

Предварительная кластеризация текстовых документов для повышения качества автоматического построения онтологий

Автор: Вороной С.М., Егошина А.А.

Источник: Вороной С.М. Предварительная кластеризация текстовых документов для повышения качества автоматического построения онтологий /С.М. Вороной, А.А.

Егошина // Бионика интеллекта: науч.-техн. журнал. 2013. №1 (66).

<http://ea.donntu.org/handle/123456789/18875>

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.

**ОНТОЛОГИЧЕСКИЙ ИНЖИНИРИНГ, КЛАСТЕРИЗАЦИЯ ДОКУМЕНТОВ,
АЛГОРИТМ LSA/LSI, СЕМАНТИЧЕСКАЯ СВЯЗЬ**

Введение

Сегодня основная часть информации, доступной 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. Данный фактор обуславливает бурный рост и развитие технологий Semantic Web. Одним из перспективных направлений в данной области является использование онтологий, которые, являясь 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.

1. Онтологический инжиниринг и существующие методы автоматического построения онтологий

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 и философию.

Для представления онтологий 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 онтологий требует экспертных знаний в исследуемой предметной области и занимает существенный объем времени, поэтому актуальной задачей является автоматизация процесса построения онтологий.

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/>).

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

Сегодня существует множество подходов к автоматизации процесса построения онтологий [1 – 4]. Рассмотрим основные из них.

1. Представление онтологий в виде конечного автомата

В работе [1] онтологии предлагается представлять 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 [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. 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 интересующей 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 [3] 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. 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 [4] предлагается подход к решению проблемы автоматического построения онтологий, преимущественно основанный на статистических методах анализа текстов на естественном языке.

Построение онтологий разделено на три 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.

2. Использование алгоритма кластеризации LSA/LSI для решения задачи автоматического построения онтологий

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 (тематике) 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 (1).



(1)

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, равный единице, реже встречающиеся слова — ранг, равный двум, 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 (2) и (3),

$$\left[\begin{array}{l} \text{Степень семантической связи} \\ \text{между словами } A \text{ и } B \\ \text{равна } C_{AB} \\ \text{где } C_{AB} = \frac{r_{AB}}{r} \end{array} \right], (2)$$

$$\left[\begin{array}{l} \text{Степень семантической связи} \\ \text{между словами } A \text{ и } B \\ \text{равна } C_{AB} \\ \text{где } C_{AB} = \frac{r_{AB}}{r} \end{array} \right], (3)$$

то степень их семантической связи получим по формуле (4).

$$\left[\begin{array}{l} \text{Степень семантической связи} \\ \text{между словами } A \text{ и } B \\ \text{равна } C_{AB} \\ \text{где } C_{AB} = \frac{r_{AB}}{r} \end{array} \right], (4)$$

где C_{AB} — степень семантической связи между словами A и B , r — количество слов в corteже (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 рис. 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».



Рис. 1. Хранение отношений между понятиями

Хранение отношений между понятиями будет организовано следующим образом. Когда для пары слов A и B рассчитывается степень семантической связи C_{AB} по формуле (4), для их хранения создается группа, если только это значение не близко к одному из уже существующих. Впоследствии, если другая пара, например, A и D , после расчета получит значение, близкое к C_{AB} , то она попадет в эту же группу.

Для того чтобы построить онтологию нужно выделить отношения, которые в рамках одной группы связывают как минимум два разных понятия. На основе выделенных понятий строится онтология. Затем выделяются те понятия, которые связаны как минимум с двумя уже присутствующими понятиями, после чего эти понятия добавляются к основе, построенной на первом шаге.

Выводы

В данной работе на основе анализа существующих методов автоматизации процесса построения онтологий установлено, что наиболее популярными являются подходы к созданию онтологий, основанные на статистическом анализе естественно-языкового текста. На качество построения онтологии влияет предварительная подготовка текста, в частности, особенности формирования коллекции текстовых документов.

Кластеризация документов по общей тематике позволит сократить время, затрачиваемое на создание онтологии. В качестве алгоритма кластеризации предлагается использовать алгоритм LSA/LSI, который является реализацией основных принципов факторного анализа применительно к множеству документов. А отношения между понятиями предметной области предложено устанавливать по степени их семантической связи, оцениваемой на основе закона Зипфа. Предложенная технология позволяет строить онтологии, нуждающиеся лишь в незначительной корректировке эксперта.

Список использованной литературы

1. Крытый С.Л., Ходзинский А.Н. Автоматное представление онтологий и операции на онтологиях [Электронный ресурс]. – Режим доступа: <http://shcherbak.net/avtomatnoe-predstavlenie-ontologij-i-operacii-na-ontologiyax>
2. Рабчевский Е.А. Автоматическое построение онтологий [Электронный ресурс]. – Режим доступа: <http://shcherbak.net/avtomaticheskoe-postroenie-ontologij>.
3. 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.
4. 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.
5. Закон Зипфа. Материал из Википедии – свободной энциклопедии [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/wiki/Закон_Ципфа