

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

Семён Григорьев

11 ноября 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**

Graph theory, formal language theory, path querying, graph database, formal grammar, parsing, algorithm optimization, parallel algorithms, mixed computations, partial evaluation, specialization.

**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 используют дополнительные, не тривиальные, ограничения 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) основан на использовании формальных языков. В данном случае рассматриваются слова, полученные конкатенацией меток рёбер пути, и задаётся язык, которому должны принадлежать такие слова. Иными

словами, возникает следующая задача: найти такие пути в графе, 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 конкретной решаемой задачи необходимо применять различные алгоритмы для достижения лучшей эффективности.

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 они требуют модификации, и 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 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.

Если рассматривать задачу поиска путей в контексте графовых баз данных, то необходимо предоставить удобные средства описания запросов, позволяющие формулировать ограничения в терминах формальных языков. Одним из наиболее естественных способов задавать такие ограничения в прикладных задачах являются парсер-комбинаторы. Традиционно парсер-комбинаторы используются для задания языка и, одновременно, синтаксического анализатора для него, путём комбинирования функций, реализующих более простые парсеры. Парсер-комбинаторы обеспечивают при этом "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 талантливых студентов к изучению соответствующих областей науки и работе над проектом.

en

Big data processing is a research area that requires strong theoretical results to create applied solutions of high-quality. A graph often serves as a model for representing data. Graphs are used to represent networks (social networks, flow networks), for static analysis and verification of software (call graphs, data-flow graphs), and are at the heart of graph databases. Although a graph is a fundamental mathematical object, it is still being actively researched. Processing of graph structured data is an actively developing field.

One of the data processing problems is determining relations between entities or the absence of specific relations. In the case of graph-structured data, this problem can be formulated in terms of searching for paths between vertices or checking that there are no paths. Usually, when solving real-world problems, some specific nontrivial constraints are applied to the paths. Constraints can take different forms. For example, one can constrain the length of a path and search for only the shortest paths, or be only interested in paths in which all vertices are distinct.

Weighted graphs are graphs which associate some value—or weight—with each edge. They play an important role in many areas: static code analysis, bioinformatics, analysis of RDF files etc. One way to formulate a path constraint in a weighted graph is with a formal language. Note that a path in a weighted graph can be naturally associated with a word over the alphabet of weights. Formal language constrained path querying is a search for paths which are associated with words from the language. There are different variations of the path querying problem: to search for paths between two particular vertices, for all paths which satisfy constraints or simply

conduct a reachability check. Depending on the particular problem, one needs to employ different algorithms to get efficient solutions.

It is natural to employ formal language theory for formal language constrained path querying. On the one hand, there are fundamental questions about the decidability of the path querying problem. It is an open question which language classes make the problem decidable. On the other hand, it is possible to utilize parsing algorithms for path querying. The algorithms themselves should be modified, and then their theoretic properties are to be determined, including their computational complexity. It is important, that the answers to these questions depend not only on the language class, but the properties of the input graphs which leads to close interaction of the formal language and graph theory. Many questions are still open, even though the constrained path querying has been studied since the early 1990s (Thomas Reps and Mihalis Yannakakis). It is still unknown whether a subcubic algorithm for context-path querying exists. The interest to graph databases, which has recently arose, spurred the development of the algorithms for specific real-world problems, but they are mostly not mature enough to be used in production.

Computationally efficient algorithms are crucial for applications. Real-world graphs are huge in terms of the number of vertices and edges, so it is natural to create parallel and distributed algorithms, including massively-parallel or GPGPU based algorithms. Parallel algorithms are being actively researched, but their application for formal language constrained path problems is not studied enough.

To apply formal language constraints for graph databases we should provide a way to formulate database queries as constraints. One natural way to do it is by using parser combinators. Usually, parser combinators are used to simultaneously specify the language and its parser by combining functions that implement simpler parsers. Parser combinators provide transparent integration with a host programming language (embedding of a domain-specific language to the general-purpose one is avoided), high level of abstraction by using higher-order functions, and type safety. This approach works well for programming language analysis, but its applicability to graph analysis has not been investigated enough.

It is important to provide methods to optimize both queries and query execution procedures for graph databases and their applications. One promising way to do it is to use mixed computations and specialization. This approach shows good results for relational databases (for example, consider the work of E. Sharygin), but its applicability for graph databases is not investigated.

This project is aimed to create and implement formal language constrained path querying algorithms, to provide methods for constraints specification, and to develop optimization techniques for graph database queries and query execution procedures. We plan to employ formal language theory and graph theory to determine decidable subclasses of the problem, and such subclasses for which efficient algorithms may be created. We will use methods of parallel algorithms construction, including methods for massively-parallel architectures, to develop practical efficient algorithms. Methods of programming language design and development are required to investigate practical ways to specify the constraints. Mixed computations will be used to optimize queries and query execution procedures.

The team includes experts in graph theory, formal language theory, compiler construction, program optimization methods, programming language development. It allows us to organize fruitful collaboration, and involve talented students to investigate respective areas of science and

to work on the project.

## 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 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 пространственной сложности. Также 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 сформулированы 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. Помимо того, 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, а для каких только приближённо. 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.

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 будут разработаны новые алгоритмы специализации.

en

The aim of the project is to study formal language constrained path problems and to develop practically efficient solutions for them. The plan is to obtain theoretical results at the junction of the formal language theory, graph theory, and parallel programming, as well as results in the field of language development and software optimization methods.

One of the problems is to create a more detailed classification of context-free path querying problems with respect to both language subclasses and the shape of the graph. The most important is to determine whether there exists a subcubic algorithm for the general context-free path querying problem. Since this question is hard and has been open for a long time, providing a complete answer cannot be guaranteed. However even a partial answer such as describing subclasses of the problem for which such algorithm exists will form a scientific result.

In the area of parallel computing, it is planned to develop new algorithms for context-free path querying for massively parallel and distributed systems. Fundamental theoretical properties of the developed algorithms are to be studied. Asymptotic time and space complexity is to be estimated. It is also planned to modify the algorithms to work with other language classes.

Combinatory parsing is to be employed as a way to both execute path queries and formulate constraints on paths. It is planned to study the limitations of this approach, as well as its advantages and disadvantages in the context of the real-world problems such as type-safety and semantics calculation. In spite of decades of active research of parsing combinators for programming languages analysis, its generalization for path querying is not trivial and new scientific results are expected. Parser combinators provide not only a way to solve a formal language path querying problem, but also serve as a constraints description mechanism. Using parser combinators as a language for describing constraints facilitates user adoption and provides a more transparent integration into the business logic. It is planned to determine the types of graphs for which it is possible to compute user semantics. Some languages which are not context-free may be analysed with parser combinators. Thus it is planned to investigate which not context-free language classes can be used as constraints in path querying.

For graph query execution procedure optimization we plan to develop a solution that is based on runtime specialization. New specialization algorithms may be developed in the process.

## 1.7 В состав научного коллектива будут входить

- 10 исполнителей проекта (включая руководителя)
- в том числе 10 исполнителей в возрасте до 39 лет включительно,
- из них: 7 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.

