

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

Семён Григорьев

16 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.

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 программных 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 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 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, 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 поиска 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, 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 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, 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 средства 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. 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, в принципе 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 в 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 поиске путей 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 пространственной 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. < 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 гн Екатерина Андреевна Вербицкая — встроенные языки, комбинаторы, ФП, суперкомпиляция Даниил Андреевич Березун — суперкомпиляция, кфмн Рустам Азимов — графы, поиск путей в графах, формальные языки, параллельные алгоритмы. Григорьев Семён — графы, формальные языки, алгоритмы поиска путей, руководство грантами, 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 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, 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 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 "монолитная" проверка типов). 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 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 "хорошие" 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 приближённое 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.

Среди причин Поиск алгоритмов для массово-параллельных и распределённых систем — создание эффективных решения для прикладных задач, графовых баз данных и т.д. Регулярки и распределённые системы

Комбинаторы и интеграция $\langle \text{Катя!} \rangle$ — надёжные решения. LINQ — Что позволило сделать обработку данных более однородной.

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. Недавно специализировали машинный код, а ещё Postgres и вообще специализация времени выполнения.

en

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

ru

Разработать методы специализации, применимые для оптимизации запросов.

Комбинаторы!

Алгоритмы! В том числе параллельные для современных архитектур и т.д.

en

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

ru

Специализацию в данном контексте не применяли. Алгоритмы запросов ещё не до конца

изучены. Практически эффективных ещё нету.

en

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

ru

Обзор алгоритмов запросов.

Обзор алгоритмов специализации.

Тра-та-та

en

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

ru

Специализация.

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, имеющих 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 200 тыс. рублей ежегодно будет тратиться на поездки с докладами на конференции.

Расходов на оборудование не предполагается.