Автоматическое построение онтологии по коллекции текстовых документов

© Е. С. Мозжерина

Санкт-Петербургский Государственный Университет
mozzherina@gmail.com

Аннотация

Рассматриваются теоретические аспекты процесса автоматического построения онтологии по коллекции тестовых документов, тематически относящихся к одной предметной области. Предлагаемый подход преимущественно основывается на статистических методах анализа текстов на естественном языке.

1. Введение

В связи с постоянным ростом объемов информации, представленной в электронном виде, задача эффективного поиска в больших коллекциях текстовых документов продолжает оставаться актуальной. Одним из перспективных направлений повышения релевантности результатов поиска является семантический поиск.

Семантический поиск — вид полнотекстового автоматизированного информационного поиска, который ведется с учетом смыслового содержания слов и словосочетаний запроса пользователя и текстов информационных ресурсов [15]. Основная проблема реализации семантического поиска заключается в необходимости смыслового анализа текстов, т.е. извлечения смысла из текста и его отображения в некоторую формальную модель, позволяющую определять близость поискового запроса и документа.

Одним из подходов к решению задачи семантизации поиска, на практике показавшим свою эффективность, является подход на основе онтологий. Например, предложенная в [20] модель организации поиска информации в сети Internet использует вывод на онтологии для формирования более эффективных запросов и, следовательно, повышения релевантности результатов.

Как правило, построение онтологии требует использования больших ресурсов, а также экспертных знаний в предметной области, и занимает существенный объем времени [18]. Таким образом, автоматизация процесса построения онтологии является актуальной задачей.

Представляется возможным автоматическое построение онтологии по коллекции текстовых документов преимущественно на основе статистических методов анализа текстов на естественном языке.

Стоит отметить, что содержание документов в коллекции непосредственно влияет на качество получаемой онтологии. Если тематически тексты документов слабо связаны, скорее всего, построенная онтология окажется невыразительной, поскольку будет описывать отдельные аспекты различных предметных областей, не создавая при этом общей картины.

Подобная ситуация может возникнуть и в случае построения онтологии на основе достаточно длинного документа, отдельные разделы которого относятся к разным предметным областям.

С другой стороны, небольшие документы обычно содержательно затрагивают только некоторую узкую часть предметной области, в связи с чем, вероятно, что онтология может оказаться неполной.

Таким образом, особое внимание должно уделяться содержанию текста для построения онтологии.

Основная цель настоящей статьи — показать теоретическую возможность автоматизации процесса построения онтологии по коллекции текстовых документов, относящихся к одной тематике, с использованием статистических методов анализа естественно-языковых текстов.

Поскольку в литературе встречается множество различных подходов к понятию «онтология», в разделе 2 уточняется содержание понятия в рамках данной статьи. В разделе 3 рассматриваются представленные в литературе подходы к процессу автоматизации построения онтологии. Раздел 4 описывает предлагаемый подход к автоматическому построению онтологии по тематически связной коллекции текстовых документов. В заключении обсуждаются отдельные вопросы практической реализации представленного подхода, а также планы по его дальнейшему расширению на оставшиеся незатронутыми этапы построения онтологии.

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

2. Понятие онтологии

В литературе по искусственному интеллекту представлено множество определений понятия «онтология». Известно, что одним из первых в области информационных технологий данное понятие начал использовать T. Gruber, который определил онтологию как «точную спецификацию концептуализации» [6].

Разнообразие практических задач, решаемых с помощью онтологий, приводит к различным содержательным трактовкам данного понятия, исчерпывающий анализ которых представлен в [16]. В данной статье под онтологией будем понимать некоторое формальное представление предметной области.

Воспользуемся определением, представленным в [13], и определим онтологию как упорядоченную тройку вида:

$$O =< T, R, F >,$$

где T — конечное множество терминов (концептов, понятий, классов) предметной области, которую представляет онтология O; R — конечное множество отношений между понятиями заданной предметной области; F — конечное множество функций интерпретации (аксиоматизация), заданных на концептах и/или отношениях онтологии O.

Заметим, что еще одним естественным ограничением множества T (помимо конечности) является его непустота, при этом на множества R и F такие ограничения не накладываются. Если множества R и F являются пустыми, то работа ведется с глоссарием. Если R состоит из единственного отношения «is-a», а F — пусто, то онтология будет представлять собой таксономию.

