Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

на правах рукописи

Герасимова Ольга Александровна

Онтологический доступ к данным с использованием дизъюнктивных аксиом

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата компьютерных наук

Москва – 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ).

Научный руководитель: Кузнецов Сергей Олегович, доктор технических наук, профессор, департамент анализа данных и искусственного интеллекта, факультет компьютерных наук, НИУ ВШЭ.

Содержание

1 Введение

В настоящее время онтологии широко используются для того, чтобы повысить удобство организации информации и доступа к ней в таких областях, как искусственный интеллект, программная инженерия, биомедицинская информатика, здравоохранение, индустриальные проекты и др. Поиск ответов на различные запросы, опосредованные онтологией в терминах дескрипционной логики (ДЛ), рассматривается как важная проблема логического вывода рассуждений в области представления знаний с начала 1990-х годов. Распространение ДЛ и её приложений, развитие языка веб-онтологии OWL (Web Ontology Language)[1] и в особенности парадигмы доступа к данным на основе онтологий [23] сделали теорию и практику поиска ответов на запросы, опосредованных онтологией, которые состоят из пары онтология плюс запрос, актуальной областью исследований, находящейся на стыке следующих направлений: представления знаний и вывода рассуждений, семантических технологий и семантической паутины, графов знаний, а также теории и технологий баз данных.

Настоящая диссертация направлена на изучение практического потенциала конкретной онтологии в задаче нахождения ответа на запрос, опосредованного онтологией, и фокусируется на проблемах, связанных с выбранным конкретным типом онтологии. Наше исследование предоставляет понимание сложности использования выбранной онтологии для доступа к данным на основе онтологий и позволяет проанализировать возможности расширения существующей практики в этой области.

1.1 Объект исследования

Мы сосредоточимся на подходе доступа к данным на основе онтологий (OBDA – Ontology-Based Data Access) [23, 26], рассмотрев более выразительную онтологию, чем базово разрешает OBDA, где онтология используется как полезный инструмент для организации поиска ответов на запросы к распределенным и гетерогенным источникам данных.

[1] https://www.w3.org/TR/owl2-overview/

Стандартный сценарий OBDA следует определенной последовательности действий (см. Рис.1), в которой:

– конечный пользователь не участвует в первоначальной организации данных;

– пользователю предоставляется онтология, разработанная экспертом в данной области, которая определяет понятия и отношения, знакомые пользователю, и предоставляет словарь для запросов пользователя;

– схемы данных (к примеру, реляционные таблицы) преобразуются в хранилище данных в терминах словаря онтологии посредством мэппинга (кодируется в таких языках, как R2RML – Relational databases (RDB) to Resource Description Framework (RDF) Mapping Language);

– задача системы OBDA – преобразовать, или *переписать*, запрос пользователя с помощью онтологии и мэппингов в эквивалентный стандартный запрос к данным и ответить на новый запрос к базе данных уже без учета онтологии.

Таким образом, онтология обеспечивает высокоуровневое концептуальное представление данных, может дополнять данные базовыми знаниями и поддерживать поиск ответов на запросы к нескольким разнородным наборам данных. Чтобы этот подход был эффективен, необходимо тщательно выбирать язык онтологии. В большинстве своем мы рассматриваем онтологии, сформулированные в терминах соответствующей дескрипционной логики.

Для обеспечения теоретической и практической применимости, парадигма OBDA предполагает, что пользовательские онтологические запросы переформулируются, или переписываются, системой OBDA в стандартные запросы к базе данных, которые достаточно эффективно обрабатываются существующими системами управления базами данных. Возможно ли вообще такое переписывание и на какой язык запросов, естественным образом зависит от конкретного онтологического запроса. Одним из способов *единообразно* гарантировать желаемую переписываемость является ограничение языка онтологи-

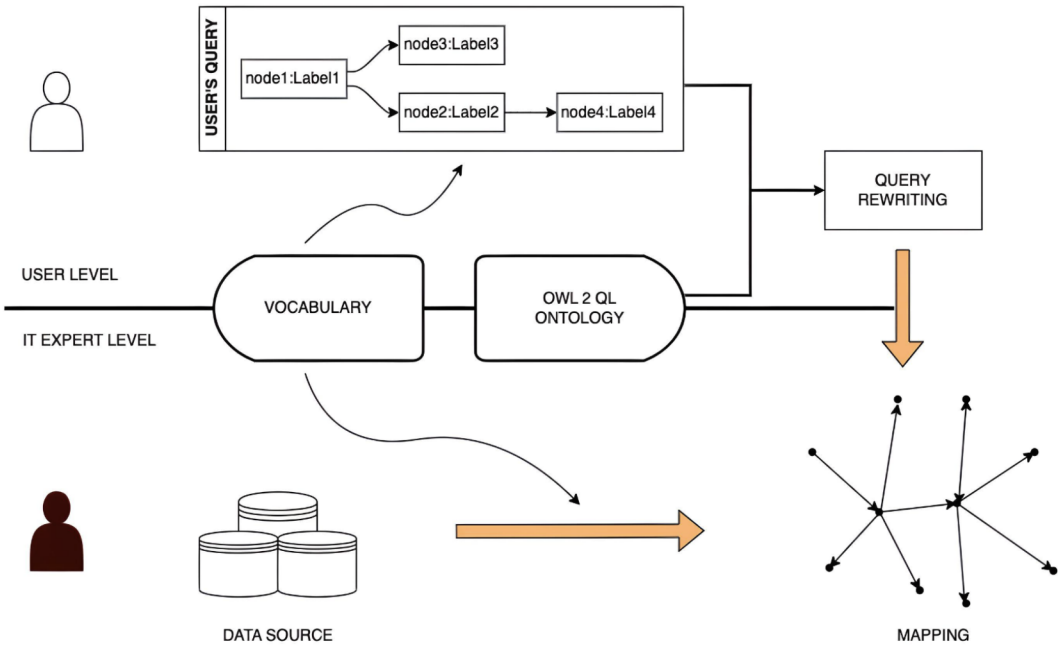


Рис. 1: Схема работы системы OBDA. Пользователь (эксперт в предметной области) формулирует запрос с учетом знания словаря и правил онтологии, описывающих зависимости данных в терминах словаря. ИТ-специалист преобразует неоднородные данные в мэппинги (mappings) с использованием заданного словаря. Такие мэппинги представляют данные в форме графа, согласованного с точки зрения синтаксиса с онтологией. Наконец, OBDA система генерирует переписывание запроса в эквивалентный стандартный запрос, который задается относительно мэппингов без учета онтологии.

ческих запросов. Таким образом, один из профилей *OWL 2 QL* [2] онтологического веб-языка *OWL 2* был разработан так, чтобы гарантировать переписываемость *всех* онтологических запросов с онтологией *OWL 2 QL* и конъюнктивных запросов (CQ – Conjunctive Query) в запросы в виде формулы первого порядка, или первопорядковые (FO – first-order) запросы, то есть, по сути, в SQL-запросы. С точки зрения теории сложности, переписываемость онтологического запроса в запросы первого порядка означает, что на него можно ответить за LogTime в AC^0 – один из наименьших классов сложности.

В нашем исследовании мы фокусируем наше внимание на проблеме опре-

[2] https://www.w3.org/TR/owl2-profiles/

деления вычислительной сложности и переписываемости при работе с конкретной онтологией, которая расширяет выразительность OBDA. Мы рассматриваем онтологию, включающую в себя единственное правило, называемое *покрывающей аксиомой* – $\{A \sqsubseteq F \sqcup T\}$, которое не описывается на языке *OWL 2 QL*. Исследования подобной онтологии обусловлены практическими потребностями, так как идея онтологии широко используется для описания реальных данных в социальных сетях, промышленных процессах, принятии решений и т.д.

1.2 Предмет исследования

Предметом нашего исследования является покрывающая аксиома, которая утверждает, что один класс покрывается объединением двух других классов. Актуальность рассматриваемого предмета заключается в его значительном влиянии на расширение выразительности OBDA в случаях, не подходящих для переписываний в стандартные SQL-запросы по данным. Проблема поиска ответа на онтологически опосредованные запросы с покрывающей аксиомой связывает между собой теорию графов, проблему удовлетворения ограничениям (CSP – Constraint Satisfaction Problem) и другие области теоретической информатики. Основная проблема заключается в эффективном определении сложности задачи, и, если она относится к вычислимому случаю, полезному для практического применения, в предоставлении алгоритма для переписывания онтологического запроса в запрос первого порядка или в datalog-программу. Однако даже для простого случая с одним правилом в дизъюнктивной онтологии вычислительная задача поиска ответа на онтологический запрос остается очень сложной.

В рамках исследования мы изучаем только Булевы конъюнктивные запросы, которые дают ответ без определения конкретных индивидуумов в данных, что как раз и необходимо из-за особенности аксиомы покрытия, не допускающей однозначной модели данных. Древовидная форма запросов была выбрана в процессе анализа сложности поставленной задачи и выбранной методологии определения сложности по данным.

