УДК 37.013 **DOI 10.34670/AR.2019.90.8.030**

Подход к обучению онтологий на основе анализа метаданных и построения универсальных зависимостей

Волчек Дмитрий Геннадьевич

Преподаватель,
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
197101, Российская Федерация, Санкт-Петербург, просп. Кронверкский, 49;
e-mail: dvolchek@yandex.ru

Аннотация

Информатизация и стремительный рост количества данных приводят к неизбежным нуждам по обработке и интеллектуальному анализу данных. Построение онтологических моделей позволяет моделировать различные предметные области, интегрировать данные из различных источников и представлять их в виде, удобном не только для людей, но и читаемом машинами. Подход к созданию онтологий не посредством работы эксперта предметной области, а основываясь непосредственно на самих данных получил название обучение онтологий. В настоящей статье рассматривается подход к обучению онтологий, основанный на анализе метаданных исходных документов, а также предлагается метод построение связей между концептами предметной области на основе анализа универсальных зависимостей. Создание онтологий различных предметных областей – весьма перспективное направление деятельности. При этом в последнее время наблюдается растущий интерес к этой технологии со стороны бизнеса. А использование методов обучения онтологий позволит удовлетворить эту потребность, минимизируя при этом как временные, так и человеческие ресурсы. Улучшение существующих и разработка новых алгоритмов, используемых при обучении онтологий позволит выйти на новый, более качественный этап цифровизации и автоматизации человеческой деятельности.

Для цитирования в научных исследованиях

Волчек Д.Г. Подход к обучению онтологий на основе анализа метаданных и построения универсальных зависимостей // Экономика: вчера, сегодня, завтра. 2019. Том 9. № 8А. С. 298-304. DOI 10.34670/AR.2019.90.8.030

Ключевые слова

Онтологии, обучение онтологий, автоматическое создание онтологий, обработка текстов на естественном языке, обучение.

Введение

Объем информации, представленной в электронном виде, продолжает увеличиваться. Во всех сферах человеческой деятельности информационные технологии занимают весьма существенную позицию. В связи с этим, возникает запрос на обработку и интеграцию большого количества данных, причем зачастую из различных источников. Данные при этом могут быть представлены как в структурированном, так и неструктурированном виде. В частности, повышается сложность программной обработки и анализа данных, представленных в виде текстов на естественных языках.

Одним из решений, которое используется все более и более обширно, является создание онтологических моделей. Такие модели могут использоваться в качестве просто формального описания некоторой предметной области, или служить для сугубо прикладных целей, например для интеграции данных из различных источников, построения базы знаний, на основе которой будут работать другие приложения или в качестве основы, опираясь на которую разработчики могли бы создавать программные решения. Для создания онтологии конкретной предметной области необходимо хорошо понимать устройство этой области: какие концепты используются, что их связывает, какие существуют ограничения и так далее.

Построение онтологии человеком требует объемных временных затрат и экспертных знаний в предметной области, помимо этого, объемы информации настолько возросли, что зачастую одного эксперта недостаточно, а привлечение нескольких людей невозможно по тем или иным причинам. Поэтому переход от ручного создания онтологий к data-driven подходу (создание онтологий на основе данных) имеет высокую актуальность.

Данные, используемые для создания онтологий, могут быть как структурированными, так и неструктурированными. В первом случае задача сводится к созданию соответствующих аксиом непосредственно в модели. При этом существуют готовые инструменты для такого отображения, например, для трансформации из SQL в RDF. В случае неструктурированных данных подход иной. Автоматическое построение онтологии по нескольким текстовым документам преимущественно на основе статистических методов анализа текстов позволяет частично или полностью отказаться от экспертов в заданной области и существенно ускоряет процесс извлечения знаний из текстового корпуса.

Стоит отметить, что онтологии имеют широкую сферу применения. На основе онтологий работают чат-боты в различных мессенджерах, онлайн магазинах. Онтологии используются для моделирования предметных областей и создания умных производств. Также онтологии лежат в основе семантической сети, которая позволяет осуществлять интеграцию и объединение различных источников информации.

Обучение онтологий

Процесс формирования онтологий на основе неструктурированного контента (в большинстве случаев текста) получил название обучение онтологий (по аналогии с машинным обучением). Среди зарубежных авторов обучение онтологий используется достаточно давно и является вполне сформированной технологией. Работ, посвященных этой теме на русском языке, а особенно описывающих использование текста на русском языке как источник данных, сравнительно немного. При этом большинство из них выполняют обзорную функцию, без детального описания рассматриваемого метода.

