

Разработка алгоритмов анализа граф-структурированных данных, основанных на теории формальных языков

Семён Григорьев

29 октября 2019 г.

1 Сведения о проекте

1.1 Название проекта

ru

Разработка алгоритмов анализа граф-структурированных данных, основанных на теории формальных языков

или

Теория формальных языков и алгоритмы синтаксического анализа для анализа граф-структурированных данных

или

Теория и практика анализа граф-структурированных данных с использованием методов теории формальных языков.

en

- 1.2 Приоритетное направление развития науки, технологий и техники в Российской Федерации, критическая технология**
- 1.3 Направление из Стратегии научно-технологического развития Российской Федерации (утверждена Указом Президента Российской Федерации от 1 декабря 2016 г. №642 “О Стратегии научно-технологического развития Российской Федерации”) (при наличии)**
- 1.4 Ключевые слова (приводится не более 15 терминов)**

ru

Теория графов, теория формальных языков, поиск путей, графовые базы данных, формальные грамматики, синтаксический анализ, оптимизации алгоритмов, параллельные алгоритмы, смешанные вычисления, специализация.

en

1.5 Аннотация проекта

ru

Эффективная обработка больших объёмов данных — актуальная прикладная область, требующая качественных теоретических результатов для решения возникающих прикладных задач. Одной из активно изучаемых в последнее время моделей для представления обрабатываемых данных является граф. На практике такая модель применяется при работе с различными сетями (социальные сети, транспортные сети), при анализе и верификации программных и аппаратных комплексов (графы вызовов, переходов и т.д.), а в общем случае является основой для графовых баз данных. Иными словами, обработка граф-структурированных данных является активно развивающейся областью.

Одна из задач при обработке данных — поиск и анализ связей между сущностями (или же установление факта отсутствия специфических связей). В случае граф-структурированных данных данная задача формулируется в терминах поиска путей между вершинами или проверки их отсутствия. При этом содержательные задачи приводят к появлению дополнительных, не тривиальных, ограничений на пути. В качестве классического примера дополнительных ограничений можно рассмотреть поиск простых путей и поиск кратчайших путей в графе.

Одним из способов задать ограничение на путь в размеченном графе (то есть в графе, рёбра которого несут некоторую нагрузку в виде метки или веса) использует формальные языки. В данном случае рассматриваются слова, полученные конкатенацией меток рёбер пути, и задаётся язык, которому должны принадлежать такие слова. Иными словами, возникает следующая задача: найти пути в графе, такие, что слова, задаваемые ими, принадлежат

заданному языку. При этом, возможны различные вариации постановки задачи (характерные для многих задач поиска путей): поиск пути между двумя заданными вершинами, поиск всех путей в графе, удовлетворяющих заданному ограничению, проверка достижимости (а не поиск непосредственно пути) и т.д. В зависимости от этого необходимо применять различные алгоритмы для достижения лучшей эффективности.

Вместе с тем, так как ограничения формулируются в терминах языков, естественным является привлечение результатов теории формальных языков. С одной стороны, возникают фундаментальные вопросы о разрешимости задачи: при использовании каких классов языков в качестве ограничений задача поиска путей разрешима. С другой стороны, оказывается возможным использовать алгоритмы синтаксического анализа для решения задачи, однако алгоритмы требуют модификации, а исследование их теоретических свойств, например, временной и пространственной сложности, оказывается нетривиальной задачей. Важно, что ответы на эти вопросы связаны не только со свойствами используемых языков, но и со свойствами обрабатываемых графов, что приводит к тесному соприкосновению двух областей науки: теории графов и теории формальных языков. Несмотря на то, что задача поиска путей с ограничениями в терминах формальных языков изучается с начала 1990-х (Томас Репс и Михалис Яннакакис), многие вопросы остаются открытыми. Например, до сих пор не решён вопрос о существовании субкубического алгоритма для поиска путей с контекстно-свободными ограничениями. А конкретные алгоритмы решения задач стали разрабатываться и изучаться совсем недавно, когда возрос интерес к графовым базам данных.

С прикладной же точки зрения, важно получение эффективных с вычислительной точки зрения алгоритмов для обработки практически важных сценариев. Так как графы, возникающие в прикладных задачах, имеют большой размер в терминах количества вершин и рёбер, то естественным путём является разработка параллельных и распределённых алгоритмов их обработки, в том числе алгоритмов, использующих массово-параллельные архитектуры, такие как GPGPU. Данное направление активно развивается в области обработки графов, однако слабо проработано в контексте обсуждаемой задачи.

Более того, если рассматривать задачу поиска путей в контексте графовых баз данных, то необходимо предоставить удобные соседства описания запросов к таким базам, позволяющие формулировать ограничения в терминах формальных языков. Одним из классических способов естественно задавать такие ограничения в прикладных языках программирования является использование парсер комбинаторов — специальных функций, позволяющих строить сложные парсеры из более простых, обеспечивая при это "бесшовную" интеграцию с основным языком программирования (нет отдельной процедуры встраивания специализированного языка в язык общего назначения), высокий уровень абстракции за счёт возможности использовать функции высших порядков, надёжность и безопасность за счёт полной интеграции с системой вывода типов используемого языка. Такой подход хорошо зарекомендовал себя при анализе языков программирования, однако его применимость для анализа графов исследована слабо.

Также, в контексте выполнения запросов к графовым базам данных, необходимо разработать методы оптимизации как самих запросов, так и процедур их исполнения. Здесь перспективным подходом является применение смешанных вычислений, в частности, специализации. Хотя в области реляционных баз данных такой подход показал себя состоятельным

(например, работы Евгения Шарыгина и соавторов), в контексте графовых баз данных данные техники не применялись.