Таким образом, мы сосредоточились на Булевых древовидных конъюнк-

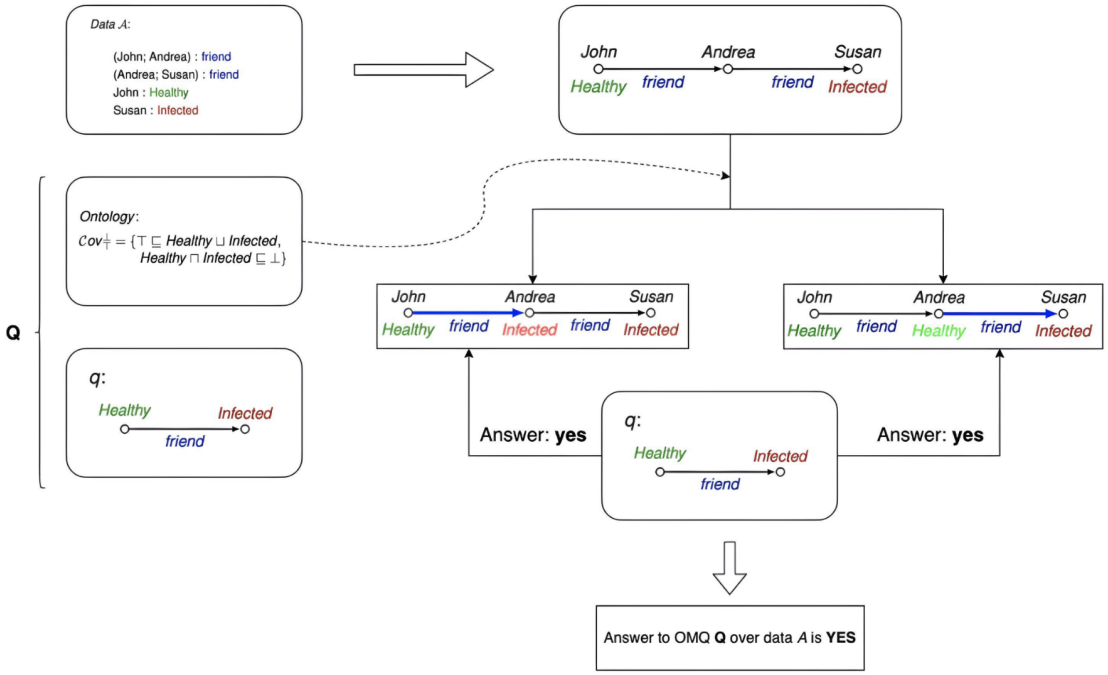


Рис. 2: Пример онтологически опосредованного запроса Q с онтологией, содержащей покрывающую аксиому. Простой запрос q дает отрицательный ответ относительно исходных данных $\mathcal{A}$. Однако, учитывая все возможные присваивания меток с использованием онтологии, ответом во всех полученных моделях данных является 'да'. Следовательно, ответом на онтологический запрос также является 'да'.

тивных запросах, опосредованных аксиомой покрытия, и исследовали для них проблемы определения сложности по данным и переписываемости в рамках расширения OBDA. Пример определения ответа на онтологический запрос представлен на Рис.2.

1.3 Задачи и цели исследования

Целью нашего исследования является проблема эффективного определения сложности по данным для поиска ответов на запросы, опосредованных онтологиями дескрипционной логики, и построения для них оптимальных переписываний в стандартные запросы к базам данных. Известно, что эта проблема в целом очень сложна с вычислительной точки зрения и для нее не определены

явные синтаксические характеристики на запросы. В то же время, она является очень важной и актуальной для доступа к данным на основе онтологий и datalog оптимизации. В нашем исследовании мы сосредоточились на Булевых конъюнктивных запросах, опосредованных простой покрывающей аксиомой, которая показывает, что каждый элемент класса принадлежит (хотя бы) одному из двух других классов.

1.3.1 Сложность по данным

Мы рассматриваем *сложность по данным* при поиске ответов на онтологически опосредованные Булевы конъюнктивные запросы с аксиомой покрытия:

(d-sirup) $\boldsymbol{Q} = (\mathsf{cov}_A, \boldsymbol{q})$, где $\mathsf{cov}_A = \{ A \sqsubseteq F \sqcup T \}$ – онтология и $\boldsymbol{q}$ – Булев конъюнктивный запрос с унарными предикатами F, T и произвольными бинарными предикатами.

Под *сложностью по данным* мы понимаем вычислительную сложность поиска ответа на $\boldsymbol{Q}$ с заданной онтологией и фиксированным конъюнктивным запросом, но относительно любого набора входных данных $\mathcal{A}$ при семантике открытого мира.

Основным изменяемым фактором, который влияет на сложность данных в нашей задачей, является структура запроса. Таким образом, мы стремимся понять, как взаимодействие между *покрывающей аксиомой* $A \sqsubseteq F \sqcup T$ и структурой $\boldsymbol{q}$ определяет сложность $\boldsymbol{Q}$.

Отдавая предпочтение терминам из теории графов для $\boldsymbol{q}$ и наборов данных из-за их удобства при анализе, моделировании и визуализации в виде размеченных ориентированных графов, мы можем сформулировать задачу ответа на $\boldsymbol{Q}$ следующим образом:

Входные данные: любой размеченный ориентированный граф $\mathcal{A}$.

Постановка задачи: определить, содержит ли каждый ориентированный граф, полученный путем разметки каждой вершины A в $\mathcal{A}$ как F или T, гомоморфный образ $\boldsymbol{q}$ (в этом случае ответ на $\boldsymbol{Q}$ относительно $\mathcal{A}$ будет ‘да’).

Задача, сформулированная выше, может быть выполнена в coNP [1], поскольку $\boldsymbol{q}$ *фиксирован* и поэтому существование гомоморфизма из $\boldsymbol{q}$ в любую разметку $\mathcal{A}$ может быть проверено за полиномиальное время путем проверки всех возможных $|\mathcal{A}|^{|\boldsymbol{q}|}$-отображений из $\boldsymbol{q}$ в $\mathcal{A}$. Для решения данной проблемы мы можем рассмотреть применение доказательства на основе метода резолюций или проверки дизъюнктивной datalog-программы $\{(1), (2)\}$, приведенной ниже, над $\mathcal{A}$. Однако оба метода могут повлечь за собой определение доказательств экспоненциального размера.

Таким образом, выбор проблемы определения сложности по данным как предмета нашего исследования приводит к следующим вопросам:

– существует ли альтернативное, более эффективное алгоритмическое решение для данного $\boldsymbol{Q}$ в принципе, и

– может ли оно быть выполнено как стандартная (линейная, симметричная) datalog-программа или запрос первого порядка, заданный относительно входных графов $\mathcal{A}$.

1.3.2 Переписываемость

Здесь необходимо поднять вопрос о второй проблеме – *переписываемости*, о которой мы уже упоминали.

Под *переписываемостью* мы подразумеваем сведение задачи нахождения конкретных ответов на $\boldsymbol{Q}$ относительно любых входных данных $\mathcal{A}$ к задаче поиска ответов на обычный запрос $\boldsymbol{Q}'$ напрямую к базе данных $\mathcal{A}$ с оптимальной сложностью по данным. В таком случае запрос $\boldsymbol{Q}'$ называется *переписыванием* онтологически опосредованного запроса $\boldsymbol{Q}$.

В терминах обозначений для datalog языка, онтологический запрос $\boldsymbol{Q} = (\mathsf{cov}_A, \boldsymbol{q})$ эквивалентен *монадическому дизъюнктивному datalog-запросу*

$$T(x) \vee F(x) \leftarrow A(x) \tag{1}$$

$$\boldsymbol{G} \leftarrow \boldsymbol{q} \tag{2}$$

с целевым предикатом нулевой арности $\boldsymbol{G}$. В 1980-х годах, пытаясь понять ограниченность (или FO-rewritability – переписываемость в запросы логики

первого порядка) и линеаризуемость (или linear-datalog-rewritability – переписываемость в линейную datalog-программу) datalog-запросов, сообщество исследователей баз данных ввело понятие *sirup* – '*datalog-запрос с единственным рекурсивным правилом*' [27, 17], которое считалось ключевым для понимания datalog рекурсии и оптимизации datalog-программ [21]. Наши онтологические запросы $\boldsymbol{Q}$, или дизъюнктивные datalog-запросы $(\{(1), (2)\}, \boldsymbol{G})$, которые в дальнейшем будут называться (*монадическими*)*дизъюнктивными sirups* или просто *d-sirups* (disjunctive sirups) – играют одну и ту же фундаментальную роль для понимания онтологических запросов с выразительными онтологиями и монадических дизъюнктивных datalog-запросов.

d-sirups могут показаться синтаксически простыми, но на самом деле они принадлежат к чрезвычайно сложному классу онтологических запросов. Например, задача о разрешимости первопорядковой переписываемости d-sirups оказывается 2ExpTime-трудной – такой же сложной, как задачи разрешимости ограниченности произвольных монадических datalog-программ [7, 3]. Стоит отметить, что одним из источников этой неожиданно высокой сложности оказываются 'сдвоенные' FT-метки вершин в конъюнктивных запросах, при условии что атомы $F(x),\ T(x) \in \boldsymbol{q}$, для некоторой переменной x. Мы можем устранить этот источник сложности, задавая стандартное *ограничение на непересекаемость* $F \sqcap T \sqsubseteq \bot$ (или $\bot \leftarrow F(x), T(x)$ в терминах datalog), часто используемое в онтологиях и концептуальном моделировании. Таким образом, мы приходим к *dd-sirups* (disjoint disjunctive sirups) вида

(dd-sirup) $\boldsymbol{Q} = (\mathsf{cov}^{\bot}_{A}, \boldsymbol{q})$, где $\mathsf{cov}^{\bot}_{A} = \{\, A \sqsubseteq F \sqcup T,\ F \sqcap T \sqsubseteq \bot \,\}$.