Необходимо особо подчеркнуть, что процесс построения онтологии представляет собой плохо формализуемую и ресурсоемкую задачу. Однако для поддержки решения задач информационного поиска применяются онтологии особого типа — лингвистические онтологии [14]. В онтологиях такого типа понятия строго связаны с терминами предметной области. Примером лингвистической онтологии является общезначимая онтология английского языка WordNet.

В дальнейшем будем рассматривать вопросы автоматического построения лингвистических онтологий.

В соответствии с определением, компоненты онтологии включают в себя классы, слоты (которые представляют собой атрибуты классов, их также иногда называют свойствами), факты (отдельные элементы, экземпляры классов) и отношения между классами и фактами.

В качестве примера рассмотрим отрывок текста, относящегося к предметной области «музыкальные инструменты»:

> *Музыкальный инструмент флейта относится к ряду деревянных духовых инструментов. Флейты представляют собой целое семейство инструментов (блок флейта, флейта пикколо). Современные флейты изготавливают из дерева, металла, пластика. Профессиональные музыканты часто предпочитают инструменты производителя Yamaha.*

На основе данного отрывка можно выделить классы «музыкальный инструмент» (атрибут класса «тип инструмента»), «флейта» (с атрибутами «материал» и «производитель» и отношением «is-a» с предыдущим классом), «блок флейта» и «флейта пикколо». Последние два класса связанны отношением «is-a» с классом флейт. Если бы в отрывке упоминалась конкретная модель флейты (например, флейта Yamaha YRA-83), то она представляла бы собой экземпляр класса (в примере — экземпляр класса блок флейт).

Таким образом, на практике разработка онтологии включает в себя четыре этапа [8]:

1. определение классов онтологии;
2. определение отношений и построение иерархии классов;
3. определение слотов и описание допустимых значений для слотов;
4. заполнение значений слотов для экземпляров.

В примере в качестве допустимых значений слота «тип инструмента» могут выступать «деревянные духовые», «медые духовые» и т.д.

Как уже было сказано, онтологии применяются для решения множества разнообразных практических задач. Наиболее масштабным проектом, ядро которого составляют онтологии, стал проект консорциума W3C Semantic Web. Основная цель проекта заключается в добавлении семантики к информации, содержащейся в Internet, с целью сделать ее доступной для автоматического восприятия компьютером.

Стоит отметить, что Semantic Web, из-за своей масштабности и принятым стандартам, оказал сильное влияние на все разработки, в которых используются онтологии. В рамках проекта, в частности, были разработаны различные языки описания онтологий (RDF [7] и RDFS [1], а также недавно принятый стандарт OWL 2 [9]).

В соответствии со стандартами были разработаны различные редакторы онтологий (такие как Protégé [5]) и процессоры для их обработки (например, Jena — средство вывода на онтологии [2]). Подробный анализ существующих инструментальных средств работы с онтологиями представлен в [3].

3. Автоматическое построение онтологий

Методы автоматического построения онтологий можно условно разделить на три основные группы в зависимости от области заимствования основного подхода: методы, основанные на подходах из области искусственного интеллекта, статистические методы и методы, использующие лингвистические подходы.

Поскольку подход, предлагаемый в статье, преимущественно основан на теории информации и относится к статистическим, в данном разделе рассматриваются методы, относящиеся к двум другим группам.

3.1 Подход на основе лексико-синтаксических шаблонов

Данный подход был предложен в [19] 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. Например, из предложения «Студент — это человек, который учится в университете», предлагаемая в [19] система выделит классы «студент», «человек» и отношение «subclass-of» между ними.

Ввиду сложности задачи, оценка результатов применения подхода проводится авторами опосредованно через анализ результатов его использования в различных приложениях Semantic Web.

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.

3.2 Подход на основе системы продукций

Данный подход был предложен в [17] и относится к группе методов автоматического построения онтологий, в основе которых лежат подходы из области искусственного интеллекта.

Автор утверждает, что эффективное автоматическое построение онтологий может быть основано на способности методов искусственного интеллекта к извлечению из текста элементов знаний и их нетривиальной переработке.

Анализ области естественно-языковой обработки текста показывает преобладание использования различных правил при решении задач в рассматриваемой предметной области. Данный факт, а также декларативный характер представления методов автоматического построения онтологий, обосновывает применение системы продукций в качестве модели представления знаний о методе.