- Семён Вячеславович Григорьев, 30 лет, к.ф.-м.н., доцент СПбГУ, трудовой договор
- Даниил Андреевич Березун, 27 лет, к.ф.-м.н., научный сотрудник ООО "ИнтеллиДжей Лабс приглашённый лектор в НИУ ВШЭ
- Екатерина Андреевна Вербицкая, 26 лет, научный сотрудник ООО "ИнтеллиДжей Лабс ассистент СПбГЭТУ "ЛЭТИ гпд
- Екатерина Николаевна Шеметова, 28 лет, лаборант-исследователь СПбГУ, трудовой договор
- Рустам Шухратуллович Азимов, 24 года, научный сотрудник ООО "ИнтеллиДжей Лабс гпд
- Юлия Алексеевна Сусанина 22 года, магистрант СПбГУ, гпд
- Никита Матвеевич Мишин, 21 год, студент СПбГУ, гпд
- Арсений Константинович Терехов, 21 год, студент СПбГУ, гпд
- Илья Владимирович Балашов, 20 лет, студент СПбГУ, гпд

### 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, РФФИ 18-11-00100). Область научных интересов включает теорию формальных языков, теорию графов, алгоритмы синтаксического анализа, разработку параллельных алгоритмов.

Даниил Андреевич Березун является кандидатом физико-математических наук, преподаёт на кафедре прикладной математики и информатики НИУ ВШЭ в Санкт-Петербурге.



Опыт руководства исследовательскими работами и преподавательской деятельности составляет более 5 лет. За это время под его руководством были защищены 3 выпускных квалификационных работы бакалавра, более 6 курсовых работ. За время преподавательской деятельности занимался подготовкой и чтением курсов по компиляции, разработке языковых процессоров, метавычислениям и семантикам языков программирования. В настоящее время под его руководством работают 2 магистранта. Имеет опыт исполнения грантов (РФФИ 18-01-00380). Область научных интересов включает анализ, разработку и реализацию языков программирования, метапрограммирование и метавычисления, математическую логику, семантику языков программирования, автоматическую генерацию программ, основанную на семантике, блокчейн и распределённые технологии.

Екатерина Андреевна Вербицкая окончила аспирантуру математико-механического факультета СПбГУ по направлению информатика, преподаёт 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.

Рустам Шухратуллович Азимов является аспирантом математико-механического факультета СПбГУ по направлению информатика. Защитил магистерскую диссертацию на тему "Синтаксический анализ графов через умножение матриц". Имеет публикации по теме проекта ("Context-Free Path Querying by Matrix Multiplication "Синтаксический анализ графов с использованием конъюнктивных грамматик "Синтаксический анализ графов и задача генерации строк с ограничениями"). Имеет опыт исполнения грантов (РНФ 18-11-00100 и РФФИ 19-37-90101). Область научных интересов: теория формальных языков, запросы к графам, языки запросов, поиск путей в графах, матричные операции, параллельные алгоритмы.

Екатерина Николаевна Шеметова в 2019 году окончила 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.

Юлия Алексеевна Сусанина является магистрантом математико-механического факультета СПбГУ по направлению "Программная инженерия". Полученные за время обучения в бакалавриате результаты, связаны с исследованием матричных алгоритмов синтаксического анализа (тема ВКР бакалавра "Оптимизация алгоритмов синтаксического анализа, основанных на матричных операциях"). Результаты были приняты к публикации в журнал "Труды ИСП РАН" и представлены на международной конференции по биоинформатике CIBB 2019. Область научных интересов включает теорию формальных языков, алгоритмы

синтаксического анализа и их применения.

Никита Матвеевич Мишин является студентом математико-механического факультета СПбГУ по направлению "Программная инженерия". 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 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. Его область научных интересов включает распределенные и параллельные вычисления, формальные языки, программирование на ГПУ (GPGPU) и функциональное программирование.

Терехов Арсений является студентом 4го курса СПбГУ по направлению "Математическое обеспечение и администрирование информационных систем" а так же студентом 3го курса Computer Science Center. Прошёл летние стажировки в компаниях Яндекс и JetBrains. Принимал участие в двух проектах под руководством работников компании JetBrains. Его область научных интересов включает формальные языки и графовые базы данных.

Илья Вадимович Балашов является студентом математико-механического факультета СПбГУ по направлению "Математическое обеспечение и администрирование информационных систем". Принимал участие в летней школе Ланит-Терком, а также в проекте TRIK. Область научных интересов включает теорию формальных языков, метапрограммирование и функциональное программирование.

**en**

The lead of the group, Semyon V. Grigorev, is an associate professor of the faculty of Mathematics and Mechanics of Saint-Petersburg State University and has a Ph.D. in mathematics and physics. He has 6 years of experience in teaching and being a leader and a manager of research projects. He has supervised 7 master dissertations, 12 graduation theses of bachelors, 2 graduation theses of specialists, and more than 10 course works. Two Ph.D. students are being supervised by him now. The following courses were prepared and taught: graph theory, formal language theory, algorithms and data structures. Semyon has experience in being a leader of both grants (RFBR 19-37-90101; FASIE, 162ГУ1/2013 and 5609ГУ1/2014), and research groups and projects. Also, he has participated in grants (RFBR 15-01-05431, RFBR 18-01-00380, RSF 18-11-00100). Research interests include formal language theory, graph theory, parsing algorithms, parallel algorithms.

Daniil A. Berezun has a Ph.D. in mathematics and physics, and is a lecturer at the Applied Mathematics and Informatics chair of NRU HSE in St. Petersburg. He has supervised 3 graduation theses of bachelor and more than 6 course works. Two master students are being supervised by him now. The following courses were prepared and taught: compiler techniques, language processors development, programming languages semantics, metacomputations. Daniil has participated in the grant RFBR 18-01-00380. Research interests include analysis, design, and implementation of programming languages, programming languages semantics, metaprogramming and metacomputations, semantic-based automated program generation, blockchain, and distributed systems.

Ekaterina A. Verbitskaia finished a Ph.D. program with specialization "Informatics" at the

faculty of Mathematics and Mechanics of Saint-Petersburg state university. She is a lecturer at the Saint-Petersburg Electrotechnical University "LETI". She has 4 years of experience in teaching and being a leader of research projects. She has supervised 2 graduation theses of bachelor. Two master students are being supervised by her now. The following courses were prepared and taught: formal language theory, compiler construction. She has participated in the grant RFBR 18-01-00380. Research interests include embedded language analysis, parser combinators and parsing algorithms, functional programming, supercompilation, and partial deduction for logical programming languages.

Rustam Sh. Azimov is a Ph.D. student at the faculty of Mathematics and Mechanics at Saint-Petersburg State University. He has a masters degree, his mathers thesis is "Graph parsing by matrix multiplication". Rustam has publications which are related to this project ("Context-Free Path Querying by Matrix Multiplication "Graph parsing by using conjunctive grammars "Graph parsing and constrained string generation problem"). He has participated in grants (RSF 18-11-00100 и RFBR 19-37-90101). Research interests include formal language theory, graph querying, query languages, linear algebra, parallel algorithms.

Ekaterina N. Shemetova has a masters degree: she graduated from the masters program "Software development" in 2019. Masters thesis: "Context-free constrained path problem". She is a Ph.D. student at Saint-Petersburg Academic University, specialization is "Informatics". She has a publication related to this project ("Boolean grammar constrained path querying in direct acyclic graphs") and has participated in the grant RSF 18-11-00100. Research interests include complexity theory, formal language theory and applications, static code analysis.