В заключение этих двух подразделов мы хотели бы подчеркнуть, что сложность и переписываемость d- и dd-sirups зависит только от структуры конъюнктивных запросов $\boldsymbol{q}$, что приводит к исследованию классификации (d)d-sirups в зависимости от типа графа, лежащего в основе $\boldsymbol{q}$ – направленный путь, дерево, их неориентированные варианты и т.д. – и определению соответствующих классов сложности по данным и переписываемости, если возможно, онтологических запросов в полученных сложностных классах.

1.3.3 Сравнение двух подходов поиска ответов на онтологические запросы на основе логических ризонеров для datalog-программ и на основе разметки данных методами машинного обучения

В дополнение к теоретическим исследованиям мы проводим эксперименты на основе машинного обучения, чтобы рассмотреть нашу задачу в терминах задачи классификации вершин на графе. Используя современные модели графовых нейронных сетей, можно восстановить пропущенные метки данных, что даст нам возможность проанализировать ответы на запросы для различных наборов данных без использования онтологии. Взяв реальные социальные графы с вершинами, размеченными по двум классам, мы маскируем (удаляем оригинальные метки классов) от 5% to 95% всей разметки, а после присваиваем метки на неразмеченные вершины в соответствии с обучением графовой нейронной сети. Затем, мы сравниваем ответы на запросы к графу без маскированных меток с результатами, полученными после применения OBDA-подхода. Таким образом, поиск ответа на Булев конъюнктивный запрос с покрывающей аксиомой может зависеть не только от сложности по данным, о которой говорили ранее, но и от количества и расположения неразмеченных данных в графе.

Для проверки наших результатов в рамках OBDA подхода мы выбираем системы, работающие с дизъюнктивными datalog-программами, чтобы протестировать полученные datalog-переписывания для онтологических запросов с дизъюнктивной аксиомой. Затем, мы сравниваем время работы систем при исходной формулировке задачи (datalog-программа, состоящая из запроса и дизъюнктивного правила онтологии) с работой datalog-программы, соответствующей переписыванию.

Наконец, мы сравниваем два различных подхода, чтобы определить пригодность трудоемкого логического подхода, основанного на переписываемости, и графовых нейронных сетей для нашей задачи.

Основными задачами диссертационного исследования являются:

– доказательство условий принадлежности конкретной задачи поиска ответов на онтологический запрос с покрывающей аксиомой к определен-

ному классу сложности в зависимости от объема данных;
- определение критериев синтаксического разделения запросов на сложностные классы на основе структуры запроса;
- нахождение практически вычислимых случаев онтологических запросов с первопорядковой или datalog-переписываемостью;
- проведение анализа эффективности datalog-переписываний для наших онтологических запросов на реальных данных с использованием datalog-ризонеров;
- предоставление сравнительного анализа работы логических систем вывода рассуждений для поиска ответов на онтологические запросы с покрывающей аксиомой и их альтернатив — моделей машинного обучения для маскирования данных и поиска ответа путем задавания запроса к уже размеченным данным без учета онтологии.

Основной **целью нашей работы** является получение классификации онтологических запросов с покрывающей аксиомой, основанной на простых и прозрачных синтаксических требованиях к виду конъюктивного запроса, и определение теоретической и эмпирической сложности по данным при поиске ответов на такие онтологические запросы.

1.4 Основные результаты и выводы

В данном разделе описываются основной научный вклад, достигнутый в настоящей работе, ее новизна, теоретическая и практическая значимость, методология исследования и достоверность результатов.

1.5 Основные положения, выносимые на защиту

1. предоставленное сведение задачи обнаружения практически вычислимых случаев онтологических запросов с путевым конъюктивным запросом и онтологией с покрывающей аксиомой к проблеме проверки вычислимости задачи удовлетворения ограничениям (CSP) и тестирование программы Polyanna, работающей с CSP конструкциями, для выявления конкретных примеров практически вычислимых запросов [15];

2. полученная полная синтаксическая классификация (d)d-sirups ($\mathsf{cov}_A^{\perp}, \boldsymbol{q}$) с путевыми конъюнктивными запросами $\boldsymbol{q}$ с точки зрения их сложности по данным и типу переписываемости [13, 12];

3. представленный описательный анализ влияния неразмеченных данных на производительность работы дизъюнктивных ризонеров; проанализированная эффективность вывода рассуждений на основе логических ризонеров по сравнению с использованием графовых нейронных сетей в качестве альтернативы из области машинного обучения для решения задачи поиска ответа на онтологический запрос [16].

Научная новизна. Новизна работы раскрывается в различных аспектах. Во-первых, основная исследовательская мотивация заключается в том, как можно расширить использование OBDA для более выразительной онтологии. Таким образом, вклад заключается в том, что были определены практически вычислимые случаи запросов для онтологии с единственной покрывающей аксиомой, когда даже простая выразительная онтология требует сложных техник для ее анализа. Во-вторых, нами был предложен новый метод для отделения вычислимых случаев от невычислимых для ограниченного семейства запросов. Предлагаемая методология основана на сочетании datalog инструментов и методов теории автоматов. Кроме того, в процессе исследования практического расширения OBDA с точки зрения покрывающей аксиомой, были получены новые теоретические результаты для установления нижних границ сложности задачи для различных случаев. Мы обнаружили несколько необходимых или достаточных условий принадлежности к определенному классу сложности или определенному типу переписывания. Мы также предоставили практические ограничения предложенных алгоритмов по поиску ответов на онтологически опосредованные запросы с покрывающей аксиомой.

Теоретическая и практическая значимость. В настоящее время многие онтологии не соответствуют ограничениям, предписанным стандартными языками для OBDA. На практике, несовместимые аксиомы часто исключаются из онтологии в надежде, что не будет слишком значимого изменения в ответах на онтологические запросы.

В ответ на эту проблему, в нашем исследовании мы пытаемся выяснить, возможен ли другой исход для покрывающей аксиомы, потому что она является широко используемым правилом в онтологиях, описывающих реальные данные. Вдохновляясь результатами исследователей Лутц и Сабеллек [25] о семантической характеристике онтологических запросов с онтологией, описанной на языке OWL 2 EL, мы сделали для нашего расширения OBDA полезное открытие и затем разработали уже собственные новые техники, которые также послужат источником вдохновения для других будущих теоретических исследований в этой области и которые доказывают теоретическую значимость полученной тетрахотомии сложностных классов в терминах сложности по данным для задачи поиска ответов на онтологические запросы с покрывающей аксиомой.

Практическая значимость заключается не только в возможности использования полученных алгоритмов для поиска ответов на онтологические запросы, но и в практической оценке того, может ли логический подход для работы с онтологическими запросами быть заменен широко распространенными методами машинного обучения для разметки данных, рассматривая случаи различного количества неразмеченных данных. Полученные результаты показали значимость использования логических ризонеров по сравнению со статистическими подходами, актуальные для текущих систем без интеграции онтологий, но и их ограничения, когда размеченных данных слишком мало.

Методология исследования. Теоретическая часть исследования основана на теории вычислительной сложности, datalog оптимизации, теории графов, методах теории автоматов и дескрипционной логики. Практическая часть включает в себя использование графовых нейронных сетей, классических методов машинного обучения и статистических методов, а также тестирование ризонеров, поддерживающих работу с дизъюнтивными datalog-программами.

Достоверность результатов. Достоверность обеспечивается полными доказательствами теорем, обеспечивающими корректность результатов. Практические эксперименты включают в себя комплексные и детальные вычисления, учитывающие доверительные интервалы метрик и сравнение двух различных подходов – основанного на машинном обучении и основанного на логике.

Финансирование. Исследование выполнено при поддержке Факультета компьютерных наук НИУ ВШЭ; Российского Научного Фонда; Программы фундаментальных исследований НИУ ВШЭ; Российского Фонда Фундаментальных Исследований.