Проект посвящён разработке и реализации алгоритмов для поиска путей с ограничениями в терминах формальных языков, а также вопросам создания средств задания таких ограничений и методам оптимизации соответствующих запросов к графовым базам данных. При разработке алгоритмов будут использоваться методы теории формальных языков и теории графов для поиска классов графов и языков, для которых, во-первых, принципиально возможно построение алгоритмов решения задач поиска путей с ограничениями в терминах формальных языков, во-вторых, возможно построение асимптотически эффективных алгоритмов. Для разработки эффективных с практической точки зрения алгоритмов будут использоваться методы построения параллельных алгоритмов, в том числе, алгоритмов для массово-параллельных архитектур. Исследование способов задания ограничений потребует использования знаний из области разработки языков программирования. При разработке методов оптимизации запросов будут использоваться техники смешанных вычислений.

Коллектив исполнителей включает специалистов по теории формальных языков, теории графов, построению компиляторов, методам оптимизации программ, и разработке языков программирования. Это позволит организовать плодотворное сотрудничество и обеспечить комплексный подход к решению задач, а также привлечь к изучению талантливых студентов к соответствующим областям науки и работе над проектом.

en

1.6 Ожидаемые результаты и их значимость

ru

Проект направлен на изучение задачи о поиске путей с ограничениями в терминах формальных языков с целью получения эффективного с прикладной точки зрения решения для неё. Ожидаются как теоретические результаты на стыке теории формальных языков и теории графов и в области построения параллельных алгоритмов, так и результаты в области разработки языков и методов оптимизации программного обеспечения.

В частности, ставится задача построить более детальную классификацию задач и поиска путей с контекстно-свободными ограничениями как с точки зрения подклассов языков, так и с точки зрения типов графов. Основная цель здесь — ответить на вопрос о существовании субкубического алгоритма для задачи в общем случае. Данный вопрос открыт уже длительное время, так что полностью ответить на него вряд ли удастся, но ценными будут и частичные ответы в терминах подклассов задачи, для которых такой алгоритм точно существует.

В области построения параллельных алгоритмов планируется получение новых алгоритмов для решения задачи поиска путей с контекстно-свободными ограничениями для массово-параллельных и распределённых систем. Будут изучены теоретические свойства предложенных алгоритмов, в частности, получены асимптотические оценки временной и пространствен-

ной сложности. Так же будет исследованы возможности расширения построенных алгоритмов для других классов языков.

При разработке прикладных способов и средств задания ограничений в терминах языков будут исследованы подходы на основе парсер-комбинаторов. Планируется, что будут получены границы применимости такого подхода, а также изучены его слабые и сильные стороны в контексте прикладных задач, такие как типобезопасность, возможность вычисления дополнительных семантических функций. Несмотря на то, что применение парсер-комбинаторов для анализа языков программирования изучено достаточно хорошо, обобщение этого подхода на графы нетривиально и ожидаются новые результаты. Парсер-комбинаторы предоставляют не только механизм для решения задачи поиска путей с ограничениями, но и формализм для описания запросов. Использование такого формализма упростит использование технологии конечными пользователями, а также предоставит более прозрачную интеграцию в логику разрабатываемой программы. Планируется разработка удобного формализма спецификации запросов. Также парсер-комбинаторы позволяют вычисление пользовательской семантики, при помощи чего можно выразить фильтрацию, агрегацию, счетчики и прочие виды обработки результата запроса. В рамках работы будет изучено, для каких классов входных графов можно точно вычислить пользовательскую семантику. Некоторые языки, не являющиеся контекстно-свободными, можно анализировать при помощи парсер-комбинаторов. Будет изучен вопрос использования более, чем контекстно-свободных ограничений для поиска путей в графах.

< Катя! >

В области оптимизации запросов и процедур их исполнения планируется разработать решение для специализации алгоритмов выполнения запросов к графовым базам данных во время выполнения. Вероятно, при этом будет необходимо разработать новые алгоритмы специализации < Даня! >

en

1.7 В состав научного коллектива будут входить

- исполнителей проекта (включая руководителя)
- в том числе !!! исполнителей в возрасте до 39 лет включительно,
- из них: !!! очных аспирантов, адъюнктов, интернов, ординаторов, студентов.

1.8 Планируемый состав научного коллектива с указанием фамилий, имен, отчеств (при наличии) членов коллектива, их возраста на момент подачи заявки, ученых степеней, должностей и основных мест работы, формы отношений с организацией (трудовой договор, гражданско-правовой договор) в период реализации проекта.

Соответствие профессионального уровня членов научного коллектива задачам проекта **ru** Руководитель, Семён Вячеславович Григорьев является доцентом кафедры информатики СПбГУ и кандидатом физико-математических наук. Опыт руководства исследовательскими работами и преподавания составляет 6 лет. За это время под его руководством защищено 7 магистерских диссертаций, 12 выпускных квалификационных работ бакалавра, 2 дипломных работы специалиста, больше 10 курсовых работ. В настоящее время под его руководством работают два аспиранта. За время преподавательской деятельности занимался подготовкой и чтением курсов по теории графов, теории формальных языков, алгоритмам и структурам данных. Имеет опыт руководства грантами (РФФИ 19-37-90101; программа УМНИК, 162ГУ1/2013 и 5609ГУ1/2014) исследовательскими группами и отдельными исследовательскими работами. Также имеет опыт исполнения грантов (РФФИ 15-01-05431, РФФИ 18-01-00380, РНФ 18-11-00100).

Екатерина Андреевна Вербицкая окончила аспирантуру математико-механического факультета СПбГУ по направлению информатика, преподает на кафедре МО ЭВМ СПбГЭТУ “ЛЭТИ”. Опыт руководства исследовательскими работами и преподавания составляет 4 года. За это время под ее руководством было защищено 2 выпускных квалификационных работы бакалавра. В настоящее время под ее руководством работают 2 магистранта. За время преподавательской деятельности занималась подготовкой и чтением курсов по теории формальных языков и разработке компиляторов. Имеет опыт исполнения грантов (РФФИ 18-01-00380). Область научных интересов включает анализ встроенных языков, синтаксический анализ при помощи парсер-комбинаторов, функциональное программирование, суперкомпиляцию и частичную дедукцию для логических языков.

Даниил Андреевич Березун — суперкомпиляция, кфмн

Рустам Азимов — графы, поиск путей в графах, формальные языки, параллельные алгоритмы.

