Поиск путей с ограничениями в терминах формальных языков

Семён Григорьев

3 июля 2019 г.

1 Основные данные проекта

1.1 Название проекта

Поиск путей с ограничениями в терминах формальных языков

1.2 Основной код (по классификатору РФФИ)

07-365 Специализированные методы и алгоритмы обработки и анализа больших данных

1.3 Ключевые слова (указываются отдельные слова и словосочетания, наиболее полно отражающие содержание проекта: не более 15, строчными буквами, через запятые)

поиск путей в графах, теория формальных языков, контекстно-свободные грамматики, конъюнктивные грамматики, матричные операции

1.4 Аннотация проекта (кратко, в том числе – актуальность, уровень значимости и научная новизна исследования; ожидаемые результаты и их значимость; аннотация будет опубликована на сайте РФФИ, если проект получит поддержку)

Графы используются в качестве структуры данных для представления больших объемов информации в компактной и удобной для анализа форме в различных областях – биоинформатике, графовых базах данных, статическом анализе кода и др. При этом оказывается необходимо вычислять запросы к большим графам с целью выявления зависимостей между их вершинами.

Наиболее популярными являются запросы, которые используют контекстно-свободные грамматики для спецификации ограничений на пути. Кроме того, существуют конъюнктивные грамматики, образующие более широкий класс грамматик. Использование конъюнктивных грамматик в задаче поиска путей позволяет формулировать более сложные запросы к графу и решать более широкий круг задач.

Одной из самых популярных техник, используемых для увеличения производительности при работе с большими объемами данных, является использование параллельных систем. Одним из подходов в данной области, позволяющим эффективно использовать параллельные системы, является матричный подход, при котором строится матрица смежности входного графа, элементами которой являются множества нетерминалов входной грамматики. Далее вычисляется транзитивное замыкание построенной матрицы, используя правила вывода входной грамматики. В процессе вычисления активно используются операции умножения и сложения булевых матриц.

Проект посвящён исследованию новых алгоритмов поиска путей с использованием контекстно-свободных и конъюнктивных языков для слабо изученных семантик запросов. Данное исследование опирается на имеющиеся результаты, которые говорят о применимости матричного подхода в задачах поиска путей. Кроме того, исследование нацелено на улучшение существующих алгоритмов поиска путей и создание новых, позволяющих описывать более широкий класс запросов к графам за счёт комбинации используемых формальных языков и семантик запросов. Также планируется доказать теоретические свойства полученных алгоритмов. Полученные новые алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, эффективные с точки зрения параллельных систем, дадут возможность построения эффективных реализаций для анализа больших графов. Планируется провести ряд экспериментов по использованию полученных алгоритмов для решения прикладных задач и для их сравнения с аналогами.

1.5 Название проекта (на английском языке)

Path querying using formal languages

1.6 Ключевые слова (на английском языке)(приводится не более 15 слов)

path querying, formal language theory, context-free grammars, conjunctive grammars, matrix operations, CFPQ, context-free path querying

1.7 Аннотация проекта на английском языке (кратко, в том числе - актуальность, уровень фундаментальности и научная новизна; ожидаемые результаты и их значимость)

Graphs are used as a data structure to represent large volumes of information in a compact and convenient for analysis form in many areas: bioinformatics, graph databases, static code analysis,

etc. In these areas, it is necessary to evaluate queries for large graphs in order to determine the dependencies between the nodes.

The most popular are queries that use context-free grammars for path constraints on the path. In addition, there are conjunctive grammars that form a wider class of grammars. The use of conjunctive grammars in path querying allows us to formulate more complex queries to the graph and solve a wider class of problems.

One of the most popular techniques used to increase performance when working with large data is the use of parallel systems. One of the approaches in this area that makes it possible to use the parallel systems effectively is the matrix approach, in which the adjacency matrix of the input graph is built, the elements of which are sets of non-terminals of the input grammar. Next, the transitive closure of the constructed matrix is calculated using the derivation rules of the input grammar. In the process of computing, the operations of multiplication and addition of Boolean matrices are actively used.