Julia A. Susanina is a masters student at the faculty of Mathematics and Mechanics at Saint-Petersburg State University, specialization is "Software engineering". The bachelors thesis is related to formal languages and parallel parsing algorithms: bachelor thesis is "Optimizing of matrix-based parsing algorithms". She presented her research at CIBB-2019 conference, and the paper she coauthored was published in the Proceedings of ISP RAS. Research interests include formal language theory, parsing algorithms and applications.

Nikita M. Mishin is a 4th-year student at the faculty of Mathematics and Mechanics at Saint-Petersburg State University, specialization is "Software engineering". He has a publication which is related to this project ("Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication GRADES-2019). Nikita participated in a number of summer schools, including the summer school of Lanit-Tercom, project RuCuHmmer which is aimed to migrate biological data analysis tool HMMER to GPGPU. Research interests include parallel algorithms, formal languages, GPGPU programming, functional programming.

Arseniy K. Terekhov is a 4th-year student at the faculty of Mathematics and Mechanics at Saint-Petersburg State University, specialization is "Information System Administration and is a 3rd-year student of Computer Science Center. Arseniy was an intern at Yandex and JetBrains software development company. He worked on two research projects led by employers of JetBrains. Research interests include formal languages and graph databases.

Ilya V. Balashov is a 3rd-year student at the faculty of Mathematics and Mechanics at Saint-Petersburg State University, specialization is "Information System Administration". He participated in the summer school organized by Lanit-Tercom and is a member of TRIK project. Research interests include formal language theory, metaprogramming, and functional programming.

## **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**

14 публикаций

из них 12 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**

- Proceedings of Joint International Workshop on Graph Data Management Experiences & Systems (Grades) and Network Data Analytics (Nda), издатель ACM, Scopus, статья
- Proceedings of International Conference on Extending Database Technology (EDBT), издатель OpenProceedings.org, !!!, статья
- СЕКР?
- SEIM?
- Что-то ещё из журналов?
- Труды Института системного программирования РАН, издатель Институт Системного Программирования РАН, РИНЦ, статья

### **Иные способы обнародования результатов выполнения проекта**

- Участие в постерных сессиях при конференциях SIGMOD,
- Проведение открытых лекций !!!!

## **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 выразительностью, а значит, от используемого класса языка зависит то, насколько сложные ограничения мы сможем задать. Например язык

сбалансированных скобочных последовательностей не является регулярным, поэтому пути такого вида, возникающие в задачах статического анализа кода и задачах анализа иерархий, не получится найти при использовании регулярных языков в качестве ограничений. Но он является контекстно-свободным, а значит, используя контекстно-свободные языки, мы сможем описать требуемое ограничение. С прикладной точки зрения, от класса языков, используемых для ограничений, зависит то, насколько выразителен тот или иной язык запросов к графовой базе данных. Вместе с этим существует и другой вопрос: насколько 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, проверка наличия хотя бы одного пути в графе, удовлетворяющего заданным ограничениям, выражается как задача проверки пустоты пересечения двух языков: языка, заданного в качестве ограничений и регулярного языка, который задаётся размеченным 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, на сколько много ложно-положительных результатов).

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) — продолжаются 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 — всё ещё 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, позволяющие задавать соответствующие ограничения в прикладных задачах, и разработать техники оптимизации, позволяющие эффективно выполнять соответствующие запросы в реальных графовых базах данных.

Так, 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 об-

щего назначения, которая удачно решена для задач синтаксического анализа с применением парсер-комбинаторов. В отличие от генераторов синтаксических анализаторов, парсер-комбинаторы решают задачу синтаксического анализа и описывают язык 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 Tijs 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.

У процедуры выполнения запроса как правило два основных параметра — запрос и данные. При этом сама процедура реализована в общем виде: она должна уметь выполнить любой корректный запрос. Это приводит к тому, что в коде присутствует большое количество операций, которые могут быть лишними при выполнении какого-либо конкретного запроса. Таким образом, в тот момент, когда запрос стал известен, можно построить более специфичную процедуру и выполнять именно её. Для решения подобных задач могут применяться смешанные вычисления, в частности, специализация. Специализатор, в данном случае, может по процедуре общего вида, которая принимает два аргумента, и запросу, построить новую, оптимизированную на данный запрос, процедуру, которая будет принимать только один аргумент — данные. Сгенерированная таким образом новая процедура позволяет существенно повысить производительность исполнения запроса на произвольных конкретных данных. Данная техника активно изучается в области языков программирования с 1972 года (работы N.D.Jones, P.Sestoft, R.Gluck и др.) и лишь в 2018 году была применена Е. Шарыгиным и соавторами для оптимизации выполнения запросов в реляционной СУБД PostgreSQL (Sharygin Eugene et.al. 2018. Runtime Specialization of PostgreSQL Query Executor). Так как графовые базы данных, языки запросов к ним и процедуры выполнения запросов существенно отличаются от реляционных, то применимость данной техники для оптимизации процедур выполнения запросов к графовым базам данных является открытым нетривиальным вопросом.

en

The project aims at researching the formal language constrained path querying problem in order to develop efficient solutions with respect to different language subclasses and types of graphs.

The expressive power of a language subclass defines the complexity of the constraints on paths that one can use. For example, the language of balanced brackets is not regular. Constraints in many problems of static analysis or hierarchical analysis can only be expressed as a balanced bracket language, thus limiting the use of regular constrained path querying. These kind of problems can be solved by context-free path querying. From a practical standpoint, the expressive power of a query language is determined by the language class used. There is a more general question: how expressive a query language can be? To answer this question, one needs to work at the junction of the formal language theory and the graph theory. The simplest case is when the problem is to determine the existence of a path in the graph which satisfies the constraints. This problem can be viewed as a check if the intersection of a constraint language and a regular language constructed from the graph is not empty. A weighted graph in this case is viewed as a finite automaton every state of which is both start and terminal. It is known that there are some useful in practice constraint language classes for which emptiness is undecidable: for example conjunctive languages introduced by Alexander Okhotin. Using this language class for constraints means that the user may write a query which is impossible to execute. It is worth to note that some approximation of the query result may be useful in practice. In this case it is necessary to estimate the quality of the result: how many false-positives and false-negatives is reported.

It is still a challenge to develop efficient algorithms even for the language classes for which emptiness is decidable. The most well understood class of constraints is regular, and the search of the efficient algorithm for distributed systems continues. Maurizio Nol  and Carlo Sartiani presented an algorithm for regular path querying based on Brzozowski derivatives which can be implemented using vertex-level parallelism (Maurizio Nol  and Carlo Sartiani, Regular Path Queries on Massive Graphs, 2016). The existence of a subcubic algorithm for context-free language remains an open problem. Jochem Kuijpers et al. implemented existing algorithms for the Neo4j database and concluded that they are not production-ready, thus the development of more efficient algorithms should be continued (Jochem Kuijpers, George Fletcher, Nikolay Yakovets, and Tobias Lindaaker, An Experimental Study of Context-Free Path Query Evaluation Methods, 2019).