Екатерина Николаевна Шеметова закончила магистратуру университета ИТМО по специальности “Разработка программного обеспечения”. Защитила магистерскую диссертацию на тему “Задача поиска путей с контекстно-свободными ограничениями”. Имеет публикацию по теме проекта (“Задача поиска путей в ациклических графах с ограничениями в терминах булевых грамматик”). Является аспирантом Санкт-Петербургского Академического университета по направлению “Информатика”. Имеет опыт исполнения грантов (РНФ 18-11-00100). Область научных интересов: теория сложности алгоритмов, теория формальных языков и её приложения, статический анализ кода.

Сусанина Юлия Алексеевна является магистрантом математико-механического факультета СПбГУ по направлению “Программная инженерия”. Ее области научных интересов включают

теорию формальных языков, алгоритмы синтаксического анализа и их применения. Полученные за время обучения в бакалавриате результаты, связанные с исследованием матричных алгоритмов синтаксического анализа, были приняты к публикации в журнал “Труды ИСП РАН” и представлены на международной конференции по биоинформатике CIBB 2019.

Мишин Никита Матвеевич является студентом математико-механического факультета СПбГУ по направлению “Программная инженерия”. Его область научных интересов включает распределенные и параллельные вычисления, формальные языки, программирование на ГПУ (GPGPU) и функциональное программирование. Имеет публикацию по теме проекта (Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication), которая представлена на международной конференции GRADES-NDA 2019 и опубликована в соответствующих материалах конференции. Участник нескольких летних школ, в частности, летней школы Ланит-Терком, проект RuCuHmmer, направленный на частичный перенос вычислений в программном пакете HMMER (анализа биологических последовательностей), связанных с обработкой большого объема данных, на GPU с целью увеличения эффективности алгоритмов.

en

1.9 Планируемый объем финансирования проекта Фондом по годам (указывается в тыс. рублей)

2020 г. - тыс. рублей, 2021 г. - введите планируемый объем финансирования в 2021 г. тыс. рублей, 2022 г. - введите планируемый объем финансирования в 2022 г. тыс. рублей.

1.10 Научный коллектив по результатам проекта в ходе его реализации предполагает опубликовать в рецензируемых российских и зарубежных научных изданиях не менее

!!! публикаций

из них !!!введите число:!!! в изданиях, индексируемых в базах данных «Сеть науки» (Web of Science Core Collection) или «Скопус» (Scopus).

Информация о научных изданиях, в которых планируется опубликовать результаты проекта, в том числе следует указать в каких базах индексируются данные издания - «Сеть науки» (Web of Science Core Collection), «Скопус» (Scopus), РИНЦ, иные базы, а также указать тип публикации - статья, обзор, тезисы, монография, иной тип

Иные способы обнародования результатов выполнения проекта

1.11 Число публикаций членов научного коллектива, опубликованных в период с 1 января 2015 года до даты подачи заявки

!!!введите число:!!!, из них !!!введите число:!!! – опубликованы в изданиях, индексируемых в Web of Science Core Collection или в Scopus.

1.12 Планируемое участие научного коллектива в международных коллаборациях (проектах) (при наличии)

Руководитель проекта подтверждает, что

- все члены научного коллектива (в том числе руководитель проекта) удовлетворяют пунктам 6, 7, 13 конкурсной документации;
- на весь период реализации проекта он будет состоять в трудовых отношениях с организацией;
- при обнародовании результатов любой научной работы, выполненной в рамках поддержанного Фондом проекта, он и его научный коллектив будут указывать на получение финансовой поддержки от Фонда и организацию, а также согласны с опубликованием Фондом аннотации и ожидаемых результатов поддержанного проекта, соответствующих отчетов о выполнении проекта, в том числе в информационно-телекоммуникационной сети «Интернет»;
- помимо гранта Фонда проект не будет иметь других источников финансирования в течение всего периода практической реализации проекта с использованием гранта Фонда;
- проект не является аналогичным по содержанию проекту, одновременно поданному на конкурсы научных фондов и иных организаций;
- проект не содержит сведений, составляющих государственную тайну или относимых к охраняемой в соответствии с законодательством Российской Федерации иной информации ограниченного доступа;
- доля членов научного коллектива в возрасте до 39 лет включительно в общей численности членов научного коллектива будет составлять не менее 50 процентов в течение всего периода практической реализации проекта;
- в установленные сроки будут представляться в Фонд ежегодные отчеты о выполнении проекта и о целевом использовании средств гранта.

2 Содержание проекта

2.1 Научная проблема, на решение которой направлен проект

ru

Проект направлен на исследование задачи о поиске путей с ограничениями в терминах формальных языков с целью получения эффективного с прикладной точки зрения решения для неё для различных классов языков и различных видов графов.

Классы языков различаются своей выразительной возможностью, а значит, от используемого класса языка зависит то, на сколько сложные ограничения мы сможем задать. Например, при использовании в качестве ограничений регулярного языка не получится найти пути, задающие сбалансированную скобочную последовательность, так как язык сбалансированных скобочных последовательностей не является регулярным. Но он является контекстно-свободным, а значит используя контекстно-свободные языки мы сможем описать требуемое ограничение. С прикладной точки зрения используемый для ограничений класс языков позволяет ответить на вопрос "на сколько выразительный тот или иной язык запросов к графовой базе данных". Вместе с этим существует и другой вопрос: на сколько выразительный язык запросов можно создать в принципе? Ответ на этот вопрос требует работы на стыке теории графов и теории формальных языков. В самом простом случае, при проверке наличия хотя бы одного пути в графе, удовлетворяющего заданным ограничениям, мы приходим к задаче проверки пустоты пересечения двух языков: языка, заданного в качестве ограничений и регулярного языка, который задаётся графом в допущении, что все вершины являются стартовыми и финальными состояниями одновременно. Известно, что существуют содержательные с прикладной точки зрения классы языков, для которых задача проверки пустоты пересечения с регулярным неразрешима в общем случае. Например, конъюнктивные языки, предложенные Александром Охотиным. Использование такого класса в качестве ограничений в языке запросов приведёт к тому, что у пользователя появится возможность писать невыполнимые запросы. Стоит отметить, что с прикладной точки зрения, в таком случае ценным результатом может быть приближённый ответ. При этом необходимо уметь оценивать "качество" приближения (сколько информации потеряно, сколько добавлено лишней).