В работе [Asim et al., 2018] описана технология и опыты различных исследователей в области обучения онтологий. Рассмотрены основные технологии, используемые для извлечения концептов, поиска синонимов, построения связей и аксиом. Исследования, рассмотренные в этой статье, относятся к различным предметным областям, в том числе производство, туризм, финансы и так далее. При этом источниками данных для обучения онтологий являются тексты на английском, китайском, испанском и других языках. Среди описываемых исследований работа с документами на русском языке отсутствует.

В работе [Ярушкина, Мошкин, 2016] описан метод обучения онтологий на основе использования механизма прецедентов для русского языка. Отдельно стоит упомянуть, что во многих работах, таких как [Ванюшкин, Гращенко, 2016] или [Kovriguina et al., 2017] решается локальная задача по извлечению концептов предметной области или построению связей, но не весь процесс обучения онтологий. Помимо этого, в работе [Ванюшкин, Гращенко, 2016] описан метод построения универсальных зависимостей для русского языка с использованием библиотеки CoreNLP.

Подробный хронологический обзор эволюции технологий обучения онтологий и рассуждения о существующих тенденциях на ближайшее будущее описаны в статье [Wong et al., 2012].

Стоит отметить, что в основе обучения онтологий лежат технологии обработки текстов на естественном языке, подробное описание и классификации которых для русского и английского языка описаны в [Корсун, Пальчунов, 2016] и [Aggarwal, 2012].

В работах, посвященных обучению онтологий, большинство авторов придерживается приблизительно одинаковой последовательности действий. Первоначальным этапом построения онтологий является выделение терминов или ключевых слов из текста, а также синонимов. Комбинация синонимов и терминов образует концепты. Затем происходит образование таксономических и нетаксономических связей между концептами. Наконец, формируются аксиомы. При этом подходы, при помощи которых осуществляется тот или иной этап могут существенно различаться и включать как статистические, так и лингвистические методы.

Метаданные используются в качестве служебной информации о различных объектах. Это могут быть геометки, метки времени, логгирование событий, записи действия пользователей и так далее. Использование метаданных в том или ином виде рассматривается в работах [Когаловский, 2012; Абросимов, 2005; Когаловский, 2013]. Основываясь на дополнительной информации относительно документов, становится возможным делать некоторые выводы, которые прямо не следуют из содержимого самого документа.

В настоящей статье предлагается подход к решению проблемы автоматического построения онтологий, преимущественно основанный на статистических методах анализа текстов на естественном языке, так как такой подход не требует привязки к определенной предметной области и использовании метаданных используемых текстовых документов. Помимо этого, предлагается метод построения нетаксономических отношений между концептами на основе механизма универсальных зависимостей.

Этапы обучения онтологий

Процесс обучения онтологий на основе неструктурированной информации считается относительно устоявшимся и получил название «слоеный пирог» обучения онтологий. Основные этапы обучения представлены на рисунке 1.

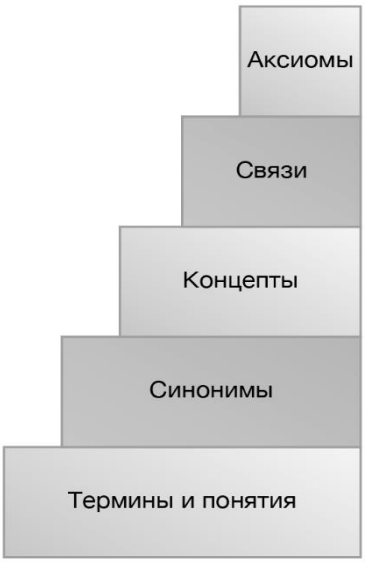

Рисунок 1 - «Слоеный пирог» обучения онтологий

Начальный этап предусматривает поиск в корпусе текстовых документов ключевых слов или терминов методами обработки естественных языков (Natural Language Processing). Осуществляется тегирование смысловых конструкций языка и помечаются части речи слов в предложениях.

Второй этап заключается в нахождении синонимов к найденным терминам с помощью вышеупомянутых методов.

Третий этап – получение концептов посредством объединения терминов и синонимов.

Четвертый этап – устанавливаются связи между концептами. Его можно разделить на 2 направления: построение таксономических и нетаксономических отношений. Формирование таксономических отношений между концептами возможно осуществить при помощи иерархической кластеризации. Этот метод чаще всего реализуется на основе использования мер подобия (Косинусное сходство или сходство Жаккара), с целью сгруппировать концепты в кластеры и построить иерархию. Для построения неткасономических отношений рассматриваются методы формального анализа концептов и поиск ассоциативных правил.

Пятый этап – предполагает использование индуктивного логического программирования для создания аксиом на основе полученных концептов, связей между концептами и возможными ограничениями (правилами).