1.6 Публикации и апробация работы

1.6.1 Основные публикации

1. Gerasimova O., Kikot S., Zakharyaschev M. *Checking the Data Complexity of Ontology-Mediated Queries: A Case Study with Non-uniform CSPs and Polyanna*, in: Description Logic, Theory Combination, and All That. Berlin: Springer, 2019. P. 329-351 [15]

2. Gerasimova O., Kikot S., Kurucz A., Podolskii V. V., Zakharyaschev M. *A Data Complexity and Rewritability Tetrachotomy of Ontology-Mediated Queries with a Covering Axiom*, in: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning. IJCAI Organization: The International Joint Conference on Artificial Intelligence (IJCAI), 2020. P. 403-413. [13]

3. Gerasimova O., Kikot S., Podolskii V. V., Kurucz A., Zakharyaschev M. *A tetrachotomy of ontology-mediated queries with a covering axiom* // Artificial Intelligence. 2022. Vol. 309. Article 103738. [12]

4. Gerasimova O., Severin N., Makarov I. *Comparative Analysis of Logic Reasoning and Graph Neural Networks for Ontology-Mediated Query Answering with a Covering Axiom* // IEEE Access. 2023. P. 1-13. [16]

1.6.2 Дополнительные публикации

5. Gerasimova O., Podolskii V. V., Kikot S., Zakharyaschev M. *On the Data Complexity of Ontology-Mediated Queries with a Covering Axiom*, in: Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. Aachen: CEUR Workshop Proceedings, 2017. Ch. 19. P. 1-12. [11]

6. Gerasimova O., Kikot S., Podolskii V. V., Zakharyaschev M. *More on the Data Complexity of Answering Ontology-Mediated Queries with a Covering Axiom*, in: Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web. Berlin: Springer, 2017. P. 143-158. [14]

7. Gerasimova O., Kikot S., Zakharyaschev M. *Towards a Data Complexity Classification of Ontology-Mediated Queries with Covering*, in: Proceedings of the 31th International Workshop on Description Logics, Tempe, Arizona, October 27-29, 2018. Aachen: CEUR Workshop Proceedings, 2018. P. 1-13. [29]

1.6.3 Доклады на конференциях и семинарах.

1. International Workshop Logic Matters, Moscow, Russia, 28 December 2021. Topic: *A Tetrachotomy of Ontology-Mediated Queries with a Covering Axiom.*

2. International Workshop Logic Matters, Moscow, Russia, 29 December 2020. Topic: *Checking the Data Complexity of Ontology-Mediated Queries: A Case Study with Non-uniform CSPs and Polyanna.*

3. The 31th International Workshop on Description Logics, Tempe, Arizona, October 27-29, 2018. Topic: *Towards a Data Complexity Classification of Ontology-Mediated Queries with Covering.*

4. The 8th International Conference on Knowledge Engineering and Semantic Web (KESW), Szczecin, Poland, November 8-10, 2017. Topic: *More on the Data Complexity of Answering Ontology-Mediated Queries with a Covering Axiom.*

5. The 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. Topic: *On the Data Complexity of Ontology-Mediated Queries with a Covering Axiom.*

1.6.4 Описание личного вклада автора диссертации в статьях

В [15] мы сформулировали задачу объединения CSP и OBDA для онтологии, содержащей аксиому покрытия, на основе подхода из [5]. Автор диссертации

использовал предложенный метод, чтобы показать, что определение практически вычислимых путевых онтологических запросов с покрывающей аксиомой может быть сведено к проверке вычислимости CSP. Автор использовал программное обеспечение Polyanna, основанное на поиске полиморфизмов, для проверки вычислительной сложности шаблонов CSP, соответствующих конкретным онтологическим запросам, чтобы определить, можен ли задача поиска ответа на заданный онтологический запрос, содержащий конъюктивный запрос в виде пути с четырьмя переменными, быть в P или является coNP-трудной. Эксперименты с Polyanna помогают определить синтаксическое условие разделения P/coNP для наших онтологических запросов.

В [13, 12] мы изучили задачу получения прозрачных синтаксических критериев разделения d-sirups с непересекающимися покрывающими классами и путеобразными Булевыми конъюнктивными запросами с точки зрения сложности по данным и типа пересываемости, если это возможно. Автор получил несколько достаточных условий для принадлежности запросов одному из классов сложности AC^0, L, NL, P и coNP. В частности, вклад автора был сосредоточен на алгоритмах сведения нашей задачи к проблеме достижимости в графе, что соответствует классам сложности L/NL. Кроме того, автор выявил гомоморфизмы внутри структуры запросов, приводящие к определению критерия разделения между классами сложности L/NL и P.

В [16] автор поставил проблему сравнения логического подхода с популярными моделями машинного обучения, такими как графовые нейронные сети, обученные для классификации вершин, чтобы обеспечить анализ того, можно ли заменить работу с онтологическими запросами с аксиомой покрытия на разметку данных с помощью машинного обучения и последующее их использование для поиска ответа на конъюнктивный запрос без онтологии. Автор диссертации сформулировал цели экспериментов, провел анализ работы логических ризонеров, подготовил статью и руководил исследованием по сравнению логических ризонеров и графовых моделей для онтологических запросов с покрывающей аксиомой.

Вклад автора включает в себя новые теоретические гипотезы, алгоритмы и методы, разработанные для классификации теоретической сложности, проведение вычислительных экспериментов на реальных данных и подготовка

научных статей.

Автор диссертации является главным автором в [16] (Q1 WoS, Scopus) и [15] (Scopus). В [13, 12] (Q1 WoS, Scopus) автор диссертации является первым автором в списке, при этом все авторы внесли значительный вклад в работу.

2 Содержание работы

В этом разделе приводится обзор диссертации. Каждая глава содержит оригинальные научные результаты в форме краткого изложения основных идей, результатов и инноваций, представленных в ней.

Объем и структура работы. Диссертация состоит из вводной главы, заключительной главы и основной части, включающей содержание четырех статей. Общий объем диссертации составляет 132 страниц, включая приложения.

2.1 Проверка сложности по данным задачи поиска ответов на запросы, опосредованные онтологией: пример использования неоднородных задач удовлетворения ограничениям и системы Polyanna

В первой главе диссертации представлены наши исследования о том, как можно свести задачу поиска ответа на запрос с учетом онтологии с покрывающей аксиомой к задаче удовлетворения ограничениям с целью различения классов сложности задачи P и coNP (сложность по данным). Также в ней излагается наш опыт работы с программой Polyanna [9] для поиска полиморфизмов. Основная цель статьи – рассмотрение нашей задачи в терминах методологии решения проблемы удовлетворения ограничениям и, среди прочего, обобщение полученных ранее результатов о сложности по данным поиска ответов на онтологические запросы для нашего примера.

Для не-хорновских онтологических языков (допускающих дизъюнктивные аксиомы) ключевым шагом для понимания сложности по данным и переписываемости был получен результат в [4], где была установлена связь между онтологическими запросами и неоднородными задачами удовлетворения ограничениям с фиксированным шаблоном через MMSNP (Monotone Monadic Strict NP) из [8]. Данная методология использовалась, чтобы продемонстрировать, что решение первопорядковой переписываемости и datalog-переписываемости для онтологических запросов с онтологией в любой дескрипционной логике

между $\mathcal{ALC}$ и $\mathcal{SHIU}$ и атомарным запросом является NExpTime-полным. Дихотомия Федера-Варди для задачи удовлетворения ограничениям [6, 30] приводит к дихотомии P/coNPдля заданных нами онтологических запросов, которая разрешима в NExpTime.

Мы показываем, как именно онтологические запросы вида $(\mathsf{cov}^{\perp}_{\top}, \boldsymbol{q})$ с путевым конъюнктивным запросом $\boldsymbol{q}$ могут быть сведены к проблеме удовлетворения ограничениям (Рис. 3). В частности, нас интересуют неоднородные проблемы удовлетворения ограничениям. Пусть $\mathcal{B}$ - это фиксированная реляционная структура, которая в данном случае называется шаблоном. Каждый шаблон $\mathcal{B}$ порождает задачу решения проблемы удовлетворения ограничениям CSP($\mathcal{B}$), которая заключается в том, чтобы решить, учитывая экземпляр данных $\mathcal{A}$, существует ли гомоморфизм из $\mathcal{A}$ в $\mathcal{B}$, в случае чего мы пишем $\mathcal{A} \to \mathcal{B}$. Мы демонстрируем, следуя [4], как при онтологическом запросе $\boldsymbol{Q} = (\mathsf{cov}^{\perp}_{\top}, \boldsymbol{q})$ с путевым конъюнктивным запросом $\boldsymbol{q}$ можно построить шаблон $\mathcal{B}_{\boldsymbol{q}}$ такой, что для любого экземпляра данных $\mathcal{A}$ выполняется $\mathcal{A} \to \mathcal{B}_{\boldsymbol{q}}$ тогда и только тогда, когда $\mathsf{cov}^{\perp}_{\top}, \mathcal{A} \not\models \boldsymbol{q}$.