Вместе с этим, даже для тех классов языков, для которых задача разрешима, предъявление эффективных алгоритмов до сих пор является нетривиальной задачей. Для самого простого и хорошо изученного класса ограничений — регулярных ограничений (используются регулярные языки) — до сих пор продолжают поиски удачного алгоритма для работы в распределённых системах. Так, в 2016 году М. Ноле и К. Сартани предложили алгоритм выполнения запросов с такими ограничениями, основанный на производных Бжзовского, который естественным образом реализуем в терминах параллелизма уровня вершин (Maurizio Nolé and Carlo Sartiani, Regular Path Queries on Massive Graphs, 2016). Для более выразительного класса языков — контекстно свободного — до сих пор открыт вопрос о существовании субкубического алгоритма. Попытки же реализовать существующие алгоритмы в рамках графовой базы данных Neo4j привели Й. Куйперса и соавторов к выводу, что они не эффективны для решения прикладных задач, а значит надо продолжать поиск эффективных алгоритмов и подклассов задач, для которых можно реализовать эффективные алгоритмы (Jochem Kuyipers, George Fletcher, Nikolay Yakovets, and Tobias Lindaaker, An Experimental

Study of Context-Free Path Query Evaluation Methods, 2019).

Помимо теоретических основ и эффективных алгоритмов необходимо предоставить механизм, позволяющие задавать соответствующие ограничения в прикладных задачах. В современном мире редко встречается анализ графов как изолированная задача. Как правило необходима интеграция с прикладными решениями, которые разрабатываются с использованием языков общего назначения. Здесь возникает задача “естественной” интеграции спецификации синтаксических ограничений в языки программирования общего назначения, которая удачно решена для задач синтаксического анализа с применением парсер комбинаторов, что дало возможность решать задачи синтаксического анализа в терминах используемого языка программирования. Использование комбинаторов обеспечивает большую гибкость (можно организовывать переиспользование и модульность всеми средствами используемого языка) и безопасность (например, благодаря тому, что происходит "сквозная" проверка типов). При этом, даже в контексте работы с линейным входом некоторые проблемы были решены сравнительно недавно несмотря на длительную историю изучения парсер комбинаторов. Так, в 2016 году А. Измайлова с соавторами представила парсер-комбинаторы, способные работать с произвольными спецификациями контекстно-свободных языков (Anastasia Izmaylova, Ali Afroozeh, and Tijds van der Storm. 2016. Practical, general parser combinators). До этого момента существовали ограничения, такие как отсутствие левой рекурсии, отсутствие неоднозначностей и т.д. Одно из преимуществ использования парсер-комбинаторов — возможность вычисления пользовательской семантики, однако вопрос о том, для каких классов входных графов возможно точное вычисление семантики, не изучен. Также не изучено, можно ли использовать парсер-комбинаторы, задающие языки, не являющиеся контекстно-свободными, в качестве ограничений для поиска путей. Разработка языка для спецификации запросов, обладающего как можно большей выразительностью без потери точности результата — нетривиальная проблема.

и интеграция $\langle \text{Катя!} \rangle$ — надёжные решения. LINQ — Что позволило сделать обработку данных более однородной. Применение данного подхода для анализа графов изучено слабо.

Что-то про специализацию $\langle \text{Даня, про то, какие проблемы есть в специализации и прочих смешанных вычислениях, которые мы попробуем зарешать!} \rangle$

en

2.2 Научная значимость и актуальность решения обозначенной проблемы

ru

Знание границ разрешимости задачи необходимо для разработки языков запросов, для оценки разрешимости прикладных задач, сводимых к данной. При этом, с практической точки

зрения могут оказаться содержательными ситуации, когда задача в общем случае не разрешима, но можно найти достаточно точные приближённые решения. Так, для статического анализа применимым оказывается приближение сверху, так как в большинстве случаев ожидаемый ответ пуст, что означает отсутствие нежелательных поведений анализируемой программы. А значит, если аппроксимация сверху пуста, то и точное решение пусто. При этом важно, чтобы приближение как можно меньше отличалось от точного решения, так как в противном случае будет большое количество ложных срабатываний — ситуаций, когда найденное нежелательное поведение на самом деле не возможно. Примером такого подхода может служить работа Ц. Чжана, в которой для статического анализа кода применялись ограничения в виде линейных конъюнктивных языков (Qirun Zhang and Zhendong Su, Context-sensitive data-dependence analysis via linear conjunctive language reachability, 2017). В такой постановке задача неразрешима, однако показано, что можно эффективно искать содержательное с практической точки зрения приближенное решение.

Знание теоретических свойств алгоритмов важно как само по себе, так и для того, чтобы создавать эффективные на практике решения. Стоит отметить, что, несмотря на то, что данная область изучается уже длительное время, совсем недавно были получены новые результаты. Так, в 2017 году Ф. Брэдфорд предъявил субкубический алгоритм для задачи достижимости в случае, когда ограничения заданы языком Дика на одном типе скобок (Phillip G. Bradford, Efficient Exact Paths For Dyck and semi-Dyck Labeled Path Reachability). Предложенное решение не обобщается на произвольные контекстно-свободные ограничения и требуется дальнейшая работа в данном направлении. В 2017 году К. Чаттерджи предъявил оптимальный алгоритм проверки достижимости для специального вида графов (двунаправленные графы) в случае, когда ограничения сформулированы в виде произвольного языка Дика (Krishnendu Chatterjee, Optimal Dyck reachability for data-dependence and alias analysis). Также К. Чаттерджи показал, что предложенный алгоритм может эффективно применяться на практике для решения задач статического анализа кода.

