ContextFree Wars: The RedisGraph Strikes Back

Arseniy Terekhov
simpletondl@yandex.ru
Saint Petersburg State University
St. Petersburg, Russia

Vlada Poguzelskaya
!!!@!!!
Saint Petersburg State University
St. Petersburg, Russia

Vadim Abzalov
!!!@!!!
Saint Petersburg State University
St. Petersburg, Russia

Semyon Grigorev
s.v.grigoriev@spbu.ru
semyon.grigorev@jetbrains.com
Saint Petersburg State University
St. Petersburg, Russia
JetBrains Research
St. Petersburg, Russia

Figure 1: Episode IV: A New Hope

ABSTRACT

A long time ago in a galaxy far far away...

1 INTRODUCTION

Language-constrained path querying [2] is a way to find paths in edge-labeled graphs when constraints are formulated in terms of language which restrict words formed by paths: the word formed by path's labels concatenation should be in the specified language. This way is very natural for navigational queries in graph databases, and one of the most popular languages which are used for constraints is a regular language. But in some cases, regular languages are not expressive enough, as a result, context-free languages gain popularity. Constraints in the form of context-free languages, or context-free path querying (CFPQ), can be used for RDF analysis [8], biological data analysis [6], static code analysis [5, 9], and in other areas.

Big amount of research done on CFPQ, a number of CFPQ algorithms were proposed, but the application of context-free constraints for real-world data analysis faced with some problems problem. The first problem is a bad performance of proposed algorithms on real-world data, as was shown by Jochem Kuijpers et al. [3]. Moreover, there are no graph databases with full-stack support of CFPQ, the main effort was made in algorithms and their theoretical properties research. This fact hinders research of problems reducible to CFPQ, thus it hinders the development of new solutions for some problems. For example, recently graph segmentation in data provenance analysis was reduced to CFPQ [?], but authors faced the problem during the evaluation of the proposed approach: no one graph database support CFPQ.

In [1] Rustam Azimov propose a matrix-based algorithm for CFPQ. This algoritm is one of promissing way to solve the first problem and provide appropriate solution for real-world data analysis, as was shown by Nikita Mishim et al. in [4] and Arseniy Terekhov et al. in [7]. But this algorithm always computes information (reachability facts or single path which satisfies constraints) for all pairs of vertices in the graph. It is unreasonable for some real-world scenarios when one can provide a relatively small set of start vertices or even single start vertex.

While all-pairs context-free path querying is a classical problem that investigates in a number of works, there is no, in our knowledge, solutions for single-source and multiple-source CFPQ. In this work we propose a matrix-based multiple-source (and single source as a partial case) CFPQ algorithm.

Also, we provide full-stack support of CFPQ for the RedisGraph[1] [?] graph database. We implement a Cypher query language extension[2] that allows one to express context-free constraints, and extend the RedisGraph to support this extension. In our knowledge, it is the first full-stack implementation of CFPQ.

To sommarize, we make the following contribution in this paper.

(1) Multiple-source matrix-based CFPQ algorithm. Single-source as a partial case.
(2) Evaluateion of two versions of this algorithm.
(3) RedisGraph extending to provide full-stack support of CFPQ.

2 PRELIMINARIES

In this section we introduce common definitions in graph theory and formal language theory which will be used in this paper.

[1] !!!
[2] !!!

Also, we provide brief description of Azimov's algorithm which will be used as a base of our solution.

2.1 Graphs

Labelled digraph, matrices, ...

Definition 2.1. Digraph

2.2 Languages

Grammars, normal forms, ...

Definition 2.2. Digraph

2.3 Matrix-Based Algorithm

Description

Pseudocode.

What about arbitrary (not CNF) CFG?

3 MATRIX-BASED MULTIPLE-SOURCE CFPQ ALGORITHM

New algo description.

3.1 Implementation Details

Algo implementation details: python, graphBLAS, smthng else? Two versions.

3.2 Algorithm Evaluation

And comparison. With combinators, GLL (.NET version).

Evaluation setup.

Tables.

Results.

Conclusion.

4 REDISGRAPH EXTENDING

In order to provide full-stack support of CFPQ we choose an appropriatr graph DB.

4.1 Cypher

parser extending, proposal, ...

Examples of queries.

4.2 RedisGraph

CFPQ to matrix expressions, etc.

Limits, restrictions, etc.

4.3 Evaluation

Small basic evalustion on real-world graph (geo?). In order to show, that performance is reasonable.

5 CONCLUSION

Conclusion

Future research

Formal translation of Cypher to linear algebra. In order to formalize limits and restrictions. There are works on a subset of SPARQL to linear algebra translation. But they are very limited.

For real-world solutions is important to provide a unified algorithm for both RPQ and CFPQ. A matrix-based algorithm is not a better choice. Experiments on a unified evaluation of RPQ and CFPQ (tensors?).

Multiple-source for Neo4j (non-linear algebra based approaches evaluation).

More applications.

REFERENCES

[1] Rustam Azimov and Semyon Grigorev. 2018. 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) (GRADES-NDA '18)*. ACM, New York, NY, USA, Article 5, 10 pages. https://doi.org/10.1145/3210259.3210264

[2] C. Barrett, R. Jacob, and M. Marathe. 2000. Formal-Language-Constrained Path Problems. *SIAM J. Comput.* 30, 3 (2000), 809–837. https://doi.org/10.1137/S0097539798337716 arXiv:https://doi.org/10.1137/S0097539798337716

[3] Jochem Kuijpers, George Fletcher, Nikolay Yakovets, and Tobias Lindaaker. 2019. An Experimental Study of Context-Free Path Query Evaluation Methods. In *Proceedings of the 31st International Conference on Scientific and Statistical Database Management (SSDBM '19)*. ACM, New York, NY, USA, 121–132. https://doi.org/10.1145/3335783.3335791

[4] Nikita Mishin, Iaroslav Sokolov, Egor Spirin, Vladimir Kutuev, Egor Nemchinov, Sergey Gorbatyuk, and Semyon Grigorev. 2019. Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication. In *Proceedings of the 2Nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA'19)*. ACM, New York, NY, USA, Article 12, 5 pages. https://doi.org/10.1145/3327964.3328503

[5] Jakob Rehof and Manuel Fähndrich. 2001. Type-Base Flow Analysis: From Polymorphic Subtyping to CFL-Reachability. *SIGPLAN Not.* 36, 3 (Jan. 2001), 54–66. https://doi.org/10.1145/373243.360208

[6] Petteri Sevon and Lauri Eronen. 2008. Subgraph Queries by Context-free Grammars. *Journal of Integrative Bioinformatics* 5, 2 (2008), 157 – 172. https://doi.org/10.1515/jib-2008-100

[7] Arseniy Terekhov, Artyom Khoroshev, Rustam Azimov, and Semyon Grigorev. 2020. Context-Free Path Querying with Single-Path Semantics by Matrix Multiplication. In *Proceedings of the 3rd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA'20)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3398682.3399163

[8] Xiaowang Zhang, Zhiyong Feng, Xin Wang, Guozheng Rao, and Wenrui Wu. 2016. Context-Free Path Queries on RDF Graphs. In *The Semantic Web – ISWC 2016*, Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil (Eds.). Springer International Publishing, Cham, 632–648.

[9] Xin Zheng and Radu Rugina. 2008. Demand-driven Alias Analysis for C. In *Proceedings of the 35th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '08)*. ACM, New York, NY, USA, 197–208. https://doi.org/10.1145/1328438.1328464