Поиск путей с ограничениями в терминах формальных языков

Семён Григорьев

1 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. Ответом на такие запросы обычно является 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. Говорят, что такой тип запросов вычислен при решении задачи поиска путей 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.

Во многих областях необходимо решать задачу поиска путей в терминах формальных языков на больших графах. Одной из самых популярных техник, используемых для увеличения производительности при работе с большими объемами данных, является использование параллельных систем. Единственным известным подходом в данной области, позволяющим эффективно использовать параллельные системы, является матричный подход, при котором строится матрица инцидентности входного графа, элементами которой являются множества нетерминалов входной грамматики. Далее вычисляется транзитивное замыкание построенной матрицы, используя правила вывода входной грамматики. В процессе вычисления транзитивного замыкания активно используются операции умножения и сложения булевых матриц.

Кроме того, существуют слабо изученые семантики запросов, отличные от реляционной. Например, при использовании семантик одного (single-path) и всех путей (all-path), необходимо для каждой тройки (A, m, n), найденной при реляционной семантике запросов, также предоставить один или все такие пути из вершины m в вершину n, соответственно.

Проект посвящён исследованию новых алгоритмов поиска путей с использованием контекстно-свободных и конъюнктивных языков, single-path и all-path 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. Кроме того, планируется доказать теоретические свойства полученных алгоритмов.

Также, стоит отметить, что применение таких классов грамматик, как конъюнктивные, в данной области изучено крайне слабо. Таким образом, будут полученны новые теоретические результаты для задач поиска путей с ограничениями в терминах конъюнктивных грамматик.

Кроме того, 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 grammars

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 a queries for large graphs in order to determine the dependencies between the nodes. The answer to such queries is usually a set of all triples (A, m, n) for which there is a path in the graph from the vertex m to the vertex n, such that the labels on the edges of this path form a string derivable from the nonterminal A in some formal grammar. This type of query is calculated using relational query semantics. The most popular are queries that use context-free grammars. 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. It is known that the path querying using relational query semantics and conjunctive grammars is undecidable problem. Existing path querying algorithms using conjunctive languages build an approximation of the solution.

In many areas, it is necessary to solve the problem of finding paths in terms of formal languages on large graphs. One of the most popular techniques used to increase performance when working with large data is the use of parallel systems. The only known approach in this area that makes it possible to use the parallel systems effectively is the matrix approach, in which the incidence 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 calculating the transitive closure, the operations of multiplication and addition of Boolean matrices are actively used.

In addition, there are poorly studied query semantics, other than relational. For example, when using the single-path and all-path semantics, it is necessary for each triple (A, m, n) found using relational query semantics to also provide one and all such paths from vertex m to vertex n, respectively.

The project is devoted to the study of new path querying algorithms using context-free and conjunctive languages, single-path and all-path query semantics. This study relies on the available 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.

Also, it is worth noting that the use of such classes of grammar as conjunctive in this area is studied poorly. Thus, new theoretical results will be obtained for path querying problems using conjunctive grammars.

In addition, 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 вычислять запросы к большим графам с целью выявления сложных зависимостей между их вершинами. Результатом вычисления таких запросов обычно является множество неявных отношений между вершинами графа, то есть путей. Естественно помечать ребра графа символами из некоторого конечного алфавита и выделять пути с помощью формальных грамматик над тем же алфавитом (регулярные выражения, контекстно-свободные грамматики).

Таким образом, задача поиска путей с ограничениями в терминах формальных языков является одной из активно развивающихся. В данной задаче на вход подается ориентированный граф с метками на ребрах и формальная грамматика. Говорят, что тип запросов вычислен с использованием реляционной семантики, если результатом является множество всех троек (A,m,n), для которых существует путь из вершины входного графа m в вершину n, метки на ребрах которого образуют строку, выводимую из нетерманала A входной грамматики. Наиболее популярными являются запросы, которые используют контекстно-свободные грамматики.

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, так как имеют в худшем случае кубическую временную сложность. 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 вычисляющий матричное транзитивное замыкание с применением матричных операций. Используемый матричный подход, заключается в том, чтобы построить матрицу инцидентности входного графа, элементами которой являются множества нетерминалов входной грамматики. 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 существуют конъюнктивные грамматики, задающие более широкий класс языков, чем контекстно-свободные. Использование конъюнктивных грамматик в задаче поиска путей позволяет формулировать более сложные запросы к графам и решать более широкий круг задач, например, задачи поиска псевдонимов и уязвимостей в исходном коде. Необходимо отметить, что задача вычисления запросов к графу с использованием реляционной семантики и конъюнктивных грамматик является неразрешимой. Один из распространенных способов найти приближенное решение неразрешимой задачи — построить аппроксимацию решения.

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. Также, 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.

Таким образом, единственным известным подходом к задаче поиска путей с ограничениями в терминах контекстно-свободных и конъюнктивных грамматик, позволяющим эффективно использовать параллельные системы, является матричный подход, при котором вычисляется матричное транзитивное замыкание.

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 испоспользования параллельных систем планируется использовать существующий матричный подход, при котором задача сводится к вычислению матричного транзитивного замыкания.

Также, планируется найти временную сложность предложенных алгоритмов. Для этого планируется оценить сверху временную сложность в худшем случае и привести пример, на котором достигается найденная оценка сверху.

Кроме того, для эффективной работы предложенных алгоритмов с большими графами планируется использовать различные матричные 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)

Во многих областях имеется необходимость в эффективных алгоритмах поиска путей с ограничениями в терминах формальных языков. Сформулированные задачи опираются на имеющиеся результаты, которые говорят о применимости матричного подхода в задачах поиска путей. Кроме того, они 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.

Также, стоит отметить, что применение таких классов грамматик, как конъюнктивные, в данной области изучено крайне слабо. Таким образом, полученные теоретические результаты для задач поиска путей с ограничениями в терминах конъюнктивных грамматик будут новыми.

Кроме того, 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

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)

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 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.

2.12 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 имеет следующий научный задел по тематике проекта.

Выступление на конференциях: одно выступление на всероссийской конференции PLC 2017, один постер на международную конференцию GRADES-NDA 2018.

Список всех публикаций:

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) Сборник трудов всероссийской конференции PLC 2017, стр. 24-27.

2.14 Дата приказа о переводе на второй курс аспирантуры