Поиск эффективных с вычислительной точки зрения алгоритмов, в том числе алгоритмов для массово-параллельных и распределённых систем, с одной стороны важен для более глубокого понимания теоретических свойств алгоритмов и развития теории, связанной с параллельными и распределёнными системами, а с другой — для создания эффективных решений для прикладных задач, например, графовых баз данных, которые становятся всё более популярными. Как уже было сказано, поиск эффективных алгоритмов даже для хорошо изученных классов задач является актуальной на сегодняшний день проблемой (например, работы Jochem Kuijpers и Maurizio Nolé).

Исследования в области способов задания ограничений связаны с разработкой языка запросов, что является актуальной задачей. С одной стороны, языки запросов к графовым базам данных только развиваются и многие даже базовые вопросы, связанные с синтаксисом и семантикой таких языков, требуют изучения. С другой стороны, существует ряд общих вопросов, связанных с интеграцией предметно-ориентированных языков в языки общего назначения. Например, вопросы о "бесшовной" интеграции, о типовой безопасности, о различных проверках времени компиляции. Для решения этих проблем регулярно предлагаются различные подходы: интегрированный язык запросов (LINQ), различного рода комбинаторы, средства "межязыкового" вывода типов.

Специализация и смешанные вычисления !!! $\langle \text{Даня!} \rangle$ — изучается давно (ещё со времён Турчина), но до сих пор много открытых вопросов как в теории так и относительно применимости. Активная область исследований. Недавно специализировали машинный код, а ещё Postgres и вообще специализация времени выполнения. Позволило улучшить производительность вычисления запросов к реляционной. Как было сказано, проблемы вычислительной эффективности стоят остро, так что почему бы и нет.

en

2.3 Конкретная задача (задачи) в рамках проблемы, на решение которой направлен проект, ее масштаб и комплексность

ru

В рамках исследования границ разрешимости задачи поиска путей с ограничениями в терминах формальных языков и изучения формальных свойств алгоритмов для решения этой задачи предполагается исследовать новые подклассы задачи для различных классов языков и типов графов. Прежде всего планируется исследовать различные подклассы контекстно-свободных языков, где преследуются две цели — как можно ближе подойти к ответу на вопрос о существовании субкубического алгоритма для решения задачи и поиск содержательных с прикладной точки зрения подклассов, для которых возможна реализация вычислительно эффективных алгоритмов. Вместе с этим планируется изучение различных типов задач и алгоритмов их решения (конструирование алгоритмов и изучение их теоретических свойств, таких как временная и пространственная сложность): поиск единственного пути, удовлетворяющего ограничениям, поиск кратчайшего пути и т.д. Кроме этого, будут изучены более широкие, чем контекстно-свободный, классы языков с точки зрения их применимости в качестве ограничений.

В области разработки параллельных и распределённых алгоритмов планируется конструирование, теоретическое и экспериментальное исследование алгоритмов для решения задачи достижимости с ограничениями в терминах формальных языков, эксплуатирующих различные типы параллелизма, такие как массовый параллелизм (SIMD), многопоточность и многоядерность. Акцент предполагается сделать на задаче с контекстно-свободными ограничениями. Предполагается, что будут рассмотрены различные подходы и модели для разработки параллельных алгоритмов, такие как параллелизм уровня вершин, сведение задач к задачам с известными эффективными параллельными алгоритмами. В ходе работы планируется изучить и сравнить в контексте решаемой задачи различные способы представления данных. Несмотря на активное развитие графовых баз данных и соответствующей теории, единого мнения относительно того, как именно лучше представлять графы, нет. Отчасти это связано с тем, что особенности решаемой задачи и используемых алгоритмов накладывают специфические ограничения, которые в области исследуемой задачи изучены фрагментарно.

Вопросы, связанные с языками запросов, поддерживающими ограничения в терминах формальных языков, будут связаны, прежде всего, с применимостью парсер комбинаторов для этой задачи, а так же с изучением ограничений, которые возникают при их использо-

вании. В частности, планируется изучить ограничения, накладываемые на семантические функции, так как они более строгие, по сравнению с линейным входом, что связано потенциальной бесконечностью путей. Также планируется экспериментальное исследование механизма интеграции запросов в код на языках общего назначения, основанного на парсер комбинаторах. Планируется сравнение с другими подходами в таких аспектах, как выразительность, модульность, предоставляемые средства повышения надёжности кода.

Планируемые к изучению вопросы оптимизации времени выполнения запросов связаны с двумя направлениями. Первое — оптимизация описания ограничений. Известно, что один и тот же язык можно описать несколькими разными грамматиками. В задачах синтаксического анализа языков программирования хорошо заметно, что свойств конкретной грамматики зависит реальное время разбора (при фиксированном инструменте и входе). подобное поведение наблюдается и при анализе графов, однако не все результаты переносимы с линейного случая. Планируется изучить способы оптимизации запросов с ограничениями в терминах контекстно-свободных языков, реализовать соответствующие алгоритмы и провести их экспериментальное исследование. Второе направление — оптимизация алгоритмов на уровне компилятора. Здесь планируется **<Даня! Специализация!>**

en

2.4 Научная новизна исследований, обоснование достижимости решения поставленной задачи (задач) и возможности получения запланированных результатов

ru

Рассматриваемая в проекте область активно развивается. Все поставленные задачи интересуют специалистов в соответствующих областях, что подтверждается наличием работ, опубликованных в недавнее время в рецензируемых профильных журналах и представленных на ведущих профильных конференциях, в том числе участниками проекта. Это позволяет гарантировать новизну ожидаемых результатов и их соответствие мировому уровню.

Поскольку некоторые задачи очень трудны, гарантировать их полное решение невозможно. Таковой, например, является задача о существовании субкубического алгоритма для задачи достижимости с контекстно-свободными ограничениями. Однако получение даже частичных результатов или улучшение существующих (например, расширение границ применимости алгоритма Брэдфорда) будет существенным вкладом. Вместе с этим, в проекте предусмотрено решение ряда интересных и ожидаемо разрешимых задач.

