Автоматическое построение онтологий

Рабчевский Евгений evgeny@ranat.ru
Пермский Государственный Университет

Введение

Для удовлетворения своих информационных потребностей, каждый Интернет пользователь периодически посещает сайты профессиональных сообществ, подписывается и просматривает тематические рассылки и RSS подачи, ищет в поисковых системах неизвестные термины. Таким образом, у каждого профессионала выстроена своя, использующая различные Интернет технологии, система интеграции знаний в интересующей его предметной области.

Однако задачи пользователей требуют более систематизированного и настраиваемого механизма интеграции распределенных и разнородных знаний в целостную картину предметной области.

Необходимо заметить, что оригинальная спецификация WWW [1] разрабатывалась именно для решения задачи интеграции научных материалов.

Очевидно, что для эффективной интеграции данных некой предметной области из различных Интернет источников, соответствующее приложение должно работать с семантикой веб ресурсов. В этой связи, в таких приложениях актуально использование различных технологий Semantic Web [2].

Стандарты Semantic Web

В Интернет используется множество языков представления данных, основанных на XML. В рамках проекта Semantic Web, для представления

данных, имеющих графовую структуру, консорциум W3 разработал язык RDF (Resource Definition Framework – Среда Описания Ресурса). RDF предоставляет средства для записи триплетов, троек данных – субъект - предикат - объект. Объект и субъект соответствуют узлам графа, а предикат или свойство - направленной дуге графа. Дуга направлена от субъекта к объекту. Каждый из элементов триплета называют RDF ресурсом и идентифицируют с помощью URI идентификаторов.

Платформа RDF активно используется для представления различных данных, в частности RSS 3.0 агрегаторы новостей собирают информацию в формате RDF.

Для машинного представления различных предметных областей в Интернет, используются онтологии и словари. Онтология – спецификация концептуализации [3], или явное, формальное описание предметной области. Как и в объектно-ориентированном описании, онтология состоит из классов и их экземпляров. У классов и экземпляров выделяются свойства, на свойства могут накладываться логические ограничения.

Поисковой системой SWOOGLE [4] на сегодня проиндексировано свыше 10 тысяч онтологий и словарей, доступных в Веб. Онтологии используются научными сообществами – для описания терминологии [5], в электронной коммерции – для описания товаров и услуг [6], и в других приложениях Интернет. Из-за своей популярности онтологии стали использоваться и в качестве баз знаний локальных интеллектуальных систем.

Для описания онтологий, доступных через Веб, созданы языки RDFS [7] (RDF Schema – RDF Схема) и OWL [8] (Ontology Web Language - Язык Сетевых Онтологий). В качестве своих базовых элементов данные языки используют RDF ресурсы. RDFS используется для записи словарей, а OWL – онтологий. Сетевые онтологии предоставляют более выразительные

возможности по сравнению с RDF словарями, например логические операции над классами и логические ограничения свойств.

Постановка задачи

Интеллектуальные системы на основе онтологий показали на практике свою эффективность, однако построение онтологии требует экспертных знаний в исследуемой предметной области и занимает существенный объем времени, поэтому актуальной задачей является автоматизация процесса построения онтологии. Для этого предлагается использовать текстовое содержание массива Веб ресурсов описательного характера определенной тематики.

Базовой является задача разработки алгоритма автоматического построения семантической карты веб ресурса с помощью анализа его текста. Семантической картой ресурса назовем отображение контента Веб ресурса в концептуализацию его содержания, представленную в виде OWL онтологии.

Для решения данной задачи был сформирован корпус англоязычных текстов, относящихся к теме Semantic Web. Ресурсы корпуса представляют собой спецификации технологий Semantic Web с сайта W3 консорциума.

Алгоритм исследовался для определенной предметной области, что объясняется профессиональными интересами автора, а также возможностью последующей оценки полученного метода сравнением результатов с онтологией, полученной с помощью экспертных знаний (параллельно с данными исследованиями автор анализировал выбранный корпус и создавал онтологию данной предметной области без средств автоматизации).

Построение семантической карты ресурса

Семантическая карта ресурса строится на основе особенностей языка, которые позволяют вытягивать семантические конструкции из текста. Исследования проводились следующим образом:

1. формировался набор пар «текст – конструкция языка OWL»;
2. по набору выявленных пар «текст – OWL конструкция» выявлялись правила, позволяющие автоматизировать процесс отображения текста в соответствующую OWL конструкцию;

Семантическая карта строится в два этапа, на первом строится формальная семантическая OWL конструкция, на втором происходит привязка полученной конструкции к конкретной предметной области.

Сформулируем правила, использующие синтаксис языка. Правила синтаксического уровня, выявляют семантику на основе принципов построения словосочетаний и предложений. Правила формулируются, как конструкции из различных частей речи, частей предложения, предлогов и союзов, а также конкретных слов. Дополнительно вводится понятие предмета – сущности, о которой говорится в предложении, предмет может состоять из нескольких слов. Понятие предмета также используется для формулировки правил.