Besides theoretical base and efficient algorithms, it is also important to develop a good way to formulate the constraints in real-world systems as well as develop optimization techniques to execute queries efficiently on the real graph databases.

Graph analysis is rarely an isolated task. As a rule, it is necessary to integrate it into some application written in a general purpose language. This is where the problem of transparent integration of the constraints specification language into a general purpose language arises. In the area of program analysis it is solved with combinatory parsing. Parser combinators serve as both a parser and a description of an object language. In contrast to parser generators, no specific DSL for the language description is needed. There are many benefits in writing the query in a general purpose language. First of all, it is possible to parameterize and reuse sub-queries which is considered a good practice in software development. Second of all, writing queries is less error-prone this way since the queries are statically analysed by the well-developed tools for the host language. In spite of the decades of studying parser combinators, some problems have been solved not so long ago even for parsing of symbolic strings. Anastasia Izmaylova et al. presented a library of parser combinators capable of processing arbitrary context-free grammars in 2016 (Anastasia Izmaylova, Ali Afroozeh, and Tijs van der Storm. 2016. Practical, general parser combinators). This work lifted the long standing limitation of parser combinators: inability to process left-recursive



or ambiguous grammars. One of the advantages of parser combinators is the ability to compute user semantics, but the language class for which it is possible to compute the exact semantics is unknown. It is also an open question whether parser combinators can express the path queries which are not context-free. Development of the query specification language which is as expressive as possible without loss of semantics precision is also a nontrivial problem.

As a rule, a query execution procedure has two parameters: a query and data. The execution procedure itself should be able to process any correct query. This leads to code containing many operations which are unnecessary for any specific query. It is possible to construct a more specific version of the query at the moment when the query becomes known which may improve the query execution performance. Mixed computations and, in particular, specialization is a way to perform such transformation. Specializer, given a general query execution procedure with two parameters, constructs a new, optimized for the given query, procedure with only one parameter (the data). The new procedure is more performant and query execution time will be better for arbitrary graphs. This technique has been actively developed since 1972 (N.D.Jones, P.Sestoft, R.Gluck etc.), but only in 2018 it has been successfully applied to optimize query execution procedures in relational database system PostgreSQL by E. Sharygin and coauthors (Sharygin Eugene et.al. 2018. Runtime Specialization of PostgreSQL Query Executor). Graph databases, graph query languages, and respective query execution procedures are significantly different from relational databases, so the applicability of specialization and related techniques for graph query execution procedures is a nontrivial open question.

## 2.2 Научная значимость и актуальность решения обозначенной проблемы

ru

Знание границ разрешимости задачи необходимо для разработки языков запросов и 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 решения 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 "межъязыкового" вывода типов.

Метавычисления, специализация и смешанные вычисления изучаются давно (работы N.D.Jones, В.Ф.Турчина, А.П.Ершова в 70–90-е года, а также работы их учеников, последователей и соавторов), но до сих пор в этой области много открытых вопросов как в теории, так и относительно применимости в прикладных задачах. Так, только в 2018 году специализация позволила существенно ускорить выполнение запросов в реляционной СУБД PostgreSQL (Sharygin Eugene et.al. 2018. Runtime Specialization of PostgreSQL Query Executor), а в 2019 было показано, что с помощью суперкомпиляции возможно построить процедуру выполнения SQL-запросов, превосходящую по производительности многие аналоги (Tiark Rumpf, Nada Amin. 2019. A SQL to C compiler in 500 lines of code). Производительность процедуры выполнения запросов в графовых базах данных важна с прикладной точки зрения, однако применимость данных подходов для ускорения исполнения запросов в графовых базах данных не изучалась. Вместе с тем, исследование данной области может привести к новым теоретическим задачам в области смешанных вычислений.

**en**

Decidability of formal language constrained path problem is important for the development of graph querying languages and also in determining the decidability of the applications which can

be reduced to this problem. Note that in practical applications, when the problem is undecidable, it may be important to provide an appropriate approximation, if possible. For example in static code analysis we expect the result to be empty, since the program is expected to not contain undesirable behaviours. Thus, if the over-approximation of the result turns out to be empty, then the precise solution is empty too, which renders over-approximating the result a good idea in this application area. But it would be better to provide an approximation which is as precise as possible, because otherwise users will face too many false-positives (undesirable behavior reported when in reality program is unable to have it) and will not be able to use the technology for their benefit. The example of such an approach is a work of Q. Zhang, who uses linear conjunctive languages for static code analysis (Qirun Zhang and Zhendong Su, Context-sensitive data-dependence analysis via linear conjunctive language reachability, 2017). Linear conjunctive constrained reachability is undecidable, but it is shown that it is possible to provide a reasonable practical approximated solutions.

Theoretical properties of algorithms are important both as a self-contained theoretical result and as a means to create efficient practical solutions. Note that despite long research history, some new results were provided in the last few years. For example, in 2019, Ph. Bradford provides a subcubic algorithm for the 1-Dyck reachability problem (Phillip G. Bradford, Efficient Exact Paths For Dyck and semi-Dyck Labeled Path Reachability). The proposed solution cannot be generalized to arbitrary context-free languages, so more research in this direction is required. In 2017, K. Chatterjee provides an optimal algorithm for (arbitrary) Dyck reachability in a specific type of graph—bidirected graph—(Krishnendu Chatterjee, Optimal Dyck reachability for data-dependence and alias analysis). He also shows that the algorithm can be efficiently applied for static code analysis.

The development of computationally efficient algorithms, including algorithms for massively-parallel and distributed systems, is important for the investigation of the theoretical properties of algorithms and the advancing of parallel and distributed system theory. It also enables the creation of efficient applied solutions, for example for graph databases, which are becoming more popular in the recent years. As we note above, efficient algorithms development is a significant problem even for well-investigated classes of formal language constrained path problem (for example, see works of Jochem Kuijpers and Maurizio Nol ).

Query language development is another problem worth attention. To answer what makes a good query language, we need to investigate ways to specify constraints. Graph database query languages are at the early stage, so there is a lot of open questions about their syntax and semantics. There is also a number of open questions about the integration of domain-specific languages into general-purpose programming languages. They deal with transparent integration, type safety, compile-time checking of correctness. Several solutions have been proposed to solve these problems (language integrated query, combinators, interlanguage type inference, etc.) and new solutions are still under development.

Partial evaluation and supercompilation have been researched since 1970s (N.D. Jones, V.F. Turchin, A.P. Ershov and their followers), but there are still many open questions both theoretical and about the applicability of these methods to real-world problems. For example, only in 2018 performance SQL query execution procedure in PostgreSQL DBMS was significantly improved by means of specialization (Sharygin Eugene et.al. 2018. Runtime Specialization of PostgreSQL Query

Executor). In 2019 it was shown that it is possible to create a SQL query execution procedure which outperforms competitors by means of supercompilation (Tiark Ropmf, Nada Amin. 2019. A SQL to C compiler in 500 lines of code). Performance of query execution in graph databases is crucial for applications, but there are no results on the application of partial evaluation, supercompilation and similar to graph databases. At the same time, investigating this area can lead to new theoretical problems and tasks in the area of mixed computations and partial evaluation.