Например, опыт участников в теории формальных языков, теории графов и алгоритмах синтаксического анализа позволит всесторонне подойти к вопросу поиска подклассов задачи о поиске путей с ограничениями в терминах формальных языков. Важно, что как положительные, так и отрицательные результаты в решении данной задачи важны: ценны как подклассы, для которых существуют эффективные алгоритмы, так и доказательства того, что для каких-то классов задач таких алгоритмов нет.

Для задачи поиска путей с контекстно-свободными ограничениями поиск эффективных с вычислительной точки зрения алгоритмов активно ведётся в настоящее время, однако удовлетворительных решений, по итогам исследования 2019 года проведённого Й. Куйперсом и соавторами, не предъявлено. Вместе с тем, у участников проекта (Р. Азимова, С. Григорьева, Е. Вербицкой) есть большой опыт разработки алгоритмов для данной задачи, в том числе, Р. Азимовым предложен алгоритм, основанный на матричных операциях, позволяющий использовать параллельные вычисления для решения задачи. Это способствует плодотворному поиску новых алгоритмов, их изучению и проведению всесторонних экспериментальных исследований.

Решение задачи интеграции языка запросов на основе парсер комбинаторов будет основано на результатах, полученных Д. Крёни, не затрагивающих, однако ряда важных вопросов, таких как класс поддерживаемых языков (какие языки можно использовать в качестве ограничений) и опыте Е. Вербицкой, занимающейся изучением парсер-комбинаторов применительно как к анализу линейного входа, так и к анализу графов. Кроме того, Е. Вербицкая разработала алгоритм поиска путей с контекстно-свободными ограничениями, основанный на восходящем синтаксическом анализе.

Планируется, что решение задачи, связанной с применением специализации для оптимизации времени выполнения запросов, будет основано на опыте Е. Ю. Шарыгина, показавшего, что данный подход позволяет существенно ускорить выполнение запросов в реляционных базах данных. Данный подход не применялся к алгоритмам выполнения запросов к графовым базам данных, поэтому может потребоваться разработка новых алгоритмов или существенная доработка существующих. Опыт участников проекта Е. А. Вербицкой и Д. А. Берерзуна в применении и разработке методов смешанных вычислений, в том числе специализации, должен помочь решить эту задачу.

en

2.5 Современное состояние исследований по данной проблеме, основные направления исследований в мировой науке и научные конкуренты

ru

В мировом научном сообществе активно ведутся работы в областях, связанных с обработкой граф-структурированных данных и, в частности, связанных с графовыми базами данных. Исследователями всего мира изучаются как теоретические аспекты задачи поиска путей с ограничениями в терминах формальных языков, так и прикладная сторона вопроса.

После двух классических работ, в которых была сформулирована общая задача поиска путей с контекстно-свободными ограничениями в разных областях — Т. Репсом в статическом анализе кода (Т. Reps, 1997, Program analysis via graph reachability) и М. Яннакакисом в графовых базах данных (М. Yannakakis, 1990, Graph-theoretic methods in database theory) — ведутся активные работы как по детальному исследованию этих задач, так и по изучению

проблемы поиска путей с языковыми ограничениями в целом (C. Barrett, R. Jacob, and M. Marathe, 2000, Formal-language-constrained path problems).

В частности, исследуются новые прикладные задачи, которые могут быть сформулированы в терминах таких запросов. Например, анализ биологических данных (P. Sevon and L. Eronen, 2008, Subgraph queries by context-free grammars), анализ онтологий или RDF (C. M. Medeiros, M. A. Musicante, and U. S. Costa, 2019, LL-based query answering over rdf databases и X. Zhang, Z. Feng, X. Wang, G. Rao, and W. Wu, 2016, Context-free path queries on rdf graphs), вывод спецификаций для программного кода (Osbert Bastani, Saswat Anand, and Alex Aiken, 2015, Specification Inference Using Context-Free Language Reachability), анализ алиасов в программном коде (Dacong Yan, Guoqing Xu, and Atanas Rountev, 2011, Demand-driven context-sensitive alias analysis for Java) и другие. Кроме этого, находят применение и более широкие классы языков, например линейные конъюнктивные, которые могут быть применены для статического анализа программ (Qirun Zhang and Zhendong Su, 2017, Context-sensitive data-dependence analysis via linear conjunctive language reachability).

Параллельно с этим ведутся теоретические исследования в области оптимальных алгоритмов для различных классов подзадач. Одним из основополагающих результатов здесь является результат Л. Валианта, показавшего, что синтаксический анализ линейного входа с применением контекстно-свободных грамматик возможен за менее чем кубическое время (L. G. Valiant, 1975, General context-free recognition in less than cubic time). Возможность обобщения этого результата с сохранением временной сложности на произвольный граф является одним из основных открытых вопросов. В последнее время получен ряд серьёзных результатов в этом направлении. Так, в 2017 году К. Чаттерджи предъявил оптимальный алгоритм для поиска путей в специальном типе графов (двунаправленные графы) с ограничениями в виде произвольного языка Дика (Krishnendu Chatterjee, Bhavya Choudhary, and Andreas Pavlogiannis, 2017, Optimal Dyck reachability for data-dependence and alias analysis). А Ф. Брэдфорд в 2017 предъявил субкубический алгоритм для задачи достижимости в произвольном графе но с ограничениями в виде языка Дика на одном типе скобок (Ph. G. Bradford, 2017, Efficient exact paths for dyck and semi-dyck labeled path reachability). Также теоретическими исследованиями в данной области занимался Й. Хеллингс (J. Hellings, 2015, Path results for context-free grammar queries on graphs и другие работы 2014-2015 годов).