Для создания методов автоматического построения онтологий автор разрабатывает модель генерации системы продукций (на основе применения генетического программирования), модель генерации преобразователей (на основе генетического и автоматного программирования), модель генерации систем логического вывода (также на основе генетического и автоматного программирования) и модель аппарата активации продукций (на основе применения автоматного программирования).

Таким образом, автором метода предлагается модель автоматического построения онтологий в виде системы продукций и применении генетического и автоматного программирования для создания требуемых моделей.

4. Подход на основе статистических методов

Рассмотренные выше подходы не являются единственно возможными. В данной статье предлагается подход к решению проблемы автоматического построения онтологий, преимущественно основанный на статистических методах анализа текстов на естественном языке.

В рамках данной статьи рассмотрим предлагаемый подход для первых двух этапов построения онтологии: выделение классов и отношений межлу ними.

4.1 Предварительная подготовка коллекции

Одной из особенностей работы с текстами на естественном языке является необходимость обязательной предварительной обработки данных. Процесс обработки может быть достаточно трудоемким и обычно состоит из нескольких этапов:

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

Однако не всегда есть необходимость в проведении всех вышеперечисленных этапов. Более подробно данные вопросы рассматриваются в [12].

В результате предварительной обработки каждый документ коллекции характеризуется вектором типов данного документа и их частотой встречаемости.

Ранее отмечалось, что особенности коллекции влияют на качество 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.

Стоит заметить, что какие-либо специальные требования к алгоритму кластеризации отсутствуют. В качестве алгоритма кластеризации предполагается использование метода Contextual Document Clustering (подробнее см. [4]), дающего хорошие результаты на больших текстовых коллекциях.

4.2 Определение классов онтологии

На первом этапе построения онтологии требуется выделить входящие в ее состав классы. Ранее было отмечено, 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 [11]. Однако первые обладают определенным преимуществом, поскольку их использование не зависит от лингвистических особенностей конкретного языка.

Подход к извлечению терминов в данной статье является преимущественно статистическим. Тем не менее, предполагается, что существующие статистические методы могут показать лучшие результаты, если дополнить их определенными эвристиками.

Предварительно в качестве базовых эвристик предлагается использовать следующие:

I. Имя класса содержит хотя бы одно существительное.
II. Общеупотребительные слова по сравнению с терминами обладают большей частотой встречаемости, приблизительно равной в различных предметных областях.
III. Количество информации термина из нескольких слов больше, чем количество информации отдельных слов, входящих в его состав.

Опишем предлагаемый подход более подробно. На первом этапе в каждой коллекции документов выделяют существительные и определяют их частоту встречаемости. Следовательно, в результате использования I, число предполагаемых классов значительно сокращается.

На втором этапе выделяют термины, состоящие из одного слова. На основании выдвинутой эвристики II, сравниваются частоты встречаемости различных существительных в рамках одной коллекции, также проводится оценка пересечения различных коллекций по используемым существительным.

Однако статистические данные — не единственный источник классов онтологии. Терминологические словари также могут стать источниками знаний при формировании ядра онтологии. В случае работы с коллекциями неспециализированных в конкретной области документов возможно использование существующих разработанных экспертами онтологий (например, для английского языка — онтологии WordNet).

Наконец, на третьем этапе на основе взаимной информации могут быть выделены термины, состоящие из нескольких слов. Стоит отметить, что в данном случае используется эвристика III.

Для случая двухсложных терминов получаем, что взаимная информация определяется по формуле:

$$mi(x, y) = \frac{P(x.y)}{P(x)P(y)},$$

где x и y представляют собой отдельные слова термина, $P(x)$ – частота встречаемости x, $P(x,y)$ – частота совместной встречаемости x и y.

В [10] подробно рассматриваются особенности использования данной формулы, а также представлен алгоритм, позволяющий статистически определить термины, состоящие из нескольких слов.

Выделенные описанным выше образом термины будут представлять собой классы будущей онтологии.

Таким образом, предлагаемый подход может быть отнесен к группе статистических методов. Предполагается, что выдвинутые эвристики (в том числе, что в состав имени класса должно входить существительное), являются достаточно универсальными и не ограничивают применение метода только русским языком.

4.3 Определение отношений между классами

Представляется, что этап выделения отношений между классами создаст наибольшие трудности. В связи с чем, первоначально имеет смысл говорить об автоматическом построении не произвольной прикладной онтологии, а тезауруса (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 еще раз воспользуемся сделанным в предыдущем пункте предположением III.