Метод обучения онтологий с использованием метаданных и алгоритма построения универсальных зависимостей

Часто исходные данные для построения онтологической модели хранятся в некоторой системе управления контентом. Это может быть система управления производством, содержащая нормативные документы, платформа онлайн-образования, где размещены тексты лекций и так далее. При этом сами данные могут быть представлены в неструктурированном виде (в виде текста). В таком случае можно извлечь не только контент, но и метаданные рассматриваемых документов.

Описываемый метод строится на предположении, что метаданные косвенно могут описывать информативность и полезность того или иного документа в контексте его

использования для обучения онтологий. В частности, количество обращений к документу может служить индикатором его полезности в контексте обучения онтологий. Тогда доля концептов, извлеченных из этого документа, должна быть выше. Аналогичную процедуру можно использовать и для других метаданных. В общем виде для этих целей можно использовать многомерную линейную регрессию.

$$f(x,\alpha) = \sum_{j=1}^{n} \alpha_j f_j(x)$$

Где входными параметрами являются значения метаданных рассматриваемого документа, веса подбираются в результате обучения модели, а выходным параметром может служить вклад документа в общее пространство терминов. Для извлечения же терминов можно рассматривать различные механизмы экстракции ключевых слов, в зависимости от эффективности того или иного метода на конкретных рассматриваемых данных.

Построение нетаксономических отношений весьма трудоемкая задача. Мы предлагаем использование универсальных зависимостей. В силу того, что в результате этапа извлечения концептов для каждого документа сформирован стек используемых терминов, в качестве связей между ними можно использовать их отношения, описанные непосредственно в тексте.

Метод построения универсальных зависимостей реализуется посредством семантического разбора предложения и установления зависимостей между акторами. Пример построения универсальных зависимостей представлен на рисунке 2.

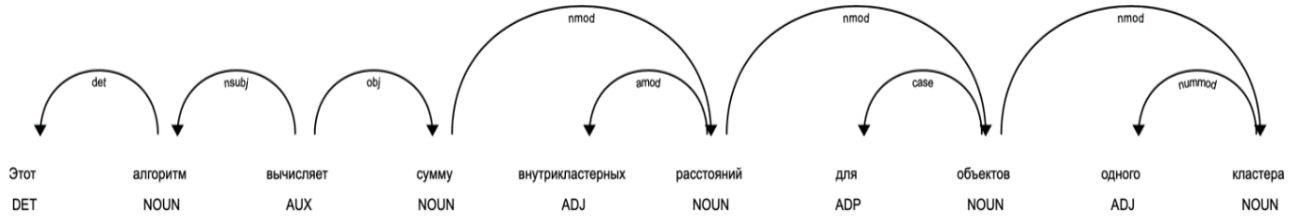

Рисунок 2 - Построение универсальных зависимостей

В результате построения такого списка зависимостей становится возможным обучить модель для установления нетаксономических отношений между концептами, извлеченными на первом этапе. При этом поиск таксономических отношений и построение аксиом возможно выполнять различными методами, в зависимости от их успешности на данных в определенной предметной области.

Заключение

Создание онтологий различных предметных областей – весьма перспективное направление деятельности. При этом в последнее время наблюдается растущий интерес к этой технологии со стороны бизнеса. А использование методов обучения онтологий позволит удовлетворить эту потребность, минимизируя при этом как временные, так и человеческие ресурсы. Улучшение существующих и разработка новых алгоритмов, используемых при обучении онтологий позволит выйти на новый, более качественный этап цифровизации и автоматизации человеческой деятельности.

Библиография

1. Абросимов А.Г. Метаданные описания коллекции периодической печати // Электронные библиотеки. 2005. Т. 8. №. 2. С. 1-7.
2. Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. №19. С. 85-93.
3. Когаловский М.Р. Метаданные в компьютерных системах // Программирование. 2013. Т. 39. №4. С. 28-46.
4. Когаловский М.Р. Метаданные, их свойства, функции, классификация и средства представления // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Переславль-Залесский, 2012. URL: elib.ict.nsc.ru/jspui/bitstream/ICT/1175/1/kogalovsky-meta.pdf
5. Корсун И.А., Пальчунов Д.Е. Теоретико-модельные методы извлечения знаний о смысле понятий из текстов естественного языка // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2016. Т. 14. №3. С. 34-48.
6. Ярушкина Н.Г., Мошкин В.С. Подход к обучению онтологии на основе гибридизации алгоритмов извлечения знаний из текстов и механизма прецедентов // Вестник Ростовского государственного университета путей сообщения. 2016. №2. С. 78-83.
7. Aggarwal C.C., Zhai C.X. (ed.) Mining text data. Springer Science & Business Media, 2012. 534 p.
8. Asim M.N. et al. A survey of ontology learning techniques and applications. URL: https://www.ncbi.nlm.nih.gov/pubmed/30295720
9. Kovriguina L. et al. Russian tagging and dependency parsing models for stanford CoreNLP natural language toolkit // International Conference on Knowledge Engineering and the Semantic Web. Springer, Cham, 2017. P. 101-111.
10. Wong W., Liu W., Bennamoun M. Ontology learning from text: A look back and into the future // ACM Computing Surveys (CSUR). 2012. Vol. 44. №4. P. 20.