Разработкой и изучением алгоритмов для поиска путей с контекстно свободными ограничениями активно занимаются группы под руководством Ф. Брэдфорда в университете Коннектикут, США (P. G. Bradford and V. Choppella, 2016, Fast point-to-point dyck constrained shortest paths on a dag), под руководством М. Мусиканте, Universidade Federal do Rio Grande do Norte, Бразилия (Fred C. SantosUmberto S. CostaMartin A. Musicante, 2018, A Bottom-Up Algorithm for Answering Context-Free Path Queries in Graph Databases), под руководством Дж. Флетчера, Technische Universiteit Eindhoven, Нидерланды. При этом, исследование группы Дж. Флетчера 2019 года показало, что существующие алгоритмы не применимы для решения прикладных задач, при том, что они являются достаточно востребованными (Jochem Kuijpers, George Fletcher, Nikolay Yakovets, and Tobias Lindaaker, 2019, An Experimental Study of Context-Free Path Query Evaluation Methods).

Разработкой языков запросов к графовым базам данных с поддержкой ограничений в терминах формальных языков занимается большая международная группа, в состав кото-

рой входят, в том числе, Т. Линдакер и Дж. Флетчер (Renzo Angles, Marcelo Arenas, Pablo Barcelo, Peter Boncz, George Fletcher, Claudio Gutierrez, Tobias Lindaaaker, Marcus Paradies, Stefan Plantikow, Juan Sequeda, Oskar van Rest, and Hannes Voigt, 2018, G-CORE: A Core for Future Graph Query Languages). При этом вопросы интеграции таких языков в языки общего назначения изучены достаточно слабо. Подход, основанный на парсер комбинаторах, изучался в работах Д. Крёни (Daniel Kröni and Raphael Schweizer, 2013, Parsing graphs: applying parser combinators to graph traversals) и Е. Вербицкой (Ekaterina Verbitskaia, Илья Kirillov, Илья Nozkin, and Semyon Grigorev, 2018, Parser combinators for context-free path querying). Данное направление находится на начальной стадии. Несмотря на то, что впервые использовать функции высшего порядка для синтаксического анализа было предложено Burge в 1975 году, а в 1990 Wadler предложил идею монадических комбинаторов, леворекурсивные спецификации и неоднозначности вывода представляли проблему. В 2016 году она была решена Анастасией Измайловой и Али Афрузехом (Practical, general parser combinators).

При этом применение парсер комбинаторов для синтаксического анализа линейного входа активно развивается, несмотря на длительную историю изучения. Так, Миркат, что-то ещё

<Катя, Обзор!>

Вопросы специализации <Даня, Обзор!> Обзор алгоритмов специализации, направления (связанные с нашей работой)

en

2.6 Предлагаемые методы и подходы, общий план работы на весь срок выполнения проекта и ожидаемые результаты

ru

При поиске подклассов задач, для которых могут быть представлены эффективные алгоритмы предполагается привлечь методы теории формальных языков, теории графов и алгоритмов синтаксического анализа и рассмотреть различные комбинации типов задач (поиск одного пути, поиск всех возможных путей, поиск путей из заданной вершины и так далее) различных подклассов контекстно-свободных языков (линейные контекстно-свободные, one-counter языки и другие), различных типов графов (деревья, ациклические, произвольные). Ожидаемые типы ожидаемых результатов здесь — нижние оценки вычислительной сложности для алгоритмов, решающих соответствующие типов задач, алгоритмы для практически интересных случаев, принадлежность или не принадлежность того или иного типа задач тому или иному классу вычислительной сложности.

Далее планируется изучить результаты, касающиеся получения субкубического алгоритма, полученные в смежных областях, таких как language editing distance, поиск кратчайших путей в различных типах графов. С использованием этих результатов предпринять попытку обобщить результаты Л. Валианта и Ф. Брэдфорда до произвольных графов и произвольных контекстно-свободных языков.

При разработке эффективных с вычислительной точки зрения алгоритмов планируется применять результаты, полученные для алгоритмов синтаксического анализа. Одно из основных направлений — попытки обобщить алгоритмы, применимые к линейному входу, до графов. Планируется, в частности, обобщить решение для регулярных ограничений, построенное на производных Бжзовского, так как сам механизм производных обобщаем для контекстно-свободных языков, а решение для регулярных, основанное на данном механизме, оказалось эффективно распараллеливаемым в модели параллелизма уровня вершин. Кроме этого, при работе над данной задачей будут привлекаться методы линейной алгебры, так как одно из перспективных направлений связано с формулировкой алгоритмов в терминах линейной алгебры. При теоретическом исследовании алгоритмов будут применяться методы теории алгоритмов.

Для решения задачи о применении парсер комбинаторов для анализа графов планируется использовать методы и результаты функционального программирования, теории типов и теории формальных языков. Для определения ограничений на семантические функции и графы для точного вычисления семантики потребуются знания из теории статического анализа кода, теории решеток. Методы программной инженерии будут использованы для разработки формализма описания запросов, допускающую большую параметризуемость и переиспользуемость запросов. <Катя!>

Что-то про специализацию <Даня!>

*** 2020 ***

*** 2021 ***

*** 2022 ***

en

2.7 Имеющийся у научного коллектива научный задел по проекту, наличие опыта совместной реализации проектов

ru

Руководитель проекта и многие его участники обладают опытом в разработке и исследовании алгоритмов синтаксического анализа, и их применении в различных областях, в том числе для анализа поиска путей в графах, что подтверждается соответствующими работами (Grigorev, Ragozina, "Context-free path querying with structural representation of result SECR-2017; Azimov, Grigorev, "Context-free path querying by matrix multiplication GRADES-NDA-2018; Verbitskaia, Kirillov, Nozkin, Grigorev, "Parser combinators for context-free path querying Scala-2018; Mishin, Grigorev, et.al. "Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication GRADES-NDA-2019).

Руководитель принимал успешное участие в совместной с Е.А. Вербицкой и Д.Березуном работе над проектом в рамках гранта РФФИ 18-01-00380. Также, С.В.Григорьев являлся исполнителем грантов РФФИ 15-01-05431, Фонда содействия развитию малых форм предпри-

ятий в технической сфере(программа УМНИК, проекты N 162ГУ1/2013 и N 5609ГУ1/2014), руководителем гранта РФФИ 19-37-90101, а также является руководителем научной группы, в соавторстве с участниками которой опубликованы указанные выше и некоторые другие работы.

