальную семантику. Но разработка сложных онтологий, а не просто таксономий терминов, требует от эксперта знания не только предметной области, но и принципов и языков онтологического моделирования, что, как правило, не выполняется.

Проблему установления соответствий можно отчасти решить применением онтологии верхнего уровня, определяющей абстрактные концепты и отношения, посредством привязки к которым можно установить соответствия между элементами разных онтологий. Однако здесь возникает вопрос об уровне абстракции концептов общей онтологии, так как их интерпретации могут быть довольно большими, что не позволить разнести по ним концепты отдельных онтологий. Если же добавлять в общую онтологию дополнительные аксиомы, сужающие множества интерпретации, то это будет накладывать дополнительные ограничения на определение концептов и отношений включаемых онтологий.

Альтернативным способом согласования может служить общий словарь терминов, постулирующий общую семантику. Однако то или иное определение, заданное в нем, может быть как неоднозначным, так и не полным, что вызовет проблемы в его формализации в виде модели объекта в онтологии. Проблемой также является будущее изменение словаря таким образом, чтобы новые определения не противоречили имеющимся.

Наряду с этим, как было отмечено ранее, остается нерешенной проблема конфликтов на уровне экземпляров – моделей наиболее конкретных сущностей. Определение какого-либо соответствия между концептами разных онтологий, например, эквивалентности, означает наличие этого отношения только между множествами их интерпретаций, а не их элементами.

Иными словами это означает, что отсутствует возможность установить соответствие между экземплярами разных онтологий, интерпретации которых представляют один и тот же объект реального мира. Для рассматриваемой предметной области это является довольно серьезной проблемой, так как в данном случае информационные источники часто содержат данные, описывающие одну и ту же сущность. Однако обработку этих данных необходимо вести совместно во избежание появления различного рода противоречий.

Для разрешения данной трудности предлагается задать множество общих атрибутов-идентификаторов у экземпляров в различных онтологиях. Это позволит устанавливать отношение эквивалентности между ними. Проблему определения атрибута и присвоения ему значения, уникального в рамках множества экземпляров всех онтологий, можно решить, если, проанализировав предметную область, выявить реальное идентификационное свойство, которое, так или иначе, уже заданно в контенте информационных ресурсов.

В данном случае предлагается использовать специфику области государственного и муниципального управления заключающуюся в том, что основные субъекты и объекты имеют заранее заданные в различных юридических документах наборы основных атрибутов, в том числе и идентификационных. При этом будут также выполнены основные требования к общезначимому атрибуту-идентификатору.

Использование общего идентификатора позволяет «склеить» различные кусочки информации для того, чтобы создать некое интегрированное представление определенного объекта реального мира.

Однако в рассматриваемой предметной области один и тот же идентификационный атрибут может использоваться для определения экземпляров, относящихся к разным классам. Например, индивидуальный номер налогоплательщика, может обозначать как гражданина, так и предприятие. В сущности, определяются два различных множества значений других атрибутов: в первом случае ФИО, адрес местожительства, год рождения, а во втором – название предприятия, юридический адрес. Для разрешения данной проблемы можно дополнительно обозначить общие описательные атрибуты, к которым предъявляется только требование общезначимости. Наличие их у экземпляра, можно считать достаточным условием для членства в определенном классе.

В результате, определив наборы общезначимых идентификационных и описательных атрибутов в тезаурусе, можно использовать их для задания концептов, а также для установления соответствия между ними с помощью функции (12). Также это позволяет устранить проблему семантических конфликтов и неопределенностей на уровне экземпляров и в то же время не накладывает ограничений на отдельные онтологические модели. Наряду с этим, задание общих атрибутов позволяет формально определить некие базовые классы в тезаурусе, которые можно конкретизировать в частных онтологиях, что позволяет сократить время разработки, задав общую модель определенных сущностей.

4 Текущие результаты и направления дальнейшей работы

В данной работе представлен подход к семантической интеграции данный в сфере государственного муниципального управления с использованием разделяемого тезауруса. На данный момент определена обобщенная структура системы интеграции и назначение ее функциональных модулей[4]. Задана концептуальная модель тезауруса, а также разработан алгоритм отображения онтологий в тезаурус, включающий оценки семантической близости концептов. Для

решения проблемы установления формального соответствия между экземплярами различных онтологий представлена методика определения и использования общих атрибутов.

Среди направлений дальнейшей работы можно выделить: имплементацию тезауруса в виде онтологии на языке OWL или в виде RDF респозитория, разработка прототипов онтологических моделей учреждений государственного и муниципального управления, выявление и определение в тезаурусе идентификационных атрибутов основных сущностей предметной области и задания прав доступа к ним, разработка языка запросов между агентами-интеграторами на основе языка запросов к RFD документам – SPARQL.

Литература

[1] Богдановская И. Ю. Концепция «электронного государства», 2006. http://www.ifap.ru/pi/04/r02.doc

[2] Виттих В. А., Волхонцев Д. В., Горбенко А. В., Гриценко Е. А., Кистанов А. М., Светкина Г. Д., Скобелев П. О., Сурнин О. Л., Шамашов М. А., Мультиагентный Интернет-портал для интеграции ресурсов департаментов социального блока Самарской области, 2006. http://www.kg.ru/support/library/portal

[3] Виттих В.А., Волхонцев Д.В., Гинзбург А.Н., Караваев М.А., Скобелев П.О., Сурнин О.Л., Шамашов М.А., Распределенные онтологии и их применение в решении задач интеграции данных, http://www.kg.ru/support/library/dataintegration/

[4] Ломов П.А., Шишаев М.Г. Семантическая интеграция информационных источников для информационной поддержки управления микросистемой. – VII Всероссийская школа-семинар «Прикладные проблемы управления макросистемами». Апатиты, 31 марта - 4 апреля 2008г. / Материалы докладов. – Апатиты: изд-во КНЦ РАН , 2008. С.25-27.

[5] Gruber, T.R. (1993) A translation approach to portable ontology specifications. Knowledge Acquisition. Vol. 5

[6] OWL - Web Ontology Language. Overview, 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/

[7] OWL 2 - Web Ontology Language. Primer, 2009. http://www.w3.org/TR/owl2-primer/

[8] OWL 2 - Web Ontology Language Manchester Syntax, 2009. http://www.w3.org/TR/owl2-manchester-syntax

[9] Resource Description Framework, http://www.w3.org/RDF/

[10] Visser U., Stuckenschmidt H., Wache H., Vogele U., «Enabling Technologies for Interoperability» – Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=102D69688CD1F1200F311A2460DE6B5A?doi=10.1.1.21.5883&rep=rep1&type=pdf

[11] Wache H., Scholz T., Stieghahn H., Кцnig-Ries B., «An Integration Method for the Specification of Rule-Oriented Mediators.» In Proc. of 1999 International Symposium of Database Applications in Non-Traditional Environments (DANTE 99), Kyoto, Japan, November 1999

Development of the method of semantic integration of the information in sphere of the state and municipal administration

Lomov P. A., Shishaev M. G.

This paper offers the approach of semantic integration of information in domain of the state and municipal administration with using shared thesaurus, which allows to eliminate some critical for the considered subject domain disadvantages of existing approaches. The conceptual model of the thesaurus was presented. The procedure of mapping concepts of ontologies in the thesaurus was described. Also, the technique of the definition and comparison of ontological contexts on the base of a set of the general attributes was introduced.

♣ Работа поддержана грантом РФФИ, проект № 08-07-00301-а