Санкт-Петербургский Государственный Университет

Программная инженерия
Кафедра системного программирования

Орачев Егор Станиславович

Реализация алгоритма поиска путей в графовых базах данных через тензорное произведение на GPGPU

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент С. В. Григорьев

Рецензент:

Санкт-Петербург
2020

Оглавление

Введение

Все чаще современные системы аналитики и рекомендаций строятся на основе анализа данных, структурированных с использованием *графовой модели*. В данной модели основные сущности представляются вершинами графа, а отношения между сущностями — ориентированными ребрами с различными метками. Подобная модель позволяет относительно легко и практически в явном виде моделировать сложные иерархические структуры, которые не так просто представить, например, в классической *реляционной модели*. В качестве основных областей применения графовой модели можно выделить следующие: графовые базы данных [3], анализ RDF данных [4], биоинформатика [12] и статический анализ кода [8].

Поскольку графовая модель используется для моделирования отношений между объектами, при решении прикладных задач возникает необходимость выявления более сложных взаимоотношений между объектами. Для этого чаще всего формируются запросы в специализированных программных средствах для управления графовыми базами данных. В качестве запроса можно использовать некоторый *шаблон* на путь в графе, который будет связывать объекты, т.е. выражать взаимосвязь между ними. В качестве такого шаблона можно использовать формальные грамматики, например, регулярные или контекстно-свободные (КС). Используя вычислительно более выразительные грамматики, можно формировать более сложные запросы и выявлять нестандартные и скрытые ранее взаимоотношения между объектами. Например, *same-generation queries* [1], сходные с сбалансированными скобочными последовательностями Дика, могут быть выражены КС грамматиками, в отличие от регулярных.

Результатом запроса может быть множество пар объектов, между которыми существует путь в графе, удовлетворяющий заданным ограничениям. Также может возвращаться один экземпляр такого пути для каждой пары объектов или итератор всех путей, что зависит от семантики запроса. Поскольку один и тот же запрос может иметь разную се-

мантику, требуются различные программные и алгоритмические средства для его выполнения.

Запросы с регулярными ограничениями изучены достаточно хорошо, языковая и программная поддержка выполнения подобных запросов присутствует в некоторых в современных графовых базах данных. Однако, полноценная поддержка запросов с КС ограничениями до сих пор не представлена. Существуют алгоритмы [4, 9, 2, 5, 10] для вычисления запросов с КС ограничениями, но потребуется еще время, прежде чем появиться полноценная высокпроизводительная реализация одного из алгоритмов, способная обрабатывать реальные графовые данные.

Работы [7, 6] в качестве реализации алгоритма [2] показывают, что возможно использовать GPGPU для выполнения наиболее вычислительно сложных частей алгоритма, что дает *существенный* прирост в производительности. Недавно представленный алгоритм [5] для вычисления запросов с КС ограничениями полагается на операции линейной алгебры, в частности, произведение Кронекера (частный случай тензорного произведения), умножение и сложение матриц в полукольце булевой алгебры. Данный алгоритм позволяет выполнять запросы для всех ранее упомянутых семантик, потенциально поддерживает бóльшие по размеру КС запросы, а также хорошо реализуется с помощью программных средств для вычисления на GPGPU.

Таким образом, важной задачей является не только реализация перспективного алгоритма [5] для выполнения запросов с КС ограничениям, но и разработка программной библиотеки для работы с примитивами линейной булевой алгебры, которая позволила бы упростить прототипирование и реализацию подобного и будущих алгоритмов на GPGPU, в частности, на платформе NVIDIA CUDA [11].

1. Постановка задачи

Цель данной работы — реализация алгоритма поиска путей в графовых базах данных через тензорное произведение на платформе NVIDIA CUDA в качестве GPGPU технологии. Для ее достижения были поставлены следующие задачи:

- Реализация библиотеки для работы с примитивами булевой алгебры на GPGPU
- Реализация алгоритма поиска путей
- Экспериментальное исследование реализации алгоритма

2. Обзор предметной области

2.1. Терминология

2.2. Поиск путей с контекстно-свободными ограничениями

2.3. Существующие решения

2.4. Поиск путей через произведение Кронекера

3. Реализация библиотеки матричных операций

4. Реализация алгоритма

5. Экспериментальное исследование

Заключение

Список литературы

[1] Abiteboul Serge, Hull Richard, Vianu Victor. Foundations of Databases. — 1995. — 01. — ISBN: 0-201-53771-0.

[2] Azimov Rustam, Grigorev Semyon. Context-free path querying by matrix multiplication. — 2018. — 06. — P. 1–10.

[3] Barceló Baeza Pablo. Querying Graph Databases // Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. — PODS '13. — New York, NY, USA : Association for Computing Machinery, 2013. — P. 175–188. — Access mode: `https://doi.org/10.1145/2463664.2465216`.

[4] Context-Free Path Queries on RDF Graphs / Xiaowang Zhang, Zhiyong Feng, Xin Wang et al. // CoRR. — 2015. — Vol. abs/1506.00743. — 1506.00743.

[5] Context-Free Path Querying by Kronecker Product / Egor Orachev, Ilya Epelbaum, Rustam Azimov, Semyon Grigorev. — 2020. — 08. — P. 49–59. — ISBN: 978-3-030-54831-5.

[6] Context-Free Path Querying with Single-Path Semantics by Matrix Multiplication / Arseniy Terekhov, Artyom Khoroshev, Rustam Azimov, Semyon Grigorev. — 2020. — 06. — P. 1–12.

[7] Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication / Nikita Mishin, Iaroslav Sokolov, Egor Spirin et al. — 2019. — 06. — P. 1–5.

[8] Fast Algorithms for Dyck-CFL-Reachability with Applications to Alias Analysis / Qirun Zhang, Michael R. Lyu, Hao Yuan, Zhendong Su // SIGPLAN Not. — 2013. — Jun. — Vol. 48, no. 6. — P. 435–446. — Access mode: `https://doi.org/10.1145/2499370.2462159`.

[9] Hellings Jelle. Path Results for Context-free Grammar Queries on Graphs. — 2015. — 02.

[10] Medeiros Ciro, Musicante Martin, Costa Umberto. An Algorithm for Context-Free Path Queries over Graph Databases. — 2020. — 04.

[11] NVIDIA. CUDA Toolkit Documentation // NVIDIA Developer Zone. — 2020. — Access mode: `https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html` (online; accessed: 01.12.2020).

[12] Quantifying variances in comparative RNA secondary structure prediction / James Anderson, Adám Novák, Zsuzsanna Sükösd et al. // BMC bioinformatics. — 2013. — 05. — Vol. 14. — P. 149.