Очевидно, что термин, находящийся на более низком уровне иерархии, обладает большим количеством информации, чем обобщающий термин.

Так, в примере из раздела 2 количество информации термина «флейта пикколо» будет больше, чем количество информации термина «флейта». Следовательно, последний термин может являться надклассом первого.

Однако для установления какого-либо отношения между терминами, знания только о

количестве информации, которое они в себе содержат, недостаточно.

Предположим, что для каждого полученного на первом этапе термина вычислено соответствующее ему количество информации. Определение того, связаны ли два различных термина с разным количеством информации отношением «is-a», можно проводить двумя способами.

Первый способ основывается на предположении, что частные термины содержат в своем составе слова из более общих терминов. Например, «блок флейта» и «флейта пикколо», содержат в себе термин «флейта». С учетом проведенного выше анализа по количеству информации этих терминов, вполне обоснованным выглядит предположение об установлении отношения «is-a» между ними (а именно, «‘флейта пикколо’ is-a ‘флейта’»).

Второй способ основывается на понятии «контекста слова». Согласно [4] контекст слова может быть определен как условная вероятность $P(Y|x)$, где Y — переменная величина, принимающая значения из словаря коллекции, а x — искомое слово.

Понятие контекста может быть расширено до «контекста термина». Тогда x будет представлять собой искомый термин, который в общем случае может состоять из нескольких слов, а переменная величина Y будет принимать значения из словаря терминов, встречающихся в рассматриваемой коллекции документов.

Иначе говоря, под контекстом термина будем понимать некоторое множество слов, которые встречаются одновременно с данным.

В случае если у терминов нет общих слов, но совпадает контекст, и при этом они обладают разным количеством информации, имеет смысл говорить об отношении «is-a» между ними.

Возвращаясь к примеру, можно предположить, что в расссатриваемом отрывке описанные условия будут выполняться для терминов «музыкальный инструмент» и «флейта». Контекст их употребления будет совпадать, в то время как количество информации последнего термина окажется выше.

Если контекст слов совпадает, но количество информации терминов приблизительно равное, то вероятнее всего между терминами действует отношение синонимии, т.е. отношение «synonym-of».

В примере из раздела 2 контекст терминов определить достаточно сложно в виду малого объема текста. Между тем можно предположить, что в рассматриваемой предметной области термины «исполнитель» и «музыкант» окажутся синонимами, контекст которых составят термины «инструмент», «произведение» и некоторые другие.

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

4.4 Оценка качества построенной онтологии

Распространенный критерий оценки качества онтологии основан на оценке работы приложения, использующего онтологию. В связи с чем, оценка автоматически построенных онтологий является отдельной сложной задачей.

Ввиду существования отработанных методик оценки качества информационного поиска (точность и полнота поиска), можно оценивать онтологии по качеству работы систем семантического поиска, использующих онтологии. На основе онтологий системы могут значительно сужать пространство поиска за счет динамического расширения запросов пользователя.

В свободном доступе находится множество различных текстовых коллекций. В основном, такие коллекции используются в исследовательских целях для оценки качества предложенных подходов к решению задач поиска, кластеризации и классификации текстов.

Имеет смысл проводить оценку качества построения онтологий на двух коллекциях: специализированной (например, коллекции MEDLINE) и коллекции с неспециализированными документами (такой как Reuters–21578).

Дополнительной сложностью для оценки качества построенной онтологии является отсутствие некоторого эталона. Экспертная оценка достаточно трудозатратна. В связи с этим, особое значение приобретает коллекция MEDLINE, для которой существует построенный тезаурус терминов, что позволит оценить работу метода на первых этапах.

5. Заключение

В статье обосновывается подход к автоматизации процесса построения онтологии по коллекции текстовых документов, относящихся к одной тематике, на основании статистических методов анализа естественно-языковых текстов.

Предполагается, что термины и некоторые базовые отношения между ними могут быть выделены автоматически из коллекции текстовых документов на основании статистических данных.

В дальнейшем планируется практическая реализация представленного теоретического подхода к автоматическому построению тезауруса по коллекциям текстовых документов, его оценка и сравнение с существующими подходами из других групп.

Также планируется расширение представленного подхода на оставшиеся незатронутыми заключительные этапы построения онтологии: определение слотов и фактов.