Чтобы проверить, находится ли проблема удовлетворения ограничениям CSP($\mathcal{B}_{\boldsymbol{q}}$) в P или coNP-трудная, можно использовать программу Polyanna [9], действующую в два этапа. Во-первых, она находит ядро шаблона $\mathcal{B}_{\boldsymbol{q}}$ с помощью свертки ('squashing'). Затем Polyanna решает вопрос о вычислимости или coNP-трудности проблемы удовлетворения ограничениям CSP($\mathcal{B}_{\boldsymbol{q}}$), проверяя, обладает ли шаблон ядра полиморфизмами определенных типов путем построения и решения соответствующих 'проблем-индикаторов' [19]. Для этого Polyanna использует различные методы декомпозиции для сокращения количества вычислений, когда задача-индикатор имеет симметрии. Задача-индикатор для полиморфизмов с арностью k и ядер с d вершинами, для сигнатуры Γ, имеет $k \cdot d^k$ переменных и $\Sigma_{R \in \Gamma} |R|^k$ ограничений. При шаблоне ядра размером d, Polyanna рассматривает полиморфизмы с арностью до $\max(3, d)$. На практике, для нашего случая, это означает, что программа может работать с ядрами размером до 4, но для некоторых ядер размером ≥ 5 память заканчивается.

Несмотря на идею, что можно перенести результаты дихотомии P/coNP для CSPs на нашу задачу, нам все еще не хватает простых и прозрачных, в

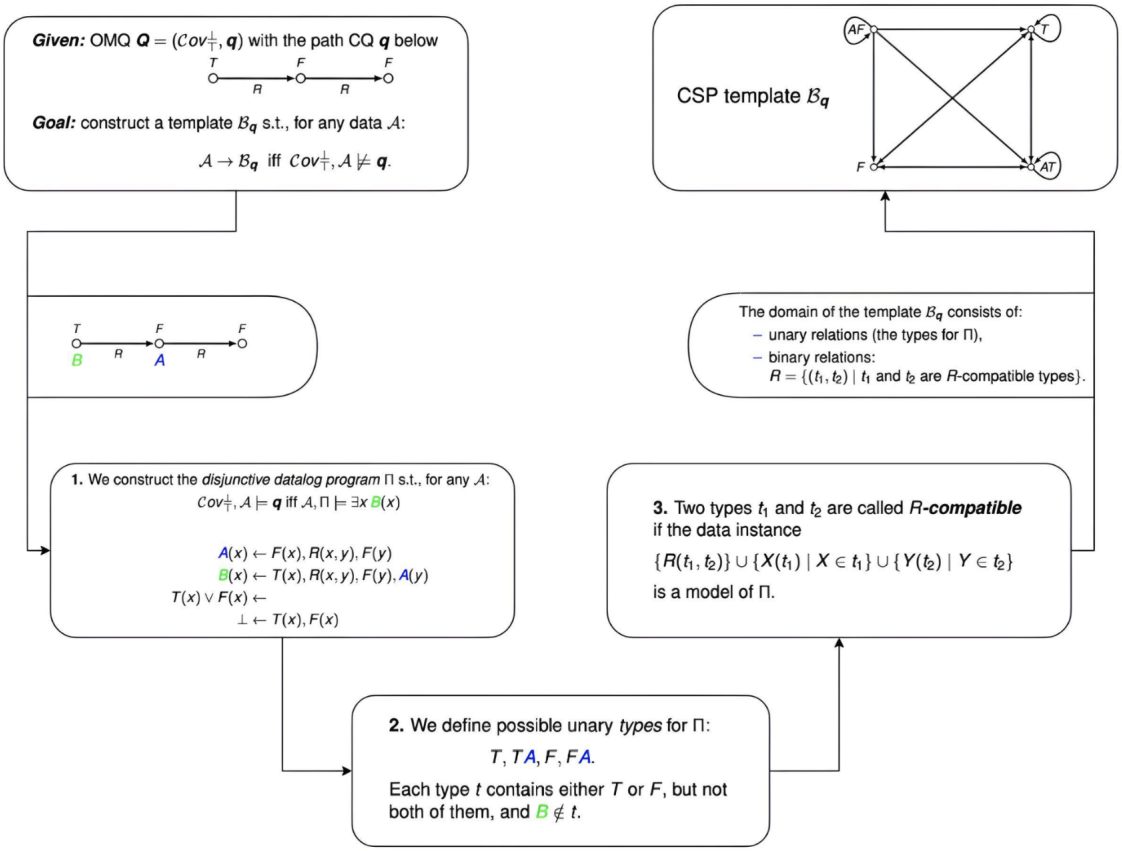


Рис. 3: Пример преобразования онтологического запроса в неоднородную проблему удовлетворения ограничениям

частности, синтаксических, условий, которые гарантировали бы ту или иную сложность по данным или тип переписываемости. Некоторые результаты в этом направлении были получены в [18, 20], и в рамках данной работы мы представили результаты по доказательству coNP-трудности для определенных запросов со специфической структурой. Однако тот факт, что до сих пор не разработана четкая классификация монадических sirups по их сложности по данным и тесная связь с проблемой удовлетворения ограничениям указывают на то, что эта проблема чрезвычайно трудна в целом.

Основные выводы работы заключаются в следующем. Во-первых, мы демонстрируем на простом примере, как определение вычислимости путевого онтологического запроса может быть сведено к проверке вычислимости

проблемы удовлетворения ограничениям. Во-вторых, мы рассматриваем, как программа Polyanna, разработанная для проверки вычислимости проблемы удовлетворения ограничениям, может быть использована в контексте нашего исследования для определения того, может ли задача поиска ответа на данный онтологический запрос с путевым конъюнктивным запросом из 4-х переменных быть в P или является coNP-трудной. В-третьих, что касается сложности по данным для задачи поиска ответов на онтологические запросы в рамках нашего исследования, мы приводим прямое доказательство coNP-трудности, используя сведение проблемы выполнимости булевых формул в 3-конъюнктивной нормальной форме (3SAT). Наконец, мы показываем, как Polyanna может быть использована для построения вычислимых монадических datalog-переписываний онтологических запросов с использованием алгоритма проверки согласованности дуг. Кроме того, в приложении мы обобщаем то, что нам известно о сложности по данным для проблемы поиска ответов на онтологические запросы в рамках нашей задачи на момент публикации данной работы.

Новизна работы заключается в двух аспектах. Во-первых, мы предлагаем исследования взаимосвязей между (d)d-sirups и задачей удовлетворения ограничениям с целью переноса результатов из одного формализма в другой. Во-вторых, новые теоретические результаты сложности по данным для задачи поиска ответов на онтологические запросы с покрывающей аксиомой [11, 14, 29], где мы классифицируем онтологические запросы по количеству вхождений одиночного атома F в их конъюнктивных запросах (случай одиночного T является симметричным).

Исследование было опубликовано в книге Springer “Description Logic, Theory Combination, and All That”, индексируемой в Scopus и WoS.

2.2 Тетрахотомия запросов, опосредованных онтологией с дизъюнктивной аксиомой

Вторая глава содержит самые важные теоретические результаты исследования. Мы получили теоретические результаты для онтологических запросов, соответствующих практически вычислимым классам от AC^0 до L/NL и P,

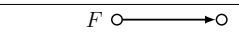
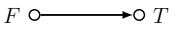
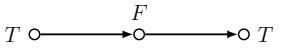
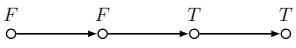
наиболее интересным с практической точки зрения. Запросы описываются и классифицируются с точки зрения их структуры (меток вершин, меток ребер, направлений ребер и т.д.). Наиболее гибкой характеристикой структуры запросов является разметка вершин, поскольку онтология работает именно с ними. Направление и метки ребер могут быть зафиксированы для упрощения задачи и уточнения процесса разметки вершин при помощи онтологии. Также представлен наиболее важный результат полной классификации широкого семейства запросов.

В этой главе мы представляем результаты подходов к получению полной классификации онтологических запросов вида $\boldsymbol{Q} = (\mathsf{cov}_A, \boldsymbol{q})$, где $\mathsf{cov}_A = \{\sqsubseteq F \sqcup T\}$, или $\boldsymbol{Q} = (\mathsf{cov}_A^{\perp}, \boldsymbol{q})$, где $\mathsf{cov}_A^{\perp} = \{\, A \sqsubseteq F \sqcup T,\ F \sqcap T \sqsubseteq \perp \,\}$, и $\boldsymbol{q}$ является Булевым конъюнктивным запросом. Мы заметили, что ответы на подобные онтологические запросы часто являются вычислимыми, причем соответствующие онтологические запросы могут быть переписаны в datalog-запросы к данным. В некоторых случаях мы даже можем получить переписываемость в линейный datalog, что гарантирует NL сложность для задачи поиска ответа на онтологический запрос. Мы привели несколько необходимых и достаточных условий для этих феноменов. Простые примеры, собранные в Таблице 1, демонстрируют, как небольшие изменения в структуре $\boldsymbol{q}$ могут кардинально повлиять на сложность онтологических запросов.