Рассмотрим несколько правил:

1. «Сложный предмет» или «noun1 + noun2» (два подряд идущих существительных), например словосочетание «ontology editor».

Проанализируем данный пример. Можно предположить существует целый класс абстрактных редакторов – Editor. Этот класс характеризуется тем, что все его экземпляры обладают неким характерным для этого класса свойством. В данном случае, это то, что они все что-либо редактируют. Назовем это характерное свойство mainPropertyOfEditor. Доменом этого свойства является класс Editor. Определим диапазон этого свойства, как класс RangeOfMainPropertyOfEditor. Выделим класс OntologyEditor,

который будет подклассом класса Editor. При этом значение свойства mainPropertyOfEditor для подкласса OntologyEditor имеет строго определенное значение – экземпляр класса RangeOfMainPropertyOfEditor, индивид Ontology. Данные утверждения можно представить следующим OWL кодом:

```
  <owl:Class rdf:ID="Editor">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >класс абстрактных редакторов</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="RangeOfMainPropertyOfEditor">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >диапазон характерного свойства реадактора (редактируемый
объект)</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="OntologyEditor">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="MainPropertyOfEditor"/>
        </owl:onProperty>
        <owl:hasValue>
          <RangeOfMainPropertyOfEditor rdf:ID="Ontology"/>
        </owl:hasValue>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >класс редакторов онтологий</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#Editor"/>
  </owl:Class>
  <owl:ObjectProperty rdf:about="#MainPropertyOfEditor">
    <rdfs:domain rdf:resource="#Editor"/>
    <rdfs:range rdf:resource="#RangeOfMainPropertyOfEditor"/>
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >характерное свойство редактора
(редактирует)</rdfs:comment>
  </owl:ObjectProperty>
```

2. «Предмет с определением» или «adjective + subject», например словосочетание «abstract syntax». Для записи соответствующего OWL кода необходимо провести рассуждения, аналогичные приведенным в предыдущем примере.

3. Простое предложение, subject1 + verb + preposition + subject2 (подлежащее, сказуемое, предлог, дополнение), например «Ontology’s incorporate information about classes».
4. subject1 + are + subject2 + that + verb + preposition + subject3 (подлежащее, are/is, дополнение, that, сказуемое, предлог, дополнение), например предложение «Decision Engineering is an emerging discipline that focuses on developing tools».

Отдельно выделяются правила, которые сами не строят семантическую конструкцию, но определяют, каким образом (к каким словам) применять правила, непосредственно выявляющие семантические конструкции. Например, правило «Если сложный предмет состоит из трех и более простых, то нужно применять правило «noun1 + noun2» начиная с конца».

Рассмотрим правило из примера 2, в котором по аналогии с примером 1 были бы введены свойство mainPropertyOfAbstract и класс RangeOfMainPropertyOfAbstract. Данные конструкции введены чисто формально, используя некие законы языка, однако данное свойство и класс имеют определенную семантику. Так определение Abstract характеризует некую особенность предмета Syntax. В данном случае эту особенность можно назвать, например как «степень детализации».

Если же подходить к анализу данного словосочетания с учетом семантики, указанные свойство и класс назывались бы «имеетСтепеньДетализации» и «СтепеньДетализации» соответственно.

Задача преобразования формальных семантических конструкций в конструкции, привязанные к семантике конкретной предметной области, на данный момент автором не решена. Автор считает, что для решения данной задачи требуется источник знаний со структурой подобной таблице, приведенной ниже:

Слово	Характерное свойство
Abstract	Степень детализации
Editor	Редактирует

Предполагается представить данный источник знаний в виде RDF представления WordNet подобного ресурса [9] компьютерной лингвистики.

Для решения подобной задачи предполагается получить правила, которые позволили бы выявить данную информацию, на основе статистики совместного использования слов содержащих название понятия и его семантику.

Для того чтобы привязать полученную семантическую модель к интересуемой предметной области, используется словарь соответствующей тематики. В итоговой онтологии фиксируются только те семантические конструкции, в которых участвуют термины из словаря предметной области. Словарь может создаваться экспертом или автоматически на основе статистических методов классификации.

Онтология «Semantic Web»

Для оценки метода автоматического построения онтологии авторы создают онтологию предметной области «Semantic Web» без средств автоматизации. Онтология создается на основе того же корпуса англоязычных тестов, из которого выявляются правила построения семантической карты ресурса. Тексты корпуса исследовались следующим образом:

- выявлялись понятия предметной области, и обозначающие их термины на русском и английском языках;
- каждое понятие дополнялось экспертным определением;
- выявлялись триплеты, содержащие найденные понятия;

- для каждого понятия и триплета фиксировался ресурс-источник.

