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

Семён Григорьев

17 октября 2019 г.

1 Сведения о проекте

1.1 Название проекта

ru

Разработка алгоритмов анализа граф-структурированных данных, основанных на 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, специализация.

en

1.5 Аннотация проекта

ru

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

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

Одним из способов задать ограничение на путь в нагруженном графе (то есть в графе, рёбра которого несут некоторую нагрузку в виде метки или веса) использует формальные языки. В данном случае рассматриваются слова, полученые 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 лчшей 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 задчи: 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, асимптотики, 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 начала изучаться в начале 90-х (Томас Репс и Михалис Яннакакис), многие вопросы остаются открытыми. Например, до сихпор не решён вопрос о сущесвовании 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, кроме теоретических результатов, 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 стредства описания запросов к таким базам, позволяющие формулировать ограничения в терминах формальных языков. Одним из классичеких 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 специальизированного 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. Токой 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. Стоит отметить, что несмотря на длительную историю исследований в области смешанных вычислений, при решении новых задач часто возникают ситуации, требующие разработки новых формальных методов.

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, впринципе 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 ипользоваться методы построения парарллелльных алгоримтов, в том числе, алгоритмов для массово-парарллелльных архитектур. Исследование способов задания ограничений потребует испльзования 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

прораммирования. Это позволит организовать плодотворное сотрудничество и обеспечить комплексний подход к решению задач, а также привлечь к изучению талантливых студентов к соответтсвующим 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 в на стыке теории формальных языков и теории графов и в области построения параллельных алгоритмов, так и результаты в области разработи языков и методов оптиммизации программного обеспечения.

В частности, ставится задача построить более детальную классификацию задач и поиске путей сконтекстно 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 удасться, но ценными будут и частичные ответы в терминах подклассов задачи, для которых такой алгоритм точно существует.

В области построения параллельных алгоритмов планируется 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 пространственной сложночсти. 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 контескте прикладных задач, такие как типобезопастность, возможность вычисления дополнительных сематнических функций. Несмотря на то, что применение прарсер-комбинаторов для анализа языков программирования изучено достаточно хорошо, обобщение этого подхода на графы нетривиально и ожидаются новые результаты. $<$ 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 алгоримты специлизации $<$ 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 Екатерина Андреевна Вербицкая — встроенные языки, комбинаторы, ФП, суперкомпиляция Даниил Андреевич Березун — суперкомпиляция, кфмн Рустам Азимов — графы, поиск путей в графах, формальные языки, параллельные алгоритмы. Григрьев Семён — графы, формальные языки, алгоритмы поиска путей, руководство грантаим, аспирантами, магистрами и т.д. кфмн

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 солько 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 опись 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 впринципе? Ответ на эот 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 к вачестве 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 сихпор является нетривиальной задачей. Для самого простого и хорошо изученного класса ограничений — решулярных ограничений (используются регулярные языки) — до сих пор продолжаются поиски удачного алгоритма для работы в распределённых системах. Так, в 2016 году М. Ноле и К Сартиани предложили алгоритм выполнения запросов с такими огранчениями, 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 Kuijpers, 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, позволяющее задавать соответствующие ограничения в прикладных задачах. В современном мире редко встречается анализа графов "сам по себе". Как правило необходима интеграция с прикладными решениями, которые разрабатывются с испльзованием языков общего назначения. Здесь возникает задача "есественной"интеграции спецификации синтаксических ограничений в языки программирования общего назначения, которая удачно решена для задач синтаксического анализа с применениеми парсер комбинаторов, что дало возможность решать задачи синтаксического анализа в терминах используемого азыка программирования. Использование комбинаторов обеспечивает большую гибкость (можно организовывать переиспользование и моульность всеми средствами используемого языка) и безопасность (например, благодаря тому, что происходит "монолитная"проверка типов). При этом, даже в контексте работы с линейным входом некоторые проблемы были решены срвнительно 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 Tijs van der Storm. 2016. Practical, general parser combinators). До этого момента существовали ограничения, такие как отсутсвие левой рекурсии, отсутсвие неоднозначностей и т.д. и интеграция <Катя!> — надёжные решения. LINQ — Что позволило сделать обработку данных более однородной. Применение данного подхода для анализа графов изучено слабо.

Что-то про специализацию <Даня, про то, какие проблемы есть в специализации и прочих смешанных вычислениях, которые мы попробуем зарешать!>

en

2.2 Научная значимость и актуальность решения обозначенной проблемы

ru

Знание границ разрешимости задачи необходимо для дизайна языков запросов, для оценки разрешимости прикланых задач, сводимых к данной. При этом, с практической точки зрения могут оказаться содержательными ситуации, когда задача в общем случае не разрешима, но можно найти "хорошие"приближённые решения. Так, для статического анализа применимым оказывается приближение сверху, так как в большинстве случаев ожидаемый ответ пуст, что означает отсутсвие нежелательных поведений анализаируемой программы. А значит, если апроксимация сверху пуста, то и точное решение пусто. При этом важно, чтобы приближение как можно меньше отличалось от точного решение, так как в противном случае будет большое количество ложных срабатываний — ситуаций, когда найденное нежелательное поведение на самом деле не возможно. Примером такого подхода может слу-

жить 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 практичекой точки зрения приближонное 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, сосвсем 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 эффетивно применяться на практике для решения задач статического анализа кода.

