

Предварительная кластеризация текстовых документов для повышения качества автоматического построения онтологий

*Орлова Е.В., Дмуховский Р.И., Егошина А.А.
Донецкий Национальный Технический Университет
Кафедра систем искусственного интеллекта*

Рассматривается онтологический инжиниринг и методы автоматического построения онтологий. Предлагается применять кластеризацию документов по общей тематике для улучшения качества онтологии с помощью алгоритма LSA/LSI. В качестве понятий, по которым будет происходить составление онтологии, предлагается использовать существительные, встречающиеся в текстах, а в качестве отношения между ними - степень их семантической связи, оцениваемой на основе закона Д. Зипфа.

Общая постановка проблемы

В настоящее время, можно наблюдать бурный рост и развитие технологий Semantic Web. Одним из перспективных направлений в исследованиях является использование онтологий. Онтологии являются новым средством представления и обработки знаний, позволяют создавать интеллектуальные средства для поиска ресурсов в сети Интернет. Они способны точно и эффективно описывать семантику данных для некоторой предметной области и решать проблему несовместимости и противоречивости понятий. Так же, онтологии обладают собственными средствами обработки (логического вывода), соответствующими задачам семантической обработки информации.

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

Известны несколько подходов к определению понятия онтологии, но общепринятого определения до сих пор нет, поскольку в зависимости от каждой конкретной задачи удобно интерпретировать этот термин по-разному: от неформальных определений до описаний онтологий в понятиях и конструкциях логики и математики [1].

Том Грубер определил онтологию как «точную спецификацию концептуализации», где средством концептуализации выступает описание множества объектов и связей между ними. Формально онтологию предметной области обычно записывают в виде формулы (1).

$$O = \{X, R, F\}, \quad (1)$$

где O – онтология предметной области, X – множество понятий предметной области, R – множество отношений между этими понятиями, F – множество функций интерпретации этих понятий и отношений между ними.

Онтологический инжиниринг

Онтологический инжиниринг – одно из популярных направлений компьютерных наук, в рамках которого разрабатываются и проектируются компьютерные онтологии, соединившие в себе различные области знания: искусственный интеллект, логику, философию.

На рынке программных средств достаточно активно продвигаются более 50 редакторов онтологий, одной из наиболее популярных систем работы с онтологиями, созданная в Стэнфордском университете является система Protege.

Protege – это одна из наиболее популярных систем работы с онтологиями, созданная в Стэнфордском университете (США). По версии разработчиков системы, все понятия предметной области делятся на классы, подклассы, экземпляры. Экземпляры могут быть как у класса, так и подкласса и описываются они фреймом. Разработка онтологий для Protege состоит из 5 шагов:

- выделение области онтологии, иначе определение границ онтологии;
- определение классов;
- организация иерархии классов;
- формирование фреймов для описания классов, подклассов, экземпляров, через определение слотов, т.е. свойств;
- определение значений.

Разработчики Protege считают, что нет правильного способа создания онтологии, так как онтология – это взгляд аналитика, т.е. всегда субъективна.

Обзор существующих методов автоматического построения онтологий

Существуют множество подходов к автоматизации процесса построения онтологий [2-3]. Рассмотрим некоторые из них.

1. Построение семантической карты ресурса.

В данном методе для автоматизации процесса построения онтологии предлагается использовать текстовое содержание массива Веб ресурсов описательного характера определенной тематики [2].

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.

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.

Для того чтобы привязать полученную семантическую модель к интересующей предметной области, используется словарь соответствующей тематики. В итоговой онтологии фиксируются только те семантические конструкции, в которых участвуют термины из словаря предметной области.

2. Автоматическое построение онтологии по коллекции текстовых документов.

В работе [3] 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 первом этапе построения онтологии требуется выделить входящие в ее состав классы. Данная задача сводится к определению терминов рассматриваемой предметной области.

Этап выделения отношений между классами создает наибольшие трудности. В

связи с чем, первоначально имеет смысл говорить об автоматическом тезаурусе (таксономии с терминами). В качестве базовых отношений, действующих между терминами, берутся отношения «is-a» и «synonym-of».

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

Такой подход позволяет выделить только базовые отношения, однако предполагается, что возможно его расширение для выделения других отношений.

Предлагаемая технология автоматического построения онтологии

Предварительный этап в построении онтологии – это подготовка коллекции документов. Особенностью работы с текстами на естественном языке является необходимость предварительной обработки данных. Процесс обработки обычно состоит из нескольких этапов:

- приведение документов к единому формату;
- токенизация;
- стемминг (лемматизация);
- исключение стоп-слов.

Для улучшения качества онтологии применяется кластеризация документов по общей тематике. Кластеризация существенно сократит время, затрачиваемое на создание онтологии.

В качестве алгоритма кластеризации предлагается алгоритм LSA/LSI. Данный метод кластеризации позволяет успешно преодолевать проблемы синонимии и омонимии, присущие текстовому корпусу основываясь только на статистической информации о множестве документов/терминов.

Латентно-семантический анализ (LSA) [4] – это метод обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами в них встречающимися, сопоставляющий некоторые факторы (тематики) всем документам и терминам.