Мы рассматриваем только путевые конъюнктивные запросы $\boldsymbol{q}$, чей ориентированный граф имеет форму пути. Одиночные вхождения вершин с метками F и T будут просто называться F- и T-*вершинами*, соответственно. Обозначим первую (корневую) вершину в $\boldsymbol{q}$ через $b_{\boldsymbol{q}}$ и последнюю (листовую) вершину через $e_{\boldsymbol{q}}$. Для вершин x и y, мы пишем $x \prec y$, чтобы обозначить, что существует направленный путь из x в y в $\boldsymbol{q}$; как обычно, $x \preceq y$ означает $x \prec y$ или $x = y$. Для $x \preceq y$, множество $[x, y]$ включает в себя те атомы из $\boldsymbol{q}$, переменные которых находятся в интервале $\{z \mid x \preceq z \preceq y\}$ и $(x, y) = [x, y] \setminus \{T(x), F(x), T(y), F(y)\}$. Для $\boldsymbol{i} = (x, y)$, пусть $|\boldsymbol{i}|$ обозначает длину пути от x до y, и $|\boldsymbol{q}| = |(b_{\boldsymbol{q}}, e_{\boldsymbol{q}})|$.

Мы разделяем путевые конъюнктивные запросы на три непересекающихся класса: 0-CQ, 1-CQ и 2-CQ. Под 0-CQ понимается любой конъюнктивный запрос, не содержащий вхождений F (или T, соответственно). 1-CQ имеют

Таблица 1: Примеры графических представлений запросов $\boldsymbol{q}$ из онтологических запросов с онтологией cov_A и соответсвующих им классов сложности с точки зрения размера данных

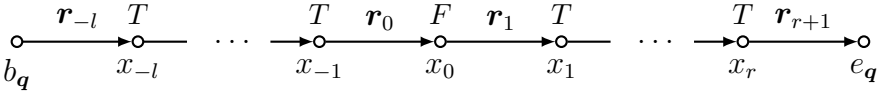
Сложность	Конъюнктивный запрос $\boldsymbol{q}$	Объяснение
AC^0		если в $\boldsymbol{q}$ есть только F, но нет T, в таком случае F можно проигнорировать
L		проверка неориентированной достижимости: ответом на $\boldsymbol{Q}$ является ‘да’
NL		проверка ориентированной достижимости: ответом на $\boldsymbol{Q}$ является ‘да’
P		вычисление монотонных Булевых схем
coNP		проверка выполнимости конъюнктивной нормальной формы

ровно одно вхождение F (или T, соответственно) и по крайней мере одно вхождение T (или F, соответственно). Если конъюнктивный запрос содержит не менее двух F-вершин и не менее двух T-вершин, то он называется 2-CQ.

Вершина конъюнктивного запроса $\boldsymbol{q}$ называется *сдвоенной*, если для некоторой переменной x выполняется $F(x), T(x) \in \boldsymbol{q}$. dd-sirups $(\mathsf{cov}_A^{\perp}, \boldsymbol{q})$ с $\boldsymbol{q}$, содержащими сдвоенные FT-метки, всегда переписываемы в запросы в виде формулы логики первого порядка.

Мы разделяем 1-CQs, не содержащие FT-меток, на *периодические* и *апериодические*, рассматривая только 1-CQs с одиночной F-вершиной и хотя бы одной T-вершиной (так как случай с одной T-вершиной и хотя бы одной F-вершиной является симметричным). При таком 1-CQ без сдвоенных вершин $\boldsymbol{q}$ и натуральных числах l, r, где $l + r \geq 1$, мы обозначаем $\boldsymbol{q} = \boldsymbol{q}_{lr}$, чтобы сказать, что $\boldsymbol{q}$ имеет l-много T-вершин $x_{-l} \prec \cdots \prec x_{-1}$, которые $\prec$-достигают его единственной F-вершины x_0, и r-много T-вершин $x_1 \prec \cdots \prec x_r$, которые $\prec$-достигают x_0. Для каждого i, $-l \leq i \leq r+1$, мы определяем множество $\boldsymbol{r}_i$ бинарных предикатов, взяв $\boldsymbol{r}_i = (x_{i-1}, x_i)$, где $x_{-l-1} = b_{\boldsymbol{q}}$ и $x_{r+1} = e_{\boldsymbol{q}}$. Заметим, что $\boldsymbol{r}_i \neq \emptyset$ для $-l < i < r+1$, но $\boldsymbol{r}_{-l} = \emptyset$, если $b_{\boldsymbol{q}} = x_{-l}$, и $\boldsymbol{r}_{r+1} = \emptyset$, если $x_r = e_{\boldsymbol{q}}$. Аналогично, для каждого i с $0 \leq i \leq l$, пусть $\boldsymbol{q}_i^- = (b_{\boldsymbol{q}}, x_{-i})$, где $\boldsymbol{q}_l^- = \emptyset$, если

$x_{-l} = b_{\boldsymbol{q}}$. Пусть $\boldsymbol{r}_i = (x_{i-1}, x_i)$, где $x_{-l-1} = b_{\boldsymbol{q}}$ и $x_{r+1} = e_{\boldsymbol{q}}$ (может случиться, что $x_{-l-1} = x_{-l}$ и $x_{r+1} = x_r$).

$\boldsymbol{q} = \boldsymbol{q}_{lr}$ 

Каждый $\boldsymbol{r}_i$ определяет конечную последовательность $\langle \boldsymbol{r}_i \rangle$ символов бинарных предикатов. Мы называем $\boldsymbol{q}$ *право-периодичным*, если $\boldsymbol{q} = \boldsymbol{q}_{0r}$ и либо $r = 1$, либо $\langle \boldsymbol{r}_i \rangle = \langle \boldsymbol{r}_1 \rangle$ для всех $i = 1, \dots, r$ и $\langle \boldsymbol{r}_{r+1} \rangle = \langle \boldsymbol{r}_1 \rangle^* \lambda$ для некоторого (возможно, пустого) префикса λ из $\langle \boldsymbol{r}_1 \rangle$. Взяв зеркальное отражение этого определения, мы получаем понятие *лево-периодичный* 1-CQ, в случае которого $\boldsymbol{q} = \boldsymbol{q}_{l\,0}$ и либо $l = 1$, либо $\langle \boldsymbol{r}_{-i} \rangle = \langle \boldsymbol{r}_0 \rangle$ для всех $i = 1, \dots, l-1$ и $\langle \boldsymbol{r}_{-l} \rangle = \lambda \langle \boldsymbol{r}_0 \rangle^*$ для некоторого (возможно, пустого) суффикса λ из $\langle \boldsymbol{r}_0 \rangle$. 1-CQ без сдвоенных вершин $\boldsymbol{q}$ называется *периодическим*, если он право- или лево-периодический, и *апериодическим* в противном случае.

Theorem ($\mathbf{AC^0}$ / NL / P / coNP-тетрахотомия (d-sirups) dd-sirups для (без сдвоенных вершин) путевых конъюнктивных запросов). *Пусть $\boldsymbol{Q}$ – любой d-sirup с путевым конъюнктивным запросом без сдвоенных вершин $\boldsymbol{q}$ или любой dd-sirup с путевым конъюнктивным запросом $\boldsymbol{q}$. Тогда имеет место следующая тетрахотомия, (где три 'if' можно заменить на 'iff' при условии, что* NL $\neq$ P $\neq$ coNP):

($\mathbf{AC^0}$) *$\boldsymbol{Q}$ является первопорядкого переписываемым и задача поиска ответа для него в* AC^0*, если $\boldsymbol{q}$ – это 0-CQ или иначе содержит сдвоенную вершину с FT-меткой*;,

(NL) *$\boldsymbol{Q}$ является переписываемым в линейную datalog-программу и задача поиска ответа на него* NL*-полная, если $\boldsymbol{q}$ – это периодический 1-CQ*;

(P) *$\boldsymbol{Q}$ является datalog-переписываемым и задача поиска ответа на него* P*-полной, если $\boldsymbol{q}$ – это апериодический 1-CQ*;

(coNP) *задача поиска ответа на $\boldsymbol{Q}$ является* coNP*-полной, если $\boldsymbol{q}$ – это 2-CQ.*

Основной результат нашего исследования – полная синтаксическая классификация dd-sirups $(\mathsf{cov}_A^{\perp}, \boldsymbol{q})$ с путевыми конъюнктивными запросами $\boldsymbol{q}$ относи-

тельно сложности по данным и типу переписываемости. Хотя AC^0/NL-часть этой AC^0/NL/P/coNP-тетрахотомии следует из наших предыдущих результатов [11, 14, 29], доказательство P- и особенно coNP-трудности оказывается сложным и потребовало разработки новых методов.