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

Исследования в области способов задания ограничений связанны с разработкой языка запросов, что являетс 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, связвнных 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 "межъязыкового"вывода типов.

Специализация и смешанные вычисления !!! $<$ Даня! $>$ – 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 поризводительность вычисления зарпосов к реляционной. Как было сказано, проблемы вычислительной эффективноси 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 преследутся две цели — как можно длиже подойти к ответу на вопрос о сущесвовании субкубического алгоритма для решения задачи и поиск содержательных с прикладной точки зрения подклассов, для которых возможна реализация вычислительно эффективных алгортмов. 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 конекстно-свободный, классы языков с точки зрения их применимости в качестве ограничений.

В области разработки параллельный и распределённых алгоритмов планируется конструирование, теоретическое и экспериментальное исследование алгоритмов для решения задачи достижимости с ограничениями в теминах 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 парарллелизм уровня вершин, сведение задач к задачам с известыми 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, бдут связаны, прежде всего, с применимостью парсер комбинаторов для этой задачи, а так же с изучением ограничений, которые возникают при их использовани. В частности, планируется изучить ограничения, накладываемые на семантические функции, так как они более строгие, по сравнению с линейным входом, что связано потенциальной бесконечностью путей, в отличие от линеного входа. 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 выррази тельность, 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 граммтиками. В задачах синтаксического анализа языков программирования хорошо заметно, что свойств конкретной грамматики зависит реальное время разбора (при фиксированном инстументе и входе). подобное поведение наблюдается и при анализе графов, однако не все результаты перерносимы 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 ожтдаемо 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 конекстно-свободными ограничениями поиск эффективных с вычислительной точки зрения алгоритмов активно ведётся в настоящее время, однако удовлетворительных решений, по итогам исседования 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 мтричных 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 к линейному входу и вопросами семантики языков программирования.

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. Бререзуна 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 ведуться работы в облостях, связанных с обработкой граф-структурированных данных и, в частности, связанных с графовыми базами данных. Исследователми 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 (T. Reps, 1997, Program analysis via graph reachability) и М. Яннакакисом в графовых базах данных (M. 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 вцелом (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. Напрмиер, анализ биологичесикх 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 про-

граммном 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, напмриер 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 ведуться 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 контекстносвободных граммтик возможен за менее чем кубическое время (L. G. Valiant, 1975, General context-free recognition in less than cubic time). Возможность обобщения этого результата на произвольный граф является одним из основных открытых вопросв. В последнее время получен ряд серьёзных результатов в этом направдении. 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 теретическими 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, исслдование 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 ограничеий в терминах формальных языков занимается большая международная группа, в состав которой входят, в том числе, Т. Линдакер и Дж. Флетчер (Renzo Angles, Marcelo Arenas, Pablo Barcelo, Peter Boncz, George Fletcher, Claudio Gutierrez, Tobias Lindaaker, 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 парер 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, Ilya Kirillov, Ilya Nozkin, and Semyon Grigorev, 2018, Parser combinators for context-free path querying). Данное направление находится на начальной стадии. При этом применение парсер комбинаторов для синатксческого 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 языко, теории графов и алгоритмов синтаксического анализа и рассмотреть тазличные 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, произвольние). Ожидаемые типы ожидаемых результатов здесь — нижние оценки вычислительной сложности для алгоритмов, решающих соответствующие типв задач, алгоритмы для практически интересных случаев, принадлежность или не пренадлежность 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 механим производных обощаем для контекстно-свободных языков, а решение для регулярных, онованное на данном механизе, оказалось эффективно распараллеливаемм в модели параллелизма уровня вершин. Кроме этого, при работе над данной задачей будут привлекаться меирды 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 <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 обладает опытом в разработке и исследовании алгоритмов синтаксического анализа, и их применении в различных областях, что подтверждается соответствующими статьями (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)

В том числе, у руководителя имеется опыт применения формальных грамматик и алгоритмов синтаксического анализа для решения задач в области биологии (биоинформатики), что подтверждается выступлениями на тематических конференциях Biata-2017/2018, BIOINFORMATICS-2019.

Кроме того, руководителем был предложен метод совмещения формальных грамматик и ИНС для анализа вторичной структуры, который предполагается развивать в рамках данного исследования. Метод был изложен в статье "The Composition of Dense Neural Networks and Formal Grammars for Secondary Structure Analysis"и представлен на конференции BIOINFORMATICS-2019.

Руководитель принимал успешное участие в совместной работе над проектами в рамках грантов РФФИ (15-01-05431 и 18-01-00380), Фонда содействия развитию малых форм предприятий в технической сфере (программа УМНИК, проекты N 162ГУ1/2013 и N 5609ГУ1/2014), а также является руководителем научной группы, в соавторстве с участниками которой опубликованы указанные выше и некоторые другие работы.

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

ru

2.9 План работы на первый год выполнения проекта

ru

en

2.10 Ожидаемые в конце первого года конкретные научные результаты

ru

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.