The project is devoted to the study of new path querying algorithms using context-free and conjunctive languages for poorly studied query semantics. This study relies on the existing results, which indicate the applicability of the matrix approach in problems of path querying. In addition, the study aims to improve the existing path querying algorithms and create new ones that allow us to describe a wider class of queries to graphs by combining the formal languages and query semantics used. In addition, it is planned to prove the theoretical properties of the obtained algorithms. The obtained new algorithms for path querying using context-free and conjunctive languages, effective from the point of view of parallel systems, will make it possible to construct effective implementations for analyzing large graphs. It is planned to conduct a series of experiments on the use of the obtained algorithms for solving applied problems and for comparing them with analogs.

2 Содержание проекта

2.1 Цель и задачи проекта

Целью проекта является разработка эффективных алгоритмов поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков.

Достижение поставленной цели обеспечивается решением следующих задач.

1) Разработать и реализовать алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, эффективные с точки зрения параллельных систем.

2) Исследовать временные сложности, и другие свойства полученных алгоритмов.

3) Провести экспериментальное исследование предложенных алгоритмов и их сравнение с аналогами.

2.2 Направление из Стратегии научно-технологического развития Российской Федерации (при наличии) (выбор из справочника)

1) Переход к передовым цифровым, интеллектуальным производственным технологиям, роботизированным системам, к новым материалам и способам конструирования, создание систем обработки больших объемов данных, машинного обучения и искусственного интеллекта;

2.3 Анализ современного состояния исследований в данной области (приводится обзор исследований в данной области со ссылками на публикации в научной литературе)

Графы используются в качестве структуры данных для представления больших объемов информации в компактной и удобной для анализа форме во многих областях, например, в биоинформатике, в графовых базах данных, при статическом анализе программ. При этом необходимо анализировать такие графы и выявлять сложные зависимости между их вершинами. Например, часто бывает необходмо проверить наличие пути, удовлетворяющего заданным свойствам, между вершинами. Один из способов задать ограничение на путь в графе с метками на рёбрах — это задать ограничение на слово, составленное из меток рёбер этого пути, для чего естественным образом могут быть использованы механизмы теории формальных языков. В таком случае, результатом будет являться множество всех троек (A, m, n), для которых существует путь в графе от вершины m до вершины n такой, что метки на ребрах этого пути образуют строку, выводимою из нетерминала A в некоторой формальной грамматике. Таким образом, возникает класс задач, называемый задачами поиска путей с ограничениями в терминах формальных языков с использованием реляционной семантики запросов. Наиболее популярными являются запросы, которые используют контекстно-свободные грамматики. Кроме того, существуют конъюнктивные грамматики, образующие более широкий класс грамматик. Использование конъюнктивных грамматик в задаче поиска путей позволяет формулировать более сложные запросы к графу и решать более широкий круг задач. Известно, что задача вычисления запросов к графу с использованием реляционной семантики и конъюнктивных грамматик является неразрешимой. Существующие алгоритмы поиска путей с использованием конъюнктивных языков строят аппроксимацию решения.

Имеется ряд алгоритмов для поиска путей с использованием реляционной семантики запросов и КС-грамматик (Hellings. J. Conjunctive context-free path queries, 2014; Sevon P., Eronen L. Subgraph queries by context-free grammars, 2008; Zhang X. et al. Context-free path queries on RDF graphs, 2016), которые, однако, демонстрируют низкую производительность на больших графах. Одной из самых популярных техник, используемых для увеличения производительности при работе с большими объемами данных, является использование параллельных систем, однако перечисленные алгоритмы не позволяют эффективно использовать данную технику.

Кроме того, существует алгоритм поиска путей (Azimov R., Grigorev S. Context-Free Path Querying by Matrix Multiplication, 2018), использующий реляционную семантику запросов и КС-грамматики, и решающий данную задачу с применением матричных операций. Используемый матричный подход, заключается в том, чтобы построить матрицу смежности

входного графа, элементами которой являются множества нетерминалов входной грамматики. В ячейку i,j матрицы смежности добавляется нетерминал A, тогда и только тогда, когда существует ребро из вершины с номером i в вершину с номером j, метка которого выводима из нетерминала A. Далее вычисляется транзитивное замыкание построенной матрицы, используя правила вывода входной грамматики. В процессе вычисления транзитивного замыкания используются операции умножения и сложения булевых матриц. Известно, что для вычислений матричных операции можно эффективно использовать параллельные системы (Che S., Beckmann B.M., Reinhardt S.K. Programming GPGPU Graph Applications with Linear Algebra Building Blocks, 2016).

