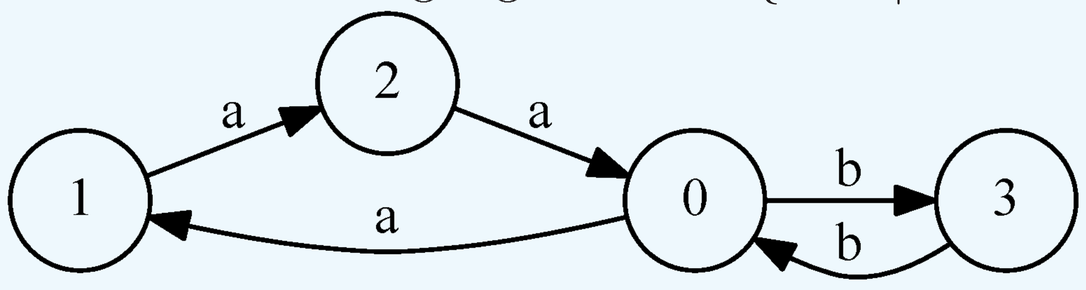
Evaluation of the Context-Free Path Querying Algorithm Based on Matrix Multiplication

Semyon Grigorev
JetBrains Research, Saint Petersburg University, Russia
s.v.grigoriev@spbu.ru, Semen.Grigorev@jetbrains.com

Contex-Free Path Querying

Find paths which satisfy constraints in form of a formal language $L = \{a^n b^n | n > 0\}$



Query = grammar for L: $S \to a\ b \mid a\ S\ b$

Result: $\{(u, v) \mid \exists p \text{ from } u \text{ to } v : \text{word}(p) \in L\}$

Results

- Dataset for CFPQ evaluation is collected and published
  - Contains both graphs and queries
  - Contains both real-world and syntetic graphs
- Several CFPQ algorithms implementations are created, evaluated and published

Future Research

- Create open extensible platform for CFPQ algorithms evaluation
- Extend dataset with new data
- Implement and evaluate destributed matrix-based CFPQ algorithms
- Implement and evaluate sparse boolean matrix-based CFPQ algorithms

Matrix-based Algorithm [1]

T is an adjacency matrix of the input graph
The grammar is in the normal form

$$T_{ij} = \{N \mid N \xRightarrow{*} \omega, \omega - \text{path bw } i \text{ and } j\}$$
$$T_{ik} \times T_{kj} = \{A \mid B \in T_{ik}, C \in T_{kj}, A \to BC\}$$
$$T^{(i)} = T^{(i-1)} \cup (T^{(i-1)} \times T^{(i-1)})$$

- Can be formulated in terms of boolean matrices multiplication
- Easy to run in parallel environments: GPUs, multithreaded CPUs, clusters

Implementations

Our implementations [2]:

[Scipy] Matrix-based algorithm which uses sparse matrices from **Scipy** library (**Python**)

[M4RI] Matrix-based algorithm which uses dense matrices multiplication from **m4ri** library (Method of Four Russians, **C**)

[GPU] Matrix-based algorithm which uses our own implementation of the naïve boolean matrix multiplication in **CUDA C**

Reference implementations:

[CuSprs] Matrix-based algorithm [1] which uses NVIDIA cuSPARSE library (**CUDA C, GPGPU**)

[CYK] CYK-based algorithm [3] implemented in **Java** (CPU)

We Need More Real-World Data

Graph: classical ontologies (RDFs)
Query: same-generation query over `type` and `SubClassOf` relations
Grammar: $S \to scor\ S\ sco \mid tr\ S\ t \mid scor\ sco \mid tr\ t$

RDF			Algorithms				
Name	#V	#E	Scipy	M4RI	GPU	