В основе метода латентно-семантического анализа лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов. При классификации/кластеризации документов этот метод используется для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки больших корпусов текстов

Существуют два основных отличия метода LSA от прочих статистических методов обработки текстов:

- в качестве исходных данных LSA использует частоту использования слов в отрывках текста, а не частоту совместного использования слов;
- метод собирает данные не о попарной совместной используемости слов, а об используемости множества слов в большом массиве отрывков.

После кластеризации коллекции документов, строим онтологию по обработанным текстам. В качестве понятий, по которым будет происходить составление онтологии будут использоваться существительные, встречающиеся в текстах.

Отношение между двумя понятиями будем представлять степенью их семантической связи, оцениваемой на основе закона Джорджа Зипфа [5] по формуле (2).

$$\frac{P * R}{N} = C, \quad (2)$$

где P – частота вхождения слова в текст, R – ранг этой частоты, N – общее количество слов в тексте, а C – встречаемость слова в языке. Ранг частоты по Зипфу определяется по частоте вхождения слова в текст. Наиболее часто встречающиеся слова имеют ранг 1, реже встречающиеся слова – ранг 2, ранг M – наименее часто встречающиеся слова, так что M – общее число рангов конкретного текста.

Джордж Зипф статистически определил, что встречаемость слова приблизительно одинакова для всех без исключения текстов в пределах одной языковой группы и подчиняется приведенному выше закону.

Из закона Зипфа для одного слова следует то, что встречаемость пары слов также будет приблизительно постоянна для любых текстов. Если рассчитать величину встречаемости для слов A и B в некотором тексте по формулам (3) и (4),

$$\frac{P_A * R_A}{N} = C_A, \quad (3)$$

$$\frac{P_B * R_B}{N} = C_B, \quad (4)$$

то степень их семантической связи получим по формуле (5).

$$C_{AB} = \frac{C_A + C_B}{2} * \rho, \quad (5)$$

где C_{AB} – степень семантической связи между словами A и B , ρ – количество слов в кортеже (A, \dots, B).

Степень семантической связи учитывает влияние всех слов между исследуемой парой. Кроме того, стоящие между двумя существительными слова имеют назначение связать их синтаксически и отразить семантическую связь. Это расстояние учитывается для того, чтобы однозначно определить, встречаются ли исследуемые пары на приблизительно одинаковых расстояниях друг от друга во всех текстах предметных областей.

На рисунке 1 показано, что пары с похожей степенью семантической связи хранятся группами. Такой способ хранения вместе с исследованиями статистики расстояний между словами позволит выяснить возможность автоматической классификации, исходя из степени семантической связи слов. Например, пара «животное — медведь» теоретически должна иметь степень семантической связи, близкую к паре «животное — слон» или «насекомое — муравей».

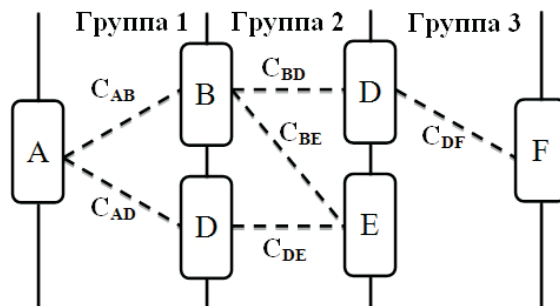


Рисунок 1 – Хранение отношений между понятиями

Хранение отношений между понятиями будет организовано следующим образом. Когда для пары слов А и В рассчитывается степень семантической связи C_{AB} по формуле (5), для их хранения создается группа, если только это значение не близко к одному из уже существующих. Впоследствии, если другая пара, например, А и D, после расчета получит значение, близкое к C_{AD} , то она попадет в эту же группу.

Для того чтобы построить онтологию нужно выделить отношения, которые в рамках одной группы связывают как минимум 2 разных понятия. На основе выделенных понятий строится онтология. Затем выделим те понятия, которые связаны как минимум с 2-мя уже присутствующими понятиями и добавим эти понятия к основе, построенной на первом шаге.

В результате работы данного алгоритма будет построена адекватная онтология, нуждающаяся лишь в незначительной корректировке экспертом.

Выводы

На основе анализа существующих методов автоматического построения онтологий показано, что существующие методы далеки до универсальности и требуют усовершенствования.

Установлено, что широко распространены подходы к созданию онтологий, 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.

На основе кластеризации документов предложена технология построения онтологии по коллекции текстовых документов. В качестве алгоритма кластеризации предлагается алгоритм LSA/LSI. Алгоритм LSA/LSI – это реализация основных принципов факторного анализа применительно к множеству документов. Отношения между понятиями предложено устанавливать по степени их семантической связи, оцениваемой на основе закона Зипфа.

Предложенная технология позволяет строить онтологии, нуждающиеся лишь в незначительной корректировке экспертом.

Литература

- [1] Клещев А.С., Артемьева И.Л. Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология».
- [2] Рабчевский Е.А. Автоматическое построение онтологий / Интернет-ресурс. – Режим доступа: <http://shcherbak.net/avtomaticheskoe-postroenie-ontologij>.
- [3] 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.
- [4] Латентно-семантический анализ. Материал из Википедии – свободной энциклопедии. / Интернет-ресурс. – Режим доступа: http://ru.wikipedia.org/wiki/Латентно-семантический_анализ.
- [5] Закон Зипфа. Материал из Википедии – свободной энциклопедии. / Интернет-ресурс. – Режим доступа: http://ru.wikipedia.org/wiki/Закон_Ципфа.