Также существует алгоритм поиска путей (Zhang Q., Su Z. Context-sensitive data-dependence analysis via linear conjunctive language reachability, 2017), работающий с конъюнктивными грамматиками. Но данный алгоритм принимает на вход только определенный подкласс конъюнктивных грамматик, а именно, линейные конъюнктивные грамматики, которые имеют не более одного нетерминального символа в каждом конъюнкте правила. Кроме того, существует алгоритм поиска путей (Р.Ш. Азимов, С.В. Григорьев, Синтаксический анализ графов с использованием конъюнктивных грамматик, 2018), работающий с любыми конъюнктивными грамматиками и использующий матричный подход, аналогичный вышеизложенному для контекстно-свободных грамматик.

Существует алгоритм поиска путей (Nole M., Sartiani C. Regular Path Queries on Massive Graphs, 2016) с использованием регулярных грамматик, который позволяет эффективно использовать параллельные системы. Этот подход основывается на идее вычисления на каждом этапе производных регулярных выражений в соответствии с символами на ребрах графа.

2.4 Предлагаемые методы и подходы к решению поставленных задач (включая детальный план проводимых исследований)

В данном проекте планируется рассмотреть различные семантики для задачи поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков. Кроме реляционной семантики также существуют семантики одного пути (single-path) и всех путей (all-path). При использовании single-path или all-path семантики требуется не только найти множество всех троек (A,m,n), но и предоставить для каждой из них один или все такие пути из вершины m в вершину n.

Далее планируется для этих семантик разработать и реализовать алгоритмы поиска путей с использованием контекстно-свободных и конъюнктивных грамматик. При этом для возможности эффективного использования параллельных систем планируется улучшить существующий матричный подход или обобщить подход с производными до контекстно-свободных грамматик.

Также планируется найти временную сложность и другие теоретические свойства предложенных алгоритмов.

Кроме того, для эффективной работы предложенных алгоритмов с большими графами планируется использовать различные оптимизации. Во многих областях, графы, для которых решается задача поиска путей с ограничениями в терминах формальных языков, являются

разреженными. Поэтому одной из таких оптимизаций, например, является использование представлений и операций, эффективно работающих для разреженных матриц.

После этого планируется провести ряд экспериментов на реальных данных, имеющихся в открытом доступе, с целью проверить практическую применимость разработанных алгоритмов. Кроме того, планируется провести сравнение предложенных алгоритмов с аналогами.

2.5 Новизна исследования, заявленного в проекте (формулируется новая научная идея, обосновывается новизна предлагаемой постановки и решения заявленной проблемы)

Сформулированные задачи нацелены на улучшение существующих алгоритмов и создание новых, позволяющих описывать более широкий класс запросов к графам за счёт комбинации используемых формальных языков и семантик запросов.

Кроме того, будут исследованны и доказанны новые теоретические свойства предложенных алгоритмов, например, такие как временная сложность.

Также полученные новые алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, эффективные с точки зрения параллельных систем, дадут возможность построения эффективных реализаций для анализа больших графов.

2.6 Ожидаемые по окончании проекта научные результаты

Предложены алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, single-path и all-path семантик запросов, эффективные с точки зрения параллельных систем. Исследованы и доказаны теоретические свойства предложенных алгоритмов, например, такие как временная сложность. Предложенные алгоритмы реализованы и проведён ряд экспериментов по их использованию для решения прикладных задач. Проведено сравнение предложенных алгоритмов с аналогами. Разработанные алгоритмы представлены на конференции и опубликованы в сборнике материалов конференции, индексируемом в Scopus.

2.7 Научный задел Научного руководителя по тематике проекта

Научный руководитель обладает большим опытом в применении теории формальных языков к различным задачам, что подтверждается рядом публикаций и выступлениями на профильных конференциях. Первые результаты были получены руководителем при работе над своей диссертацией "Синтаксический анализ динамически формируемых программ". В дальнейшем им были предложены алгоритмы поиска путей с контекстно-свободными ограничениями, основанные на нисходящем (Generalized LL) и восходящем (Generalized LR) алгоритмах синтаксического анализа. Результаты изложены в работах Grigorev S., Ragozina A. "Context-free path querying with structural representation of result"и Verbitskaia E., Grigorev S., Avdyukhin D. "Relaxed parsing of regular approximations of string-embedded languages".