Было опубликовано две работы, посвященных описанным выше результатам. Первая была опубликована в материалах 17-й международной конференции Principles of Knowledge Representation and Reasoning (Core A*), вторая - в журнале Artificial Intelligence, индексируемом Q1 (2022) в Scopus и Web of Science.

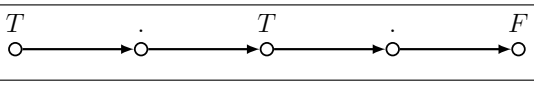
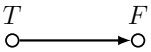
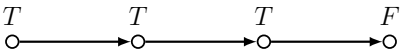
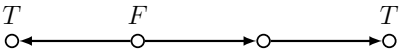
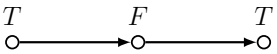
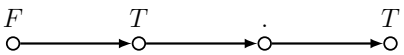
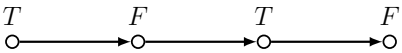
2.3 Сравнительный анализ метода вывода логических рассуждений и графовых нейронных сетей для поиска ответов на запросы, опосредованные онтологией с дизъюнктивной аксиомой

Третья глава посвящена онтологическим запросам, принадлежащим к интересным с практической точки зрения классам сложности L/NL и P, для которых существует способ переписать исходный запрос в datalog-программу с учетом информации из дизъюнктивной онтологии.

В Таблице 2 приведены классы сложности и соответствующие им графические представления конъюнктивных запросов, для которых мы проводим эксперименты по оценке производительности datalog-ризонеров в данном исследовании.

Переписав онтологический запрос в datalog-программу, мы стремимся понять эффективность ризонеров при поиске ответов на онтологические запросы для различных наборов данных и шаблонов конъюнктивных запросов. Наш выбор остановился на двух хорошо известных и простых в использовании datalog ризонерах DLV (DataLog с $\vee$- дизъюнкцией)[24], основанного на дедуктивном выводе рассуждений, используя дизъюнктивное логическое программирование и Clingo [10], основанного на Answer Set Programming (ASP). Мы сосредоточились на следующих вопросах: как размер неразмеченных данных (без меток F и T) влияет на производительность ризонеров, какие гра-

Таблица 2: Графические представления запросов как направленные размеченные графы

Запрос	Граф запроса	Сложность
$\boldsymbol{q}_0$		NL
$\boldsymbol{q}_1$		NL
$\boldsymbol{q}_2$		NL
$\boldsymbol{q}_3$		NL
$\boldsymbol{q}_4$		P
$\boldsymbol{q}_5$		P
$\boldsymbol{q}_6$		coNP

фовые нейронные сети лучше всего справляются с классификацией вершин, и как подход поиска ответа на запрос к данным, обогащенным метками, полученными от графовых нейросетевых моделей, сравнивается с использованием логических ризонеров совместно с истинной разметкой в данных.

Мы проанализировали производительность datalog-ризонеров в зависимости от запущенных практически вычислимых запросов, учитывая непосредственные их переписывания в datalog-программы, и от доли данных, маскируемых для оценки работы онтологии.

Для логических ризонеров становится понятно, что время работы datalog систем на неразмеченных данных различного размера напрямую зависит от структуры конъюнктивного запроса. Clingo работает быстрее, чем DLV на больших графах благодаря эффективной реализации ASP. Кроме того, важно отметить, что если объем неразмеченных данных слишком велик, то онтологический подход может выдавать неверный ответ из-за отсутствия достаточного числа известных меток.

Для получения ответов на основе графовых нейронных сетей было выяснено, что уровень предсказания для классификации вершин, особенно для графов с отрицательной ассортативностью, недостаточен, чтобы заменить логи-

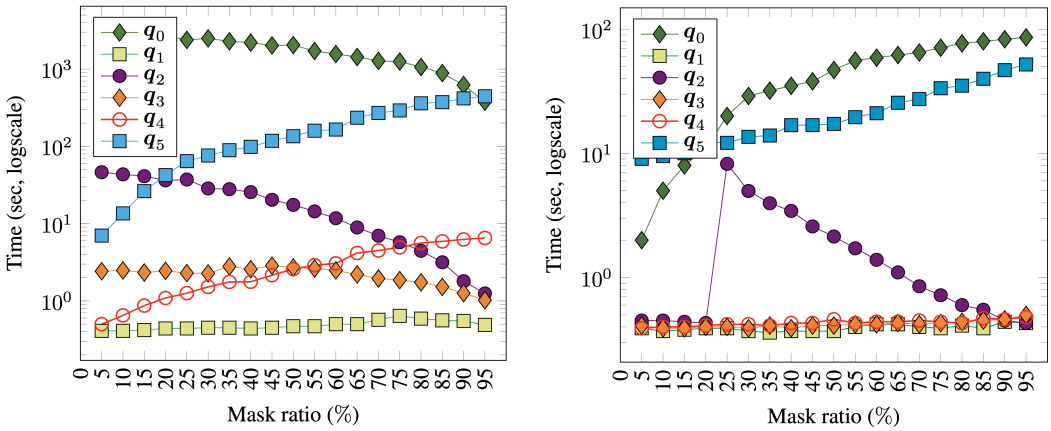


Рис. 4: Сравнение времени работы систем DLV (слева) and Clingo (справа) для различных запросов на датасете Polblogs.

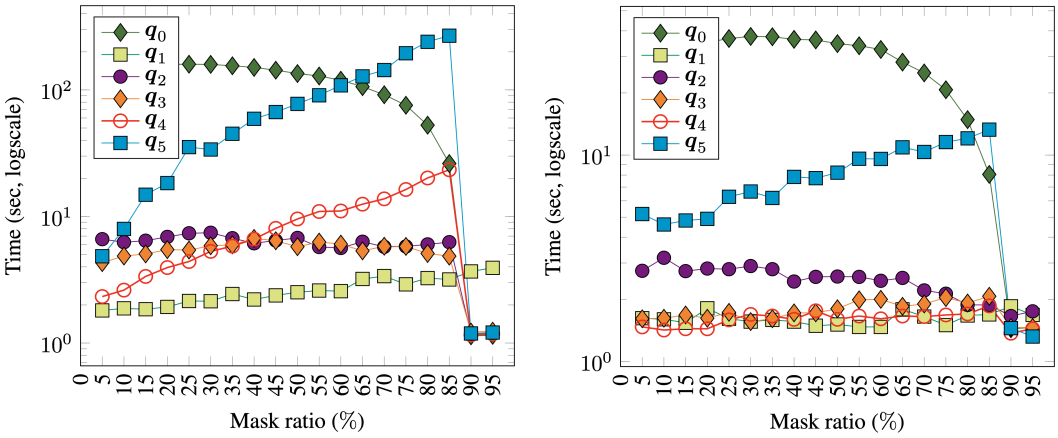


Рис. 5: Сравнение времени работы систем DLV (слева) and Clingo (справа) для различных запросов на датасетеDeezer.

ческий вывод рассуждения. Таким образом, на больших графах даже графовые нейросетевые модели с высокой точностью такие, как Graph Convolutional Network (GCN) [22] и Graph Attention Network (GAT) [28], терпят неудачу, поскольку получаемая ими разметка вершин покрывает большинство простых структур для конъюнктивных запросов, которые гарантированно находятся напрямую в данных.

Наши результаты подчеркивают важность совместного анализа графовых структур данных и запроса, а также количества и расположения неразмеченных данных для выбора между точным и приблизительным методами поиска

ответа.

Новизна нашей работы заключается в том, что мы указали на ограничения использования подхода классификации вершин для решения задачи поиска ответа на онтологические запросы с покрывающей аксиомой, а также доказали преимущества ризонеров и datalog переписываний онтологических запросов для обеспечения согласованности данных.

Статья по результатам исследования опубликована в журнале IEEE Access. Он входит в рейтинг Q1 (2022) в Web of Science и Scopus.

3 Заключение

Диссертация основана на четырех опубликованных статьях: двух журнальных статьях уровня Q1 [12, 16], статьи в сборнике трудов конференции уровня Core A* [13] и статьи из книги Springer [15]. В статьях [15, 13, 12] представлены новые подходы в области OBDA для решения конкретного случая задачи поиска ответов на Булевы конъюнктивные древовидные запросы, опосредованные онтологией с покрывающей аксиомой. В статье [16] предлагается новое понимание важности OBDA по сравнению с методами машинного обучения, устраняющими онтологии путем разметки данных.

В совокупности, эти работы позволяют интегрировать покрывающую аксиому в OBDA, обозначая теоретические границы практических случаев нашей задачи для OBDA и конструктивные переписывания онтологических запросов с покрывающей аксиомой в datalog-программы. Исследование позволяет соединить в едином подходе семантические технологии, теоретическую информатику и управление базами данных.

