УДК 004.82

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА КЛАСТЕРИЗАЦИИ LSA/LSI ДЛЯ РЕШЕНИЯ ЗАДАЧИ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЙ

Егошина А.А., Орлова Е.В., Дмуховский Р.И.
Донецкий Национальный Технический Университет
Кафедра систем искусственного интеллекта
E-mail: orlova-evgenia@mail.ru

Аннотация

Егошина А.А., Орлова Е.В., Дмуховский Р.И. Использование алгоритма кластеризации LSA/LSI для решения задачи автоматического построения онтологий. *Рассмотрены подходы к решению задачи автоматического построения онтологий с последующей интеграцией данных на их основе. Показано, что предварительная кластеризация документов текстовой коллекции улучшает качество построенной онтологии. В качестве базового подхода к интеграции данных предлагается использовать метод построения онтологии по коллекции текстовых документов*

Общая постановка проблемы

На сегодняшний день практически вся информация, доступная во всемирной паутине не содержит семантики и поэтому ее поиск, релевантный запросам пользователя, а также интеграция в рамках конкретной предметной области затруднены. Для обеспечения эффективного поиска, веб-приложение должно четко понимать семантику документов, представленных в сети. В связи с этим, можно наблюдать бурный рост и развитие технологий Semantic Web, происходящий в настоящее время. Консорциумом W3C была разработана концепция, которая базируется на активном использовании метаданных, языке разметки XML, языке RDF (Resource Definition Framework – Среда Описания Ресурса) и онтологическом подходе. Все предложенные средства позволяют осуществлять обмен данными и их многократное использование.

Одним из перспективных направлений в исследованиях является использование онтологий для решения задач интеграции данных. Методы интеграции данных 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.

Анализ подходов к автоматизации процесса построения онтологий

Известны несколько подходов к определению понятия онтологии, но общепринятого определения до сих пор нет. Согласно определению Т. Грубера, онтология – это спецификация концептуализации предметной области [1]. Это формальное и декларативное представление, которое включает словарь понятий и соответствующих им терминов предметной области, а также логические выражения (аксиомы), которые описывают множество отношений между понятиями. Для описания отношений в онтологиях используются весь арсенал формальных моделей и языков, разработанных в области искусственного интеллекта – исчисление предикатов, системы продукций, семантические сети, фреймы и т.п.

Онтологии получили широкое распространение в решении проблем представления знаний и инженерии знаний, семантической интеграции информационных ресурсов, информационного поиска и т.д. 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.

На данный момент существует не менее десятка зарубежных систем, относимых к классу инструментов онтологического инжиниринга, которые поддерживают различные формализмы для описания знаний и используют различные машины вывода из этих знаний. Среди уже разработанных онтологий наиболее известными и объемными являются CYC (http://www.cyc.com) и SUMO (http://www.ontologyportal.org/).

Существуют множество подходов к автоматизации процесса построения онтологий [2 – 5]. Рассмотрим основные из них.

1. Представление онтологий в виде конечного автомата

В работе [2] предполагается, что онтологии представляются в виде орграфа G, где множество вершин V представляет множество предметных областей, а множество ребер E – бинарное отношение между этими предметными областями.

Представление онтологий в виде конечного автомата без выходов позволяет ввести операции на онтологиях. Операции на автоматах означают операции на регулярных языках, которые акцептируются этими автоматами. Основными такими операциями являются следующие: объединение, пересечение, конкатенация или умножение двух автоматов, итерация, обращение.

Алгебраические свойства введенных операций на онтологиях вытекают из соответствующих свойств операций алгебры регулярных языков. Это значит, что данные операции удовлетворяют следующим законам: коммутативность и ассоциативность операций объединения и пересечения, ассоциативность умножения, дистрибутивность операции умножения относительно операций объединения и пересечения.

Данное множество операций (в случае надобности) можно расширять по крайней мере в двух направлениях. Одним из таких направлений является расширение операциями на графах (введение и удаление вершины и ребра, соединение графов, изоморфного соединения декартового произведения и т. д.). Другим направлением является алгебра отношений. Поскольку каждая онтология является представлением некоторой совокупности отношений (в частности: одного), то можно вводить операции реляционной алгебры.

2. Построение семантической карты ресурса

В данном методе для автоматизации процесса построения онтологии предлагается использовать текстовое содержание массива Веб ресурсов описательного характера определенной тематики [3].

Базовой является задача разработки алгоритма автоматического построения семантической карты веб ресурса с помощью анализа его текста. Семантическая карта ресурса – это отображение контента Веб ресурса в концептуализацию его содержания, представленное в виде OWL онтологии.

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

- формировался набор пар «текст – конструкция языка OWL»;
- 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. Формулируются правила, использующие синтаксис языка. Правила синтаксического уровня, выявляют семантику на основе принципов построения словосочетаний и предложений. Отдельно выделяются правила, которые сами не строят семантическую конструкцию, но определяют, каким образом (к каким словам)

применять правила, непосредственно выявляющие семантические конструкции.

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.

3. Подход на основе лексико-синтаксических шаблонов

Данный подход был предложен в [4] и относится к группе методов автоматического построения онтологий, использующих лингвистические средства.