Также, им предложено решение, позволяющее использовать парсер-комбинаторы для поиска путей с контекстно-свободными ограничениями (Verbitskaia E. et al. "Parser combinators for context-free path querying").

Под его руководством Рустамом Азимовым был разработан алгоритм для поиска путей с контекстно-свободными ограничениями, основанный на умножении булевых матриц (Azimov R., Grigorev S. "Context-free path querying by matrix multiplication"). В дальнейшем было проведено экспериментальное исследование этого алгоритма (Mishin N. et al. "Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication")

2.8 Педагогический задел Научного руководителя (обязательно указать, количество аспирантов, из них – количество защитивших диссертацию; количество ученых, защитивших диссертации на соискание ученой степени доктора 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 работает один аспирант. Всего аспрантов за время работы — 1, из них в настоящее время — 1, защитивших диссертацию — 0.

Также руководителем регулярно читаются следующие курсы: теория формальных языков, теория графов, алгоритмы и структуры данных, практика программирования.

2.9 Список основных публикаций Научного руководителя в рецензируемых журналах (не менее 5)

Научный руководитель имеет следующие основные публикации в рецензируемых журналах.

1) Р.Ш. Азимов, С.В. Григорьев, Синтаксический анализ графов с использованием конъюнктивных грамматик, Труды Института системного программирования РАН, том 30, вып. 2, стр. 149-166, 2018 г.

2) Polubelova M., Grigorev S., Lexical analysis of dynamically generated string expressions, Systems and Means of Informatics, pp. 43-62, 2016.

3) Polubelova M., Bozhko S., Grigorev S., Certified grammar transformation to Chomsky normal form in F, Труды Института системного программирования РАН, том 28, вып. 2, стр. 127-138, 2016.

4) С.В. Григорьев, Е.А. Вербицкая, М.И. Полубелова, А.В. Иванов, Е.В. Мавчун, Инструментальная поддержка встроенных языков в интегрированных средах разработки, Моделирование и анализ информационных систем, том 21, вып. 6, стр. 131-143, 2015 г.

5) С.В. Григорьев, А.К. Рагозина, Обобщенный табличный LL-анализ, Системы и средства информатики, том 25, вып. 1, стр. 89-107, 2015 г.

2.10 Название диссертационной работы Аспиранта

Поиск путей с ограничениями в терминах формальных языков

2.11 Основные цели и задачи диссертационного исследования

Целью диссертационного исследования является разработка алгоритмов поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, эффективных с точки зрения параллельных систем.

Достижение поставленной цели обеспечивается решением следующих задач.

1) Разработать и реализовать алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, основанные на матричном подходе.

2) Разработать и реализовать алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, основанные на вычислении производных формальных языков.

3) Разработать и реализовать алгоритмы поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных языков, основанные на нисходящем (Generalized LL) алгоритме синтаксического анализа.

2.12 Список основных (не более 5) публикаций Аспиранта в рецензируемых журналах

Аспирант имеет следующую публикацию в рецензируемом журнале.

1) Р.Ш. Азимов, С.В. Григорьев, Синтаксический анализ графов с использованием конъюнктивных грамматик, Труды Института системного программирования РАН, том 30, вып. 2, стр. 149-166, 2018 г.

2.13 Научный задел Аспиранта по тематике проекта (необходимо указать сколько выступлений на конференциях; список всех публикаций; прочие достижения (премии, награды, гранты))

Аспирант обладает большим опытом в применении теории формальных языков к различным задачам, что подтверждается рядом публикаций и выступлениями на профильных конференциях. А именно, выступление на всероссийской конференции PLC 2017 и выступление и участие в постерной сессии на международной конференции GRADES-NDA 2018.

Аспирант имеет следующие публикации.

1) SCOPUS, Azimov R., Grigorev S. Context-Free Path Querying by Matrix Multiplication, In Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), 2018;

2) ВАК, Р.Ш. Азимов, С.В. Григорьев, Синтаксический анализ графов с использованием конъюнктивных грамматик, Труды ИСП РАН, том 30, вып. 2, 2018 г., стр. 149-166;

3) РИНЦ, Р.Ш. Азимов, С.В. Григорьев, Синтаксический анализ графов с использованием конъюнктивных грамматик, Сборник трудов всероссийской конференции PLC, 2017 г., стр. 24-27.

2.14 Дата приказа о переводе на второй курс аспирантуры

01.07.2019