Для оценки программной реализации метода автоматического построения онтологии предполагается обработать исследуемый корпус полученным программным средством и сравнить результаты с онтологией, полученной ручным способом. Создан RDF словарь для хранения результатов ручной разработки онтологии. Словарь представляет собой модель для хранения понятий с их определениями, триплетов и оригинальных источников понятий и триплетов.

Семантическая разметка, RDF/A, GRDDL

RDF графы и онтологии на их основе могут размещаться в отдельных Веб ресурсах, например файлах или RDF хранилищах, доступ к которым осуществляется через RDF сервера. Также RDF графы могут встраиваться в другие XML документы, например в XHTML. Встраивание RDF данных в XHTML используется для спецификации семантики (семантической разметки) контента.

Семантическая разметка или аннотирование представляет собой явное описание семантики контента ресурса при помощи понятий семантической модели (онтологии или словаря). Такое явное описание семантики выполняется указанием четкого соответствия между определенной частью контента ресурса и его семантикой, описанной в семантической модели.

Рабочая группа развертывания Семантического Веба W3 консорциума разработала технологию RDF/A [10], которая позволяет встраивать RDF данные в XHTML. RDF/A является одним из множества микроформатов [11] или диалектов языков, расширений языка HTML, в котором определяется, каким образом использовать конструкции языка HTML, чтобы интерпретировать записанный таким образом HTML код, как RDF данные.

Существуют микроформаты для записи таких словарей, как vCard, DC, RDF Calendar, RSS, GeoInfo. Все указанные словари записываются в виде RDF графов, RDF/A является микроформатом для записи непосредственно RDF синтаксиса, и может быть использован для записи терминов любых RDF словарей, например тех же vCard, DC, RDF Calendar, RSS, GeoInfo.

Ниже следует пример использования терминов словаря набора данных DC (словарь DC описывает мета свойства электронных документов) в XHTML.

```
<head profile="http://www.w3.org/ 2003/g/data-view">
<link rel="schema.DC" href="http://purl.org/dc"/>
<meta name="DC.Title" xml:lang="en" lang="en"
content="Использование терминов словаря DC в XHTML коде" />
</head>
```

Данный XHTML соответствует триплету субъектом которого является URI самого ресурса, предикатом – свойство Title, описанное в словаре DC по адресу http://purl.org/dc, объектом – строка "Использование терминов словаря DC в XHTML коде". Вставка такого RDF триплета в заголовок HTML страницы позволит соответствующим приложениям понять, что название документа - "Использование терминов словаря DC в XHTML коде". При этом это название может отличаться от того, которое представлено пользователю с помощью тега <title>. Таким образом, в XHTML можно вставлять любые RDF графы. Использование профиля profile=http://www.w3.org/2003/g/data-view необходимо для возможности указания значения "transformation" у тега rel, что необходимо для указания ссылки на механизм GRDDL извлечения (см. следующий абзац).

Для извлечения RDF данных из различных микроформатов W3 консорциум разработал технологию GRDDL [12] (Gleaning Resource

Descriptions from Dialects of Languages - Извлечение Описания Ресурса из Диалектов Языков). Для работы GRDDL-скреперов (программ, извлекающих RDF данные из XHTML) в XHTML коде необходимо указать ссылку на механизм извлечения:

<link rel="transformation" href=http://www.w3.org/2000/06/dc-extract/dc-extract.xsl/>

Механизм извлечения основан на технологии преобразования XML документов XSLT [18], в данном случае XHTML преобразуется в RDF.

Литература

1. Tim Berners-Lee, World Wide Web: Proposal for HyperText Project. 1990. // http://www.w3.org/Proposal.html
2. Сообщество Semantic Web // http://www.w3.org/2001/sw
3. Gruber, T.R. (1993) A translation approach to portable ontology specifications. Knowledge Acquisition. Vol. 5.
4. Swoogle - Semantic Web Search Engine. // http://swoogle.umbc.edu/
5. Е.М. Бениаминов "Алгебраические методы в теории баз данных и представлении знаний". М.: Научный мир, 2003.
6. Реестр товаров и услуг ООН. // http://www.unspsc.org/.
7. RDF Schema 1.0, Язык описания RDF словарей. Рекомендация W3C 10 Февраля 2004. // http://www.w3.org/TR/rdf-schema/
8. Язык OWL. // http://www.w3.org/2004/OWL/
9. RDF/OWL представление WordNet, Рабочий документ W3C 19 Июня 2006 http://www.w3.org/TR/wordnet-rdf/
10. Встраивание RDF в XHTML RDFa. Рабочий документ W3C 12 марта 2007. // http://www.w3.org/TR/xhtml-rdfa-primer
11. Сообщество пользователей микроформатов. // http://microformats.org/
12. Рабочая группа GRDDL http://www.w3.org/2001/sw/grddl-wg/