Сторонники подхода утверждают, что для построения онтологий следует активно использовать все уровни анализа естественного языка: морфологию, синтаксис и семантику. Таким образом, для автоматического построения онтологии автором используется один из методов семантического анализа текстов на естественном языке – лексико-синтаксические шаблоны.

Как метод семантического анализа лексико-синтаксические шаблоны давно используются в компьютерной лингвистике и представляют собой характерные выражения и конструкции определенных элементов языка. Данная методика семантического анализа не является специализированной на определенную предметную область.

На основе лексико-синтаксических шаблонов выделяются онтологические конструкции. В целом отмечается, что лексико-синтаксические шаблоны как метод семантического анализа текстов на естественном языке – в случае большого объема коллекции шаблонов – является эффективным средством для автоматического построения онтологий.

4. Автоматическое построение онтологии по коллекции текстовых документов

В работе [5] предлагается подход к решению проблемы автоматического построения онтологий, преимущественно основанный на статистических методах анализа текстов на естественном языке.

Построение онтологий разделено на 3 этапа:

- предварительная подготовка коллекции;
- определение классов онтологии;
- определение отношений «is-a» и «synonym-of», построение иерархии классов.

На качество построения онтологии влияет предварительная подготовка текста, в частности, особенности коллекции документов. Кластеризация документов по общей тематике может сократить время, затрачиваемое на создание онтологии. Для улучшения получаемой в результате работы системы онтологии, предлагается провести предварительную кластеризацию документов коллекции таким образом, чтобы в один кластер попадали тематически близкие документы, а дальнейшую работу проводить отдельно с каждым полученным кластером.

На первом этапе построения онтологии требуется выделить входящие в ее состав классы. Следует отметить, что понятия лингвистической онтологии строго связаны с терминами. Таким образом, данная задача сводится к определению терминов рассматриваемой предметной области.

Алгоритмы извлечения терминов из текстов на естественном языке можно разделить на две группы: статистические и лингвистические. Однако первые обладают определенным преимуществом, поскольку их использование не зависит от лингвистических особенностей конкретного языка. Подход к извлечению терминов в рассматриваемом методе является преимущественно статистическим. Предполагается, что существующие статистические методы могут показать лучшие результаты, если дополнить их определенными эвристиками.

Предварительно в качестве базовых эвристик предлагается использовать следующие:

- имя класса содержит хотя бы одно существительное;

- общеупотребительные слова обладают большей частотой встречаемости, и приблизительно равной в документах из различных кластеров;

- количество информации термина из нескольких слов больше, чем количество информации отдельных слов.

Этап выделения отношений между классами создаст наибольшие трудности. В связи с чем, первоначально имеет смысл говорить об автоматическом тезаурусе (таксономии с терминами). В качестве базовых отношений, действующих между терминами, определим отношения «is-a» и «synonym-of».

Для выделения отношения «is-a» можно воспользоваться количественным подходом к информации. Для этого было использовано предположение, что количество информации термина из нескольких слов больше, чем количество информации отдельных слов, входящих в его состав.

Предложенный подход позволяет выделить только базовые отношения, необходимые для построения таксономии. Однако предполагается, что возможно его расширение для выделения других отношений.

Выводы

В результате исследований было установлено, что широкое распространение получили подходы, основанные на статистическом анализе текста на естественном языке. В таких подходах онтология строится по коллекции текстовых документов.

На качество построения онтологии влияет предварительная подготовка текста, в частности, особенности коллекции документов. Кластеризация документов по общей тематике может сократить время, затрачиваемое на создание онтологии.

В качестве алгоритма кластеризации предлагается алгоритм LSA/LSI. Алгоритм LSA/LSI – это реализация основных принципов факторного анализа применительно ко множеству документов. Данный метод кластеризации позволяет успешно преодолевать проблемы синонимии и омонимии, присущие текстовому корпусу основываясь только на статистической информации о множестве документов/терминов.

На основе построенной онтологии можно проводить интеграцию данных. Существует множество методов, многие из которых учитывают так же алгоритм построения онтологии. В ходе исследований было выявлено, что интеграция на основе автоматически построенной онтологии проходит значительно проще, быстрее и качественнее. В частности, для интеграции текстовых данных, наиболее подходящим является метод построения онтологии по коллекции текстовых документов.

Список литературы

1. Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы. Учебное пособие. Казань, Москва. 2006;

2. Крывый С.Л., Ходзинский А.Н. Автоматное представление онтологий и операции на онтологиях / Интернет-ресурс. - Режим доступа: http://shcherbak.net/avtomatnoe-predstavlenie-ontologij-i-operacii-na-ontologiyax.

3. Рабчевский Е.А. Автоматическое построение онтологий / Интернет-ресурс. - Режим доступа: http://shcherbak.net/avtomaticheskoe-postroenie-ontologij.

4. Рабчевский Е. А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска. // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL 2009. – Петрозаводск, 2009. – С. 69–77.

5. Мозжерина Е. С. Автоматическое построение онтологии по коллекции текстовых документов // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции – RCDL 2011 – Воронеж, 2011 – С. 293 – 298.