УДК 004.89

ОБНОВЛЕНИЕ ОНТОЛОГИЙ С ПОМОЩЬЮ СЕМАНТИЧЕСКИХ СЕТЕЙ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Вороной С.М., Калинин А.С., Охрименко К.С.
Донецкий национальный технический университет
кафедра систем искусственного интеллекта
E-mail: fuckel-sanny@rambler.ru

Аннотация

Вороной С.М., Калинин А.С., Охрименко К.С. Обновление онтологий с помощью семантических сетей текстов на естественном языке. *В данной работе определены подходы и методы для реализации корректного механизма обновления онтологий на основе семантических сетей. Рассмотрен метод автоматического построения семантических сетей текстов.*

Общая постановка проблемы

Онтологии являются эффективным средством представления и систематизации знаний. Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную предметную область.

Онтология состоит из терминов (понятий), их определений и атрибутов, а также связанных с ними аксиом и правил вывода.

Формально онтология может быть представлена тройкой [1]:

$$O =< T, R, F > \quad (1)$$

где T – концепты (термины) предметной области, описываемые онтологией O;

R – отношения между терминами предметной области;

F – функции интерпретации, заданные на терминах и отношениях онтологии.

Чаще всего онтологии используются в качестве:

1. Словаря предметной области. Онтология содержит общую терминологическую базу предметной области, поэтому разработчики программного обеспечения могут использовать термины из онтологии для документирования своего продукта и для формирования пользовательского интерфейса, в том числе и многоязычного.
2. Отображения на базу данных. Онтология предоставляет набор базовых терминов предметной области, с которыми приходится иметь дело в любом процессе измерения.
3. Формата хранения метаданных. Свойства онтологических терминов определяют состав и формат представления метаданных, содержащихся в системе. Эффективная поддержка метаданных является одной из ключевых задач инженерии информационных систем. Привлечение онтологии позволяет повысить эффективность реализации различных средств обработки данных.
4. Формата обмена данными. Открытые форматы обмена данными с внешними системами, основанные на онтологии, существенно упрощают задачу интеграции систем, относящихся к различным областям либо созданных различными разработчиками.

Так как онтологии широко применяются для ответственных проектов, то можно с уверенностью утверждать, что онтологические данные (термины и отношения) должны соответствовать действительности, то есть быть актуальными на текущий момент времени.

Современное научное общество быстро развивается, ежедневно появляются новые открытия, теории и гипотезы, поэтому остро стоит вопрос обновления данных онтологии и

средств автоматизации этого процесса, так как корректность онтологии играет важную роль в правильном функционировании системы.

Задача обновления онтологических знаний нетривиальна и в основном может решаться в полуавтоматическом режиме, совместно с экспертом.

В задачах автоматического изменения онтологий можно выделить такие основные проблемы:

1. Семантические различия (синонимия и многозначность) между исходной онтологией и новыми извлеченными данными.
2. Структурные отличия между данными.
3. Отсутствие априорных знаний (априорные знания нужны для обновления онтологий с использованием контролируемых методов извлечения знаний).

Цель исследования

Главной целью исследования является анализ методов и алгоритмов, позволяющих обновлять онтологию по семантической сети в полуавтоматическом режиме.

Исследования

Особый акцент в работе делается на поиск оптимальных путей обновления онтологии с использованием семантических сетей. Для реализации конкретных вопросов на первом этапе необходимо выбрать наиболее эффективный и в то же время быстрый способ построения семантической сети.

Семантические сети – наиболее мощная математическая модель для представления знаний о предметной области (ПО), одно из важнейших направлений искусственного интеллекта. В настоящее время в научной литературе описано множество альтернативных представлений моделей семантических сетей. Они предназначены для решения разнообразных задач в различных ПО.

В общем случае под семантической сетью понимается выражение, приведенное в формуле 2.

$$S = (O, R_1, R_2, ...R_\kappa), \tag{2}$$

где O – множество объектов конкретной предметной области;

$R_i \quad i-1,n$ – множество отношений между объектами.

i – тип отношений.