## **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 и сведение 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 также с изучением ограничений, которые возникают при их использовании. В частности, планируется изучить ограничения, накладываемые на семантические функции, и их связь с потенциальной бесконечностью множества путей. Также планируется экспериментальное исследование механизма интеграции запросов в код на языках общего на-

значения, 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 от 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 изучить применимость существующих техник специализации и супекомпиляции для оптимизации процедуры выполнения запросов. Возможно, будут разрабатываться новые методы специализации.

**en**

As a part of investigating of decidibility of formal language constraint path problems and the properties of the algorithms for them, it is planned to study new subclasses of the problem for various language classes and types of graphs. First of all, it is planned to study different subclasses of context-free languages. Here the aim is to get closer to answering the question about the existence of subcubic algorithm and also to find language subclasses expressive enough to be used in the real-world applications for which one can develop efficient algorithms. Moreover, it is planned to study different formulations of the problem: single path search, shortest path search. Algorithms developed for these different formulations may vary in their formal properties such as time and space complexity. Finally, it is planned to study languages which are not context-free from the standpoint of their applicability as constraint languages.

In the area of parallel and distributed algorithms, it is planned to create algorithms for formal language constraint path querying which take advantage of different types of parallel architectures: massively-parallel architectures (SIMD), multithreading, multicore. Here the focus lays on context-free constraints. It is planned to consider different approaches and models of parallel programming such as vertex-level parallelism and reduction to the problem with known efficient algorithms. As a part of this work, different data representations are to be compared. Despite active development of graph databases and the underlying theory, there exists no consensus on the best graph representation. It is related to the fact that the specific constraints of the problem are not studied well.

Concerning the constraints language, it is planned to study the applicability and limitations of parser combinators. In particular, it is planned to study the constraints for the semantic functions in the context of potentially infinite number of paths. It is also planned to conduct an experimental study of the ways to integrate queries into a host language based on parser combinators. It is planned to compare it with the existing approaches in such aspects as expressive power, modularity and safety.

There are two directions to optimize queries in runtime. First one is the optimization of the

queries themselves. It is known that a context-free language can be described with numerous grammars. The execution time of parsing of programming languages depends on the properties of the particular grammar (having the tool and input fixed). The same behavior is observed in path querying, but it is not always the case that the best grammar for parsing is good for path querying. It is planned to study ways to optimize queries, implement the corresponding algorithms, and run an experimental study. The second direction is compile-time optimization of algorithms implemented. Here it is planned to study the applicability of the existing specialization and supercompilation techniques and it is possible that it will be necessary to develop new specialisation methods.

## **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 вопро-

сов, 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

The problems listed here are of interest to researchers: a number of papers is published in the last few years in peer-reviewed journals and presented at the high-rated conferences. Some of these papers are written by the members of this project. This guarantees the attention of the scientific community to the results and their good quality.

Since some of the problems are nontrivial, we cannot guarantee that a complete solution will be found during this project. The example of a nontrivial problem is determining the existence of a subcubic algorithm for context-free reachability. But even finding a partial solution or improving the existing results (for example, an adaptation of Bradford’s solution for a wider class of languages) will be a significant contribution. Note, that we also plan to work on problems which are likely to be solved completely.

The experience of project members in such areas as formal language theory, graph theory, parsing algorithms, allows us to investigate subclasses of formal languages constrained path problem comprehensively. Both negative and positive results in this area are important: finding subclasses for which efficient algorithms exist as well as proving that there are not such algorithms for particular subclasses.

In 2019 Jochem Kuijpers et al showed that no computationally efficient and applicable for real-world problems algorithm for context-free path querying currently exists, despite the algorithms for context-free path querying being under active development in the last few years. Members of the team (R. Azimov, S. Grigorev, E. Verbitskaia) have strong experience in developing such algorithms. For example, R. Azimov proposed the matrix-based algorithm for context-free path querying which utilizes parallel hardware. This allows us to develop new algorithms, investigate their formal properties, and provide comprehensive evaluation and comparison of them.

Integration of the query language base on parser combinators will be based on the results of D. Kröni and on the experience of E. Verbitskaia who worked on parser combinators for both linear and graph-structured input and also developed a context-free path querying algorithm which is based on a bottom-up parsing algorithm. Note that D. Kröni does not discuss a class of languages which can be used as constraints and other important details.

We plan that we will use the results of E. Sharigin who shows that specialization can be used for query execution procedure optimization in relational databases. Specialization has not been yet applied for graph database, so it may be necessary to develop new algorithms or significantly improve the existing solutions. The experience of E. Verbitskaia and D. Berezun in partial evaluation and specialization is expected to be instrumental in solving this problem.

## 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 (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 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 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 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, Ilya Kirillov, Ilya 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 У. Бердж (William Burge) в 1975 году, а в 1990 Ф. Вадлер (Philip Wadler) предложил идею монадических комбинаторов, леворекурсивные спецификации и неоднозначности вывода до недавнего времени представляли проблему. В 2016 году она была решена Анастасией Измайловой и Али Афрузехом (Anastasia Izmaylova, Ali Afroozeh, and Tijs van der Storm, 2016, Practical, general parser combinators).

Разработкой метавычислителей занималось множество исследовательских групп. Так например, группа из университета Копенгагена под руководством N.D.Jones и его коллег (С.К. Gomard, P.Sestoft, O.Danvy, Т.Æ.Mogensen, R.Glück и другие) занималась изучением специализации программ, реализацией самоприменимых специализаторов для различных языков программирования, вопросами сложности и производительности специализаторов. Классическим трудом по статической специализации программ является работа N.D.Jones, С.К.Gomard, P.Sestoft, L.O.Andersen, Т.Mogensen, Partial evaluation and automatic program generation, 1993. Группы под руководством В.Ф.Турчина занимались изучением автоматического преобразования программ посредством суперкомпиляции. Создание применимых на практике решений,

основанных на данных методах оптимизации программ является активно исследуемой областью. Так, например, И.Г.Ключников и С.А.Романенко в 2009 году сумели применить идеи суперкомпиляции для функций высших порядков (И.Г.Ключников. Суперкомпиляция функций высших порядков. 2010), а затем предложили многорезультатную и многоуровневую суперкомпиляцию (I.Klyuchnikov, S.A.Romanenko. Multi-result supercompilation as branching growth of the penultimate level in metasystem transitions. 2011; I.Klyuchnikov, S.A.Romanenko. Higher-level supercompilation as a metasystem transition. 2012) позволяющие ещё лучше оптимизировать программы. V.Srinivasan и T.Reps применили идеи специализации для оптимизации машинного кода (V.Srinivasan, T.Reps. Partial Evaluation of Machine Code. 2015). В 2000 году М. Спербер применил смешанные вычисления для построения синтаксических анализаторов (Michael Sperber, Peter Thiemann. 2000. Generation of LR parsers by partial evaluation). Применительно к оптимизации процедур выполнения запросов, наиболее существенные прикладные результаты принадлежат Т. Ромпфу (Tiark Rompf, Nada Amin. 2015. Functional pearl: a SQL to C compiler in 500 lines of code) и Е. Шарыгину (Sharygin E., Buchatskiy R., Zhuykov R., Sher A. 2018. Runtime Specialization of PostgreSQL Query).

en

Graph-structured data processing and particularly graph database querying are actively researched in the community. Both theoretical and practical aspects of the formal language constrained path problem are investigated.