Основными результатами данной диссертации, выносимыми на защиту, являются:

1. представленная методология поиска практически вычислимых случаев путевых онтологических запросов с покрывающей аксиомой, в том числе сведение нашей задачи к CSP, и разработанные datalog-переписывания для вычислимых случаев со сложностью AC^0, L/NL и P.

2. определение полной синтаксической классификации (d)d-sirups $(\mathsf{cov}^{\perp}_{A}, \boldsymbol{q})$ с путевыми конъюнктивными запросами $\boldsymbol{q}$ в рамках сложности по данным и типу переписываемости среди таких классов сложности, как AC^0 / NL / P / coNP.

3. анализ эффективности онтологических запрос с покрывающей аксиомой путем сравнения OBDA подхода с методами машинного обучения, применяемых для разметки данных и поиска ответа на запрос напрямую по размеченным данным.

Направления будущих исследований

Наши результаты на синтаксические ограничения запросов порождают направления для будущих исследований. Представляется интересным распространить основную теорему о полной классификации на более общие семейства онтологических запросов, наподобие (i) d-sirups с путевыми конъюнктивными запросами, которые могут содержать сдвоенные FT-метки, (ii) неориентированные путевые, (iii) ориентированные древовидные и (iv) неориентированные древовидные dd- и d-sirups. После этого было бы важно урегулировать высокую сложность разрешения первопорядковой переписываемости запросов и других типов переписываемости для произвольных (d)d-sirups.

Следующим шагом может быть рассмотрение сложности задачи разрешения переписываемости для более сложных онтологий, таких как Schema.org со множественными дизъюнкциями, *DL-Lite*$_{krom}$ и *DL-Lite*$_{bool}$ [2] с ограниченным использованием экзистенциального квантора в правой части импликаций.

Также интересно проанализировать размер переписываний в виде формул логики первого порядка для онтологических запросов с дизъюнктивными аксиомами, начиная с d- и dd-sirups.

Еще одним направлением является рассмотрение проблем сложности по данным и переписываемости для (d)d-sirups с означиванием одной или нескольких переменных при ответе на запрос.

Кроме того, нахождение более низких границ сложности по данным для классов онтологических запросов, для которых алгоритмы переписывания являются полными, может иметь большое значение для подведения итогов для случая дизъюнктивных онтологий и их роли в OBDA.

Наконец, нам нужно предложить новые подходы, чтобы избежать эффекта, при котором графовые нейронные сети, насыщающие данные метками, могут дать очень точные результаты для разметки данных, но не дают правильных ответов на запрос с учетом онтологии.

Список литературы

[1] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[2] A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The DL-Lite family and relations. *Journal of Artificial Intelligence Research (JAIR)*, 36:1–69, 2009.

[3] Michael Benedikt, Balder ten Cate, Thomas Colcombet, and Michael Vanden Boom. The complexity of boundedness for guarded logics. In *30th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2015, Kyoto, Japan, July 6-10, 2015*, pages 293–304. IEEE Computer Society, 2015.

[4] M. Bienvenu, B. ten Cate, C. Lutz, and F. Wolter. Ontology-based data access: A study through disjunctive datalog, CSP, and MMSNP. *ACM Transactions on Database Systems*, 39(4):33:1–44, 2014.

[5] Meghyn Bienvenu, Balder ten Cate, Carsten Lutz, and Frank Wolter. Ontology-based data access: a study through disjunctive datalog, CSP, and MMSNP. In *Proc. of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2013)*, pages 213–224. ACM, 2013.

[6] Andrei A. Bulatov. A dichotomy theorem for nonuniform csps. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 319–330. IEEE Computer Society, 2017.

[7] Stavros S. Cosmadakis, Haim Gaifman, Paris C. Kanellakis, and Moshe Y. Vardi. Decidable optimization problems for database logic programs (preliminary report). In *STOC*, pages 477–490, 1988.

[8] Tomás Feder and Moshe Y. Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory. *SIAM J. Comput.*, 28(1):57–104, 1998.

[9] Richard Gault and Peter Jeavons. Implementing a test for tractability. *Constraints*, 9:139–160, 2004.

[10] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Multi-shot asp solving with clingo. *Theory and Practice of Logic Programming*, 19(1):27–82, 2019.

[11] O. Gerasimova, S. Kikot, V. Podolskii, and M. Zakharyaschev. On the data complexity of ontology-mediated queries with a covering axiom. In *Proceedings of the 30th International Workshop on Description Logics*, 2017.

[12] Olga Gerasimova, Stanislav Kikot, Agi Kurucz, Vladimir Podolskii, and Michael Zakharyaschev. A tetrachotomy of ontology-mediated queries with a covering axiom. *Artificial Intelligence*, 309:103738, 2022.

[13] Olga Gerasimova, Stanislav Kikot, Agi Kurucz, Vladimir Podolskii, Michael Zakharyaschev, et al. A data complexity and rewritability tetrachotomy of ontology-mediated queries with a covering axiom. In *Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning*. International Joint Conferences on Artificial Intelligence Organization, 2020.

[14] Olga Gerasimova, Stanislav Kikot, Vladimir V. Podolskii, and Michael Zakharyaschev. More on the data complexity of answering ontology-mediated queries with a covering axiom. In *Knowledge Engineering and Semantic Web - 8th International Conference, KESW 2017, Szczecin, Poland, November 8-10, 2017, Proceedings*, pages 143–158, 2017.

[15] Olga Gerasimova, Stanislav Kikot, and Michael Zakharyaschev. Checking the data complexity of ontology-mediated queries: A case study with non-uniform CSPs and Polyanna. In Carsten Lutz, Uli Sattler, Cesare Tinelli, Anni-Yasmin Turhan, and Frank Wolter, editors, *Description Logic, Theory Combination, and All That*, volume 11560 of *Lecture Notes in Computer Science*, pages 329–351. Springer, 2019.

[16] Olga Gerasimova, Nikita Severin, and Ilya Makarov. Comparative analysis of logic reasoning and graph neural networks for ontology mediated query answering with a covering axiom. *IEEE Access*, pages 1–13, 2023.

[17] Georg Gottlob and Christos H. Papadimitriou. On the complexity of single-rule datalog queries. *Inf. Comput.*, 183(1):104–122, 2003.

[18] André Hernich, Carsten Lutz, Ana Ozaki, and Frank Wolter. Schema.org as a description logic. In Diego Calvanese and Boris Konev, editors, *Proceedings of the 28th International Workshop on Description Logics, Athens,Greece, June 7-10, 2015.*, volume 1350 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[19] Peter Jeavons, David Cohen, and Marc Gyssens. A test for tractability. *CP*, 1118:267–281, 1996.

[20] Mark Kaminski, Yavor Nenov, and Bernardo Cuenca Grau. Datalog rewritability of disjunctive datalog programs and non-Horn ontologies. *Artif. Intell.*, 236:90–118, 2016.

[21] Paris C. Kanellakis. Elements of relational database theory. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, pages 1073–1156. Elsevier and MIT Press, 1990.

[22] N. Kipf, T. and M. Welling. Semi-supervised classification with graph convolutional networks, 2017.

[23] Roman Kontchakov and Michael Zakharyaschev. An introduction to description logics and query rewriting. In *Reasoning Web International Summer School*, pages 195–244. Springer, 2014.

[24] Nicola Leone, Gerald Pfeifer, Wolfgang Faber, Francesco Calimeri, Tina Dell'Armi, Thomas Eiter, Georg Gottlob, Giovambattista Ianni, Giuseppe Ielpa, Christoph Koch, et al. The dlv system. In *Logics in Artificial Intelligence: 8th European Conference, JELIA 2002 Cosenza, Italy, September 23–26, 2002 Proceedings 8*, pages 537–540. Springer, 2002.

[25] Carsten Lutz and Leif Sabellek. Ontology-mediated querying with the description logic EL: trichotomy and linear datalog rewritability. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1181–1187. ijcai.org, 2017.

[26] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *Journal on Data Semantics*, X:133–173, 2008.

[27] Moshe Y. Vardi. Decidability and undecidability results for boundedness of linear recursive queries. In Chris Edmondson-Yurkanan and Mihalis Yannakakis, editors, *Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, March 21-23, 1988, Austin, Texas, USA*, pages 341–351. ACM, 1988.

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks, 2017.

[29] Michael Zakharyaschev, Stanislav Kikot, and Olga Gerasimova. Towards a data complexity classification of ontology-mediated queries with covering. In Magdalena Ortiz and Thomas Schneider, editors, *Proceedings of the 31st International Workshop on Description Logics co-located with 16th International Conference on Principles of Knowledge Representation and Reasoning (KR 2018), Tempe, Arizona, US, October 27th - to - 29th, 2018.*, volume 2211 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[30] Dmitriy Zhuk. A proof of CSP dichotomy conjecture. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 331–342. IEEE Computer Society, 2017.