С.В.Григорьевым и А.К.Рагозиной предложен алгоритм поиска путей с контекстно-свободными ограничениями на основе обобщённого нисходящего синтаксического анализа, доказана его корректность, получены оценки временной и пространственной сложности.

Е.А.Вербицкой и С.В.Григорьевым предложен алгоритм поиска путей с контекстно-свободными ограничениями на основе обобщённого восходящего синтаксического анализа, доказана его корректность, проведены экспериментальные исследования. Также начато изучение применимости парсер-комбинаторов для анализа графов. < Мы же тут даже что-то опубликовали, не? > Кроме этого Е.А.Вербицкая предложила механизм поддержки левой рекурсии в библиотеке парсер-комбинаторов Ostop. <Катя! что-то про другие свои работы>

Р.Ш.Азимовым и С.В.Григорьевым предложен алгоритм поиска путей с контекстно-свободными ограничениями на основе матричных операций, доказана его корректность, получена оценка временной сложности. Кроме того, предложено обобщение данного алгоритма, в котором в качестве ограничений над путями используются конъюнктивные грамматики, позволяющие выражать более сложные запросы к графам. Для обобщенного алгоритма также доказана корректность и получена оценка временной сложности.

2.8 Перечень оборудования, материалов, информационных и других ресурсов, имеющих у научного коллектива для выполнения проекта

ru

Использование особых ресурсов не предполагается.

2.9 План работы на первый год выполнения проекта

ru

Планируется работа над заранее намеченными на этот год исследовательскими задачами, предоставление результатов на конференциях и подготовка результатов к печати. Также будет проведено осмысление полученных результатов с возможной формулировкой новых задач. В целом предполагается следующее распределение задач между основными исполнителями проекта.

С.В.Григорьев займётся разработкой параллельных алгоритмов для поиска путей с контекстно-свободными ограничениями и изучением их свойств. Будут исследоваться алгоритмы, основанные на различных матричных операциях. Планируется рассмотреть различные способы

представления разреженных матриц, выяснить масштабируемость таких алгоритмов, провести экспериментальное исследование, сравнение между собой и с аналогами.

Р.Ш. Азимов займётся разработкой алгоритма для поиска путей с контекстно-свободными ограничениями, использующего тензорное произведение матриц (произведение Кронекера) и работающего с матрицами существенно большего размера. В основе подхода лежит использование рекурсивных сетей или рекурсивных автоматов в качестве представления контекстно-свободных грамматик. Планируется найти теоретические свойства полученного алгоритма и провести экспериментальное исследование.

Изучением применимости парсер-комбинаторов для анализа графов займётся Е. А. Вербицкая. <Катя! подробнее задачи>

Вопросами оптимизации алгоритмов выполнения запросов будет заниматься Д. А. Березун. <Даня! подробнее задачи>

Шеметова Е. Н. будет заниматься изучением вопроса существования и построения субкубического алгоритма для задачи достижимости с контекстно-свободными ограничениями. В рамках исследования существования субкубического алгоритма будут даны условные нижние оценки времени работы классического кубического алгоритма на предмет его оптимальности (fine-grained complexity), также будут изучены возможности сведения задачи к другим известным задачам, для которых субкубического или субквадратического алгоритма не существует. Также планируется разработка алгоритмов, вычисляющих аппроксимацию решения задачи за субкубическое и более оптимальное время, а также алгоритмов, эффективно решающих задачу для полезных на практике подклассов графов или контекстно-свободных грамматик. В рамках построения субкубического алгоритма для общего случая будут изучена возможность сведения вычислений к матричному умножению в $(\min, +)$ -полукольце, в котором свойства элементов матриц позволяют осуществить данное умножение эффективно за субкубическое время.

Ю.А. Сусанина займётся вопросами сведения алгоритма поиска путей сконтекстно-свободными ограничениями, основанного на матричных операциях, к задаче решения (систем) матричных уравнений. Далее предполагается рассмотреть возможность ускорения процесса поиска путей за счет применения известных численных методов для нахождения корней уравнений (например, метод Ньютона). Также планируется исследовать частные случаи сведения задачи решения (систем) матричных уравнений к задаче синтаксического анализа графов.

< Непонятно, что писать, написал обобщенно, оптимизации вроде же Даня> Н.М. Мишин будет заниматься вопросами оптимизации и эффективной реализации алгоритмов выполнения запросов и адаптации их под эффективное вычисления на GPU. В рамках реализации алгоритмов будет учитываться необходимость обработки огромного объема данных и распараллеливание алгоритмов на несколько видеокарт. Для достижения максимальной эффективности будет проведена серия экспериментальных исследований с целью сравнения реализаций, выяснения скрытых констант и масштабируемости соответствующих алгоритмов.

К обсуждению всех задач, работе над ними, и написанию статей будут привлекаться включённые в состав научного коллектива студенты, магистры и аспиранты.

Также на первый год планируется К поездок с докладами на международные конференции.

en

2.10 Ожидаемые в конце первого года конкретные научные результаты

ru

Будет описан и реализован алгоритм поиска путей с контекстно-свободными ограничениями, основанный на пересечении рекурсивного автомата и графа. Изучены его свойства, проведено экспериментальное исследование. Сколько-то работ будут представлены на конференциях. Они будут опубликованы в сборниках докладов, индексируемых !!!

Будет представлен алгоритм поиска путей с контекстно-свободными ограничениями, основанный на решении (систем) матричных уравнений. Будут изучены свойства полученного алгоритма и предложена реализация с применением метода Ньютона. Также планируется рассмотреть сведение задачи решения (систем) линейных уравнений к задаче поиска путей в графах для частных случаев — взвешенных стохастических грамматик.

<От остальных, что сделаем?>

en

2.11 Перечень планируемых к приобретению руководителем проекта за счет гранта Фонда оборудования, материалов, информационных и других ресурсов для выполнения проекта

ru

Не более 800 тыс. рублей ежегодно будет тратиться на поездки с докладами на конференции. Расходов на оборудование не предполагается.