The context-free constrained path querying has been under active development since it was introduced in two different areas: by T. Reps in the area of static code analysis (T. Reps, 1997, Program analysis via graph reachability) and by M. Yannakakis as applied to graph databases (M. Yannakakis, 1990, Graph-theoretic methods in database theory). The general problem of formal language constrained path querying is also investigated (C. Barrett, R. Jacob, and M. Marathe, 2000, Formal-language-constrained path problems).

One research direction is investigating which new practical problems can be formulated in terms of context-free constrained path queries. These problems include biological data analysis (P. Sevon and L. Eronen, 2008, Subgraph queries by context-free grammars), RDF and ontology analysis (C. M. Medeiros, M. A. Musicante, and U. S. Costa, 2019, LL-based query execution over RDF databases и X. Zhang, Z. Feng, X. Wang, G. Rao, and W. Wu, 2016, Context-free path queries on RDF graphs), inference of program code specifications (Osbert Bastani, Saswat Anand, and Alex Aiken, 2015, Specification Inference Using Context-Free Language Reachability), alias analysis in program code (Dacong Yan, Guoqing Xu, and Atanas Rountev, 2011, Demand-driven context-sensitive alias analysis for Java) and others. Moreover, more expressive classes of languages, such as linear conjunctive, can be used for static code analysis (Qirun Zhang and Zhendong Su, 2017, Context-sensitive data-dependence analysis via linear conjunctive language reachability).

At the same time, theoretical aspects of efficient algorithms creation for different classes of the problem are investigated. One fundamental result was presented by L. Valiant: it is possible to create a subcubic algorithm for context-free parsing (L. G. Valiant, 1975, General context-free recognition in less than cubic time). One of the most important open questions is whether it is possible to generalize Valiant's result for context-free path querying. Recently, several results

in this direction have been published. In 2017 K. Chatterjee proposed an optimal algorithm for Dyck-constrained path querying in a special type of graphs—bidirected graphs—(Krishnendu Chatterjee, Bhavya Choudhary, and Andreas Pavlogiannis, 2017, Optimal Dyck reachability for data-dependence and alias analysis). Ph. Bradford in 2017 provides a subcubic algorithm for 1-Dyck (Dyck language over one type of braces) constrained path querying for arbitrary graphs (Ph. G. Bradford, 2017, Efficient exact paths for Dyck and semi-Dyck labeled path reachability). Also, J. Hellings provides some theoretical results in this area (J. Hellings, 2015, Path results for context-free grammar queries on graphs and other papers in 2014-2015).

Several research groups actively work on context-free path querying algorithms research and development. For example, a group led by Ph. Bradford from the University of Connecticut, USA (P. G. Bradford and V. Choppella, 2016, Fast point-to-point Dyck constrained shortest paths on a dag), group led by M. Musicante, Universidade Federal do Rio Grande do Norte, Brasil (Fred C. SantosUmberto S. CostaMartin A. Musicante, 2018, A Bottom-Up Algorithm for Answering Context-Free Path Queries in Graph Databases), group led by G. Fletcher, Technische Universiteit Eindhoven, Netherland. Note. The group led by G. Fletcher concluded in their study of 2019 that the existing context-free path querying algorithms are not applicable for real-world problems. At the same time, they note that the development of such algorithms is an important task (Jochem Kuijpers, George Fletcher, Nikolay Yakovets, and Tobias Lindaaker, 2019, An Experimental Study of Context-Free Path Query Evaluation Methods).

A big international research team which includes T. Lindaaker and G. Fletcher are working on languages which support formal language constraints (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). However, integration of such languages into general-purpose programming language is not investigated enough. Parser combinators based approach to solve the integration problem is discussed in papers of D. Kröni (Daniel Kröni and Raphael Schweizer, 2013, Parsing graphs: applying parser combinators to graph traversals) and E. Verbitskaia (Ekaterina Verbitskaia, Ilya Kirillov, Ilya Nozkin, and Semyon Grigorev, 2018, Parser combinators for context-free path querying). This topic is still at the early stage of research. Despite the fact that W. Burge proposed to utilize higher-order functions for passing in 1975, and Ph. Wadler proposed monadic combinators in 1990, left-recursive specifications and ambiguities were a problem until recently. This problem was solved in 2016 by A. Izmaylova and A. Afroozeh (Anastasia Izmaylova, Ali Afroozeh, and Tijs van der Storm, 2016, Practical, general parser combinators).

A number of groups contribute to the research of partial evaluation. For example, a group led by N.D.Jones from Copenhagen university (C.K. Gomard, P.Sestoft, O.Danvy, T.Æ.Mogensen, R.Glück) workes on program specialization, implementation of self-applicable specializers for different programming languages, and estimations of time complexity and performance of specializers. One of the classical papers in this area is the paper of N.D.Jones, C.K.Gomard, P.Sestoft, L.O.Andersen, T.Mogensen, Partial evaluation and automatic program generation, 1993. A group led by V.F. Turchin worked on automatic program transformation by means of supercompilation. The creation of tools which are based on these techniques and can be used for real-world problems is an actively developing area. For example, in 2009 I.G. Klyuchnikov and S.A. Romanenko showed that supercompilation can be applied for higher-order functions (I. Klyuchnikov, Higher-Order Supercompilation, 2010), and after that, they propose multi-result and multilevel supercompilation

(I.Klyuchnikov, S.A.Romanenko. Multi-result supercompilation as branching growth of the penultimate level in metasystem transitions. 2011; I.Klyuchnikov, S.A.Romanenko. Higher-level supercompilation as a metasystem transition. 2012). V.Srinivasan and T.Reps apply specialization for machine code optimization (V.Srinivasan, T.Reps. Partial Evaluation of Machine Code. 2015). In 2000 M. Sperber applied partial evaluation for parser generation (Michael Sperber, Peter Thiemann. 2000. Generation of LR parsers by partial evaluation). In the applications for databases, the most important results are by Tiark Rompf (Tiark Rompf, Nada Amin. 2015. Functional pearl: a SQL to C compiler in 500 lines of code) and by E. Sharygin (Sharygin E., Buchatskiy R., Zhuykov R., Sher A. 2018. Runtime Specialization of PostgreSQL Query).

## 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 результатов здесь — нижние оценки вычислительной сложности для алгоритмов, решающих соответствующие типы 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 планируется 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.

Для оптимизации процедур выполнения запросов будут использованы методы оптимизации программ. В частности будут использованы смешанные вычисления, специализация, суперкомпиляция. Будут использоваться как методы статической оптимизации, так и оптимизации времени выполнения. Планируется изучить применимость существующих методов в контексте графовых баз данных, привести их экспериментальное исследование. Ожидается, что в ходе этих работ будут сформулированы новые задачи, которые будут решаться в рамках данного исследования.

\*\*\* 2020 \*\*\*

Конструирование матричных алгоритмов поиска путей с контекстно-свободными ограничениями, попытки улучшить асимптотические оценки их временной сложности, изучение возможности построения таких алгоритмов для массово-параллельных и распределённых систем (С.В. Григорьев, Н.М. Мишин).