An approach to training ontologies based on the analysis of metadata and the construction of universal dependencies

Dmitrii G. Volchek

Lecturer,
Saint Petersburg National Research University of Information Technologies, Mechanics and Optics,
197101, 49, Kronverskii av., Saint Petersburg, Russian Federation;
e-mail: dvolchek@yandex.ru

Abstract

Informatization and the rapid growth in the amount of data lead to the inevitable needs for data processing and data mining. Building ontological models allows you to simulate various subject areas, integrate data from various sources and present them in a form convenient not only for people, but also readable by machines. The approach to creating ontologies is not through the work of an expert in a subject field, but based directly on the data themselves, is called ontology training. This article discusses an approach to training ontologies based on an analysis of the metadata of the source documents, and also proposes a method for constructing relationships between domain concepts based on the analysis of universal dependencies. The creation of ontologies of various subject areas is a very promising area of activity. At the same time, there has recently been a growing interest in this technology on the part of business. And the use of ontology training methods will satisfy this need, while minimizing both temporary and human resources. Improving existing and developing new algorithms used in training ontologies will allow us to enter a new, better stage of digitalization and automation of human activity, conclude the author of this research.

For citation

Volchek D.G. (2019) Podkhod k obucheniyu ontologii na osnove analiza metadannykh i postroeniya universal'nykh zavisimostei [An approach to training ontologies based on the analysis of metadata and the construction of universal dependencies]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 9 (8A), pp. 298-304. DOI 10.34670/AR.2019.90.8.030

Keywords

Ontologies, ontology training, automatic ontology creation, natural language text processing, training.

References

1. Abrosimov A.G. (2005) Metadannye opisaniya kollektsii periodicheskoi pechati [Metadata for the collection of periodicals]. *Elektronnye biblioteki* [Electronic Libraries], 8, 2, pp. 1-7.
2. Aggarwal C.C., Zhai C.X. (ed.) (2012) *Mining text data.* Springer Science & Business Media.
3. Asim M.N. et al. *A survey of ontology learning techniques and applications*. Available at: https://www.ncbi.nlm.nih.gov/pubmed/30295720 [Accessed 06/06/2019]
4. Kogalovskii M.R. (2013) Metadannye v komp'yuternykh sistemakh [Metadata in computer systems]. *Programmirovanie* [Programming], 39, 4, pp. 28-46.
5. Kogalovskii M.R. (2012) Metadannye, ikh svoistva, funktsii, klassifikatsiya i sredstva predstavleniya [Metadata, their properties, functions, classification and means of presentation]. In: *Trudy 14-i Vserossiiskoi nauchnoi konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii»* [Proceedings of the 14th All-Russian Scientific Conference: Electronic Libraries: Advanced Methods and Technologies, Electronic Collections]. Pereslavl-Zalesskii. Available at: elib.ict.nsc.ru/jspui/bitstream/ICT/1175/1/kogalovsky-meta.pdf [Accessed 06/06/2019]
6. Korsun I.A., Pal'chunov D.E. (2016) Teoretiko-model'nye metody izvlecheniya znanii o smysle ponyatii iz tekstov estestvennogo yazyka [Model-theoretic methods for extracting knowledge about the meaning of concepts from natural language texts]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii* [Bulletin of the Novosibirsk State University. Series: Information Technology], 14, 3, pp. 34-48.
7. Kovriguina L. et al. (2017) Russian tagging and dependency parsing models for stanford CoreNLP natural language toolkit. In: *International Conference on Knowledge Engineering and the Semantic Web*. Springer, Cham.
8. Vanyushkin A.S., Grashchenko L.A. (2016) Metody i algoritmy izvlecheniya klyuchevykh slov [Keyword extraction methods and algorithms]. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh* [New information technologies in automated systems], 19, pp. 85-93.
9. Yarushkina N.G., Moshkin V.S. (2016) Podkhod k obucheniyu ontologii na osnove gibridizatsii algoritmov izvlecheniya znanii iz tekstov i mekhanizma pretsedentov [An approach to learning ontology based on hybridization of algorithms for extracting knowledge from texts and the use-case mechanism]. *Vestnik Rostovskogo gosudarstvennogo universiteta putei soobshcheniya* [Bulletin of the Rostov State University of Railway Engineering], 2, pp. 78-83.
10. Wong W., Liu W., Bennamoun M. (2012) Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44, 4, pp. 20.