Из множества существующих методов построения семантической сети был выбран метод создания семантической сети из коллекции текстовых документов определенной предметной области [2]. Суть метода заключается в пошаговом анализе текста, который приведен на рисунке 1.

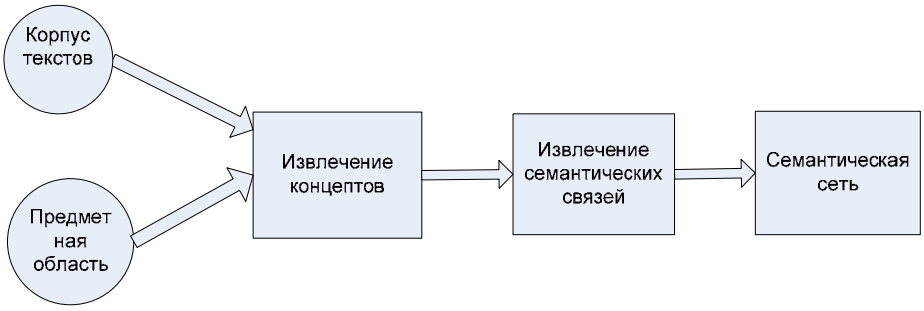

Рисунок 1 – Процесс создания семантической сети

На этапе извлечения концептов происходит выделение ключевых слов, выделение ключевых словосочетаний и группирование словосочетаний. В свою очередь группирование ключевых слов разбивается на несколько этапов, приведенных ниже.

1. Нормализация, токенизация, лемматизация.
2. Фильтрация на основе лингвистической информации: удаление стоп слов, имен собственных, чисел, дат, всего остального кроме существительных и прилагательных.
3. Ранжирование слов-кандидатов с использованием статистической информации.

Выделение ключевых словосочетаний также делится на отдельные шаги.

1. Извлечение свободных словосочетаний.
2. Группирование словосочетаний-кандидатов, путем поиска наибольших общих подстрок.
3. Ранжирование словосочетаний.

После реализации всех шагов, описанных выше, получим семантическую сеть. В результате семантическая сеть будет представлять граф, состоящий из концептов и связей между ними. Данная структура сети очень схожа с начальной структурой имеющейся онтологии (термины (классы) - отношения). Основываясь на подобии структур можно переходить к следующему шагу, поставленной выше задачи.

Следующий важный шаг – обновление онтологии из полученной семантической сети. Эта процедура состоит из трех этапов [3].

1. Выравнивание порядка графов онтологии и семантической сети. К графу с меньшим порядком добавляются фиктивные классы в количестве, равном разности порядков графов.
2. Разрешение конфликта имен. Разрешение конфликта имен проводится экспертом.
3. Объединение классов и их свойств и отношений. Происходит путем сравнения элементов графа семантической сети со всеми элементами исходной онтологии, которые находятся на одном уровне. В результате сравнения формируются два списка: список совпадающих имен концептов и список концептов, которые не совпадают, содержащихся в семантической сети и дополняют базовую онтологию.

Добавление нового понятия в онтологию.

1. Проводится обход каждого элемента второго списка, определяются их прямые родители.
2. Производится поиск в структуре онтологии на наличие класса, сходного с определенным родителем, если такой класс есть, то происходит присоединение нового концепта, иначе выполняем шаг 3.
3. Находим следующего ближайшего родителя и выполняем шаг 2.

Выводы

В данной работе были определены подходы и методы для реализации корректного механизма обновления онтологий на основе семантических сетей. Рассмотрен метод автоматического построения семантических сетей текстов. В результате проведения исследования был предложен метод, позволяющий обновлять онтологию по семантической сети в полуавтоматическом режиме.

Список литературы

1. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб. : Питер, 2001.
2. Панченко А. Построение семантической сети из разнородных данных.
3. Вороной А.С., Егошина А.А. Средства интеграции онтологий предметных областей для создания баз знаний интеллектуальных обучающих систем.