Построение алгоритма поиска путей с контекстно-свободными ограничениями, основанного на пересечении конечных автоматов, исследование его теоретических свойств (Р.Ш. Азимов).

Изучение применимости парсер-комбинаторов для анализа графов, построение прототипов решений, использующих парсер-комбинаторы для поиска путей с контекстно-свободными ограничениями, проведение их экспериментальных исследований (Е.А. Вербицкая).

Изучение применимости существующих техник специализации, в частности, результатов Е. Шарыгина, для оптимизации процедур поиска путей с контекстно-свободными ограничениями, для алгоритмов, реализуемых на центральном процессоре, создание прототипов, проведение экспериментальных исследований (Д.А. Березун, И.В. Балашов).

Построение алгоритма поиска путей с контекстно-свободными ограничениями, основанного на решении (системы) полиномиальных уравнений, исследование его теоретических свойств, создание прототипа, его экспериментальное исследование (Ю.А. Сусанина).

Изучение частных случаев задачи поиска путей с ограничениями в терминах формальных языков, для которых возможно построение эффективных алгоритмов. При обнаружении соответствующих классов задач, построение соответствующих алгоритмов и изучение их теоретических свойств (Е.Н. Шеметова).

\*\*\* 2021 \*\*\*

Поиск алгоритмов оптимизации запросов, содержащих ограничения в терминах формальных языков, изучение их теоретических свойств, проведение экспериментальных исследований (С.В. Григорьев, Р.Ш. Азимов).

Реализация прототипа алгоритма поиска путей с контекстно-свободными ограничениями, основанного на пересечении конечных автоматов, его экспериментальное исследование (Р.Ш. Азимов).

Изучение ограничений на пользовательские семантические действия при использовании парсер-комбинаторов для поиска путей с контекстно-свободными ограничениями (Е.А. Вербицкая).

Изучение применимости существующих техник специализации, в частности, результатов Е. Шарыгина, для оптимизации процедур поиска путей с контекстно-свободными ограничениями, для алгоритмов, использующих графические сопроцессоры (GPGPU), создание прототипов, проведение экспериментальных исследований. Также будет проводиться анализ результатов, полученных в данной области в предшествующем году, и на основании анализа формулирование новых направлений и конкретных задач (Д.А. Березун).

Поиск подкласса (систем) полиномиальных уравнений, задача решения которых сводится к задаче поиска путей с контекстно-свободными ограничениями (Ю. А. Сусанина, С.В. Григорьев).

Поиск подклассов языков, для которых возможно построение субкубического алгоритма поиска путей, попытки обобщить результаты Ф. Брэдфорда (Е.Н. Шеметова, С.В. Григорьев).

\*\*\* 2022 \*\*\*

Исследование различных структур для представления разреженных матриц и их применимости в матричных алгоритмах поиска путей с контекстно-свободными ограничениями. В частности, планируется исследование Quad-tree представления (С.В. Григорьев)

Реализация прототипа алгоритма поиска путей с контекстно-свободными ограничениями, основанного на производных Бжзовского, его экспериментальное исследование (Р.Ш. Азимов).

Изучение возможности использования парсер-комбинаторов для задания более чем контекстно-свободных ограничений (Е.А. Вербицкая).

Изучение применимости существующих техник специализации, для оптимизации процедур, активно использующих операции линейной алгебры, в случае необходимости, разработка новых техник и алгоритмов. Решение задач, поставленных в предшествующем году (Д.А. Березун).

Попытка построить взаимное сведение между задачами (или соответствующими подзадачами) решения (систем) полиномиальных уравнений и поиска путей с контекстно-свободными ограничениями (Ю. А. Сусанина, С.В. Григорьев).

Поиск взаимной сводимости между задачами Language Editing Distance (LED) и задачей достижимости с контекстно-свободными ограничениями с целью отыскать пути построения субкубического алгоритма для задачи достижимости с контекстно-свободными ограничениями в общем виде (Е.Н. Шеметова, С.В. Григорьев).

**en**

We plan to use methods of formal language theory, graph theory, and parsing algorithms to find subclasses of language constrained path problems for which efficient algorithms exist. We plan to investigate different combinations of graph-theoretic subproblems (single path or all paths problem, single-source problems, etc.), different subclasses of context-free languages (for example,

linear context-free, one counter languages), different types of graphs (directed acyclic graphs, trees, general graphs). Expected results here are the following: lower bounds of computational complexity for algorithms for respective subclasses of the problem, develop algorithms for practical cases, classify the problem subclasses to computational complexity classes.

Also, we plan to investigate results on subcubic algorithms in related areas, such as language editing distance and shortest path problems for different graph types. Using these results we want to extend the results of L. Valiant and Ph. Bradford to arbitrary graphs and arbitrary context-free languages.

For computationally efficient algorithms development we plan to use results from parsing algorithms. One of the main directions is to adopt parsing algorithms for context-free path querying. Namely, we plan to adopt a solution for regular queries, based on Brzozowski derivatives, to context-free queries. We hope that this way can lead to efficient parallel solution because the original solution for regular queries is efficient in the vertex-level parallelism model, and Brzozowski derivatives can be generalized from regular languages to context-free ones. A promising direction is to formulate graph querying problems in terms of linear algebra, thus we plan to use linear algebra to build efficient algorithms. We plan to use algorithmic information theory and complexity theory to determine the theoretical properties of the algorithms constructed.

We plan to employ methods and results of functional programming, type theory, and formal language theory to investigate the applicability of parser combinators to graph querying. We will use the lattice theory and theory of static code analysis to formulate restrictions for semantic functions and graphs. We plan to use software engineering methods to develop an abstraction mechanism for extensible and reusable queries specification.

Program optimization methods will be used to optimize query execution procedure. Namely, we plan to use partial evaluation, specialization, and supercompilation. We will use both compile-time and run-time optimizations. We plan to investigate the applicability of existing methods for graph database query engine optimization and evaluate them. We expect that new tasks and problems will arise during this particular subtask, and these problems will be investigated in this project.

\*\*\*2020\*\*\*

Development of matrix-based algorithms for context-free path querying, attempting to improve estimations of time complexity for them, investigating the ability to develop massively-parallel and distributed versions of such algorithms (S.V. Grigorev, N.M. Mishin).

Creation of the algorithm for context-free path querying which is based on the intersection of finite automata, investigation of its theoretical properties (R. Sh. Azimov).

Investigation of the applicability of parser combinators to graph querying, development of the prototype solution for context-free path querying by means of parser combinators, and its evaluation (E.V. Verbitskaia).

Investigation of the applicability of specialization, namely results of E. Sharigin, for context-free path query execution procedures and for CPU-based algorithms, development of prototype solutions and their evaluation (D.A. Berezun, I.V. Balashov).

Creation of the context-free path querying algorithm which is based on solving of polynomial equations (system of equations), and investigation of its theoretical properties, development of prototypes and evaluation (J.A. Susanina).

Investigation of subcases of formal language constrained path problem for which it is possible to create efficient algorithms. If such cases are detected, the respective algorithms will be developed and their theoretical properties investigated (E.N. Shemetova).

\*\*\*2021\*\*\*

Development of algorithms for language constrained path queries optimization, investigation of their theoretical properties, and evaluation (S.V. Grigorev, R.Sh. Azimov).