На этапе реализации необходимо особо учесть, что генерируемая онтология должна отвечать существующим стандартам и должна допускать экспорт в популярные инструментальные системы для работы с онтологиями.

Литература

[1] Brickley, D., and Guha, R. V. RDF Vocabulary description language 1.0: RDF Schema. [Электронный ресурс]. — URL: http://www.w3.org/TR/rdf-schema/(дата обращения: 15.04.2011).

[2] Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. Jena: implementing the Semantic Web recommendations. // Proc. of the 13th Int. World Wide Web Conf. on Alternative Track Papers & Posters. USA, 2004. Pp. 74–83.

[3] Ding, L., Kolari, P., Ding, Z., Avancha, S., Finin, T., and Joshi, A. Using ontologies in the Semantic Web: a survey. // Ontologies: a handbook of principles, concepts and applications in information systems. Springer US, 2006. Pp. 79–114.

[4] Dobrynin, V., Patterson, D. W., and Rooney, N. Contextual Document Clustering. // Proc. of ECIR. 2004. Pp. 167–180.

[5] Gennari, J. H., Musen, M. A., Fergerson, R. W., Grossolo, W. E., Crubézy, M., Eriksson, H., Noy, N. F., and Tu, S. W. The evolution of Protégé: an environment for knowledge-based systems development. // International Journal of Human–Computer Studies. 2003. Vol. 58. № 1. Pp. 89–123.

[6] Gruber, T. R. A translation approach to portable ontology specification. // Knowledge Acquisition.1993. Vol. 5. № 1. Pp. 199–220.

[7] Klyne, G., and Carroll, J. J. Resource Description Framework (RDF): concepts and abstract syntax. [Электронный ресурс]. — URL: http://www.w3.org/TR/rdf-concepts/ (дата обращения: 15.04.2011).

[8] Noy, N. F., and McGuinness, D. L. Ontology Development 101: a guide to creating your first ontology // Technical Report KSL-01-05 and Stanford Medical Informatics Technical Reports SMI-2001-0880. 2001. Pp. 1–25.

[9] OWL 2 Web Ontology Language document overview. [Электронный ресурс]. — URL: http://www.w3.org/TR/owl2-overview/,(дата обращения: 15.04.2011).

[10] Pantel P., and Lin D. A statistical corpus-based term extractor. // Proc. of Canadian Conf. on AI. 2001. Pp. 36–46.

[11] Syafrullah, M., and Salim, N. Improving Term Extraction Using Particle Swarm Optimization Techniques. // Journal of Computer Science. 2010. Vol. 6. № 3. Pp. 323–329.

[12] Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. J. Text Mining: predictive methods for analyzing unstructured information. Springer, 2005.

[13] Гаврилова Т. А., Хорошевский В. Ф. Базы знаний интеллектуальных систем: учеб. для вузов. — СПб.: Питер, 2000. — 384 с.

[14] Добров Б. В., Лукашевич Н. В., Синицин М. Н., Шапкин В. Н. Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска. // Труды 7й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL‘2005. — Ярославль, 2005.

[15] Захарова И. В. Математическая модель семантического поиска с использованием онтологического подхода: Автореф. дис. канд. физ.–мат. наук. — Челябинск, 2009. — 20 с.

[16] Клещев А. С., Артемьева И. Л. Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология». // Научно–техническая информация, серия 2 «Информационные процессы и системы». — М.: ВИНИТИ, 2001. — № 2. — С. 20–27.

[17] Найханова Л. В. Методы и модели автоматического построения онтологий на основе генетического и автоматного программирования: Автореф. дис. докт. тех. наук. — Красноярск, 2008. — 36 с.

[18] Рабчевский Е. А. Автоматическое построение онтологий. // Научно–технические ведомости Санкт–Петербургского государственного политехнического университета. — СПб.: Издательство Политехнического Университета, 2007. — № 52–2. — С. 22–26.

[19] 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.

[20] Россеева О. И., Загорулько Ю. А. Организация эффективного поиска на основе онтологий. // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. — Аксаково, 2001. — Т. 2. — С. 333–342.

Automatic Ontology Learning from Text Document Collection

© Elena St. Mozzherina

The paper considers theoretical aspects of automatic ontology learning problem from text document collection where documents belong to the same application domain. The suggested approach is mostly based on statistical methods of natural language texts analysis.