Implementation of context-free path querying algorithm which is based on the intersection of finite automata and its evaluation (R.Sh. Azimov).

Investigation of restrictions on user-defined semantic actions for parser combinators as applied to graph querying (E.A. Verbitskaia).

Investigation of applicability of existing specialization techniques, including results of E. Sharigin, to optimization of context-free path query execution procedure for algorithms that run on GPGPUs. Development and evaluation of prototypes. Also results of the previous year will be analyzed and new tasks and problems will be formulated (D.A. Berezun).

Determining the subclasses of (a system of) polynomial equations, such that solving these equations can be reduced to context-free path querying (J.A. Susanina, S.V. Grigorev).

Determining the subclasses of context-free languages such that it is possible to create a subcubic algorithm for respective path querying problem. Attempting to generalize results of Ph. Bradford (E.N. Shemetova, S.V. Grigorev).

\*\*\*2022\*\*\*

Investigation of data structures for sparse matrices representation and their applicability for matrix-based algorithms for context-free path querying. Particularly, the investigation of Quad-tree (S.V. Grigorev).

Development and evaluation of a context-free path querying algorithm which is based on Brzozowski derivatives (R.Sh. Azimov).

Investigation of parser combinators applicability for more than context-free constraints specification (E.A. Verbitskaia).

Investigation of applicability of the existing specialization techniques to optimization of programs which are based on linear algebra operations. Development of new algorithms and methods, if necessary. Working on tasks formulated in the previous year (D.A. Berezun).

Attempting to determine bidirectional reduction between solving (a system of) polynomial equations and context-free path querying (or between respective subclasses of these problems) (J.A. Susanina, S.V. Grigorev).

Determining the bidirected reduction between Language Editing Distance and Context-Free Path Querying in order to create a subcubic algorithm for the general case of context-free path querying (E.N. Shemetova, S.V. Grigorev).



## 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;
- Shemetova, Grigorev, "Path querying on acyclic graphs using Boolean grammars" Proceedings of the Institute for System Programming, 2019;
- 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.А. Березуном работе над проектом в рамках гранта РФФИ 18-01-00380. Также, С.В. Григорьев являлся исполнителем грантов РФФИ 15-01-05431 и Фонда содействия развитию малых форм предприятий 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 (Verbitskaia E., Grigorev S., Avdyukhin D. 2016. Relaxed Parsing of Regular Approximations of String-Embedded Languages). Также начато изучение применимости парсер-комбинаторов для анализа графов. Кроме этого Е.А. Вербицкая предложила механизм поддержки левой рекурсии в библиотеке парсер-комбинаторов Ostar.

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 (Rustam Azimov and Semyon Grigorev. 2018. Context-free path querying by matrix multiplication). 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.

Е.Н. Шеметова имеет опыт исследований задач поиска путей с ограничениями в терминах формальных языков и ограничений. В частности, она провела исследование данной задачи для ациклических графов и булевых грамматик, опубликованное в 2019 году в работе "Path querying on acyclic graphs using Boolean grammars".

Д.А. Березун имеет опыт исследований в области семантики языков программирования, метавычислений и программной специализации. В частности, им предложен алгоритм компиляции нетипизированного лямбда исчисления в низкоуровневое представление посредством игровой семантики программ и частичных вычислений (D.Berezun, N.D.Jones. Compiling untyped lambda calculus to lower-level code by game semantics and partial evaluation. 2017; D.Berezun, N.D.Jones. Working Notes: Compiling ULC to Lower-level Code by Game Semantics and Partial Evaluation. 2016). Кроме того, им была показана корректность предложенного алгоритма и его обобщения, а также предложено обобщение понятия головной линейной редукции термов (Д.Березун. Полная головная линейная редукция. 2017).

Кроме этого, участники проекта создали набор данных, необходимый для экспериментального исследования разрабатываемых решений. Он представлен и используется в работе "Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication". В ходе исследований планируется его расширение.

## **2.8 Перечень оборудования, материалов, информационных и других ресурсов, имеющихся 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. Распределение задач между основными исполнителями проекта приведено в следующем разделе.

Также на первый год планируется 5 поездок с докладами на международные конференции (в среднем по 100000 рублей).

**en**

During the first year, it is planned to work on research questions listed in this plan, to present results at conferences, and prepare results for publication. Also it is planned to collaborate for

understanding the new results. As a result, some new problems will be formulated. Detailed plan for each team member is presented below.

Also, 5 trips to international conferences are planned (on average, 100000 rub. each) in order to give talks.

## **2.10 Планируемое на первый год содержание работы каждого основного исполнителя проекта (включая руководителя проекта)**

**ru**

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, и рассматриваться различные подходы к построению параллельных алгоритмов. Планируется 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 исследовать теоретические свойства полученного алгоритма.

Изучением применимости парсер-комбинаторов для анализа графов займётся Е. А. Вербицкая. А именно, будет вестись работа по изучению основных сценариев анализа графов, в которых применение комбинаторов может оказаться востребованным. Также будет вестись работа над прототипом, демонстрирующим эти сценарии, и его экспериментальным исследованием.

Адаптацией результатов Е. Шарыгина для алгоритмов выполнения запросов в графовых базах данных будет заниматься Д.А. Березун. Необходимо будет исследовать возможности такой адаптации и провести экспериментальное исследование решения, полученного в результате.

Е. Н. Шеметова будет заниматься разработкой 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 займется 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 засчет применения известных численных методов для нахождения корней уравнений (например, метод Ньютона).

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

## **2.11 Ожидаемые в конце первого года конкретные научные результаты**

**ru**

Будет описан алгоритм поиска путей с контекстно-свободными ограничениями, основанный на пересечении рекурсивного автомата и графа. Будут изучены его теоретические свойства: доказана корректность и получена оценка временной и пространственной сложности. По итогам, одна работа будет представлена на конференции. Результаты будут опубликованы в сборнике докладов, индексируемом в scopus. Начнётся работа над журнальной статьёй по этим результатам.

Будет проведено экспериментальное исследование матричного алгоритма, использующего массово-параллельные архитектуры (GPGPU). Результаты данного исследования будут представлены на конференции и опубликованы в сборнике материалов, индексируемом в scopus.

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. Результаты будут представлены на конференции и опубликованы в сборнике материалов, индексируемом в scopus.

Над прочими заявленными темами будет вестись работа, однако результаты будут опубликованы во второй год проекта.

**en**

Context-free path querying algorithm based on the intersection of a recursive state machine and a graph will be described. Theoretical time and space complexity will be determined, algorithm correctness will be proven. Results will be presented at a conference and published in proceedings indexed in Scopus. A full-length journal article on these results will be prepared for publication.

Implementation of matrix-based context-free path querying algorithm which utilize massively parallel hardware (GPGPU) will be evaluated. Results of evaluation will be presented at a conference and published in proceedings indexed in Scopus.

Context-free path querying algorithm based on solving of (systems of) matrix equations will be described. Theoretical properties of the algorithm will be investigated, implementation based on Newton method will be provided and evaluated. Results will be presented at a conference and published in proceedings indexed in Scopus.

The work on other topics will have a progress, but the results will be published during the second year of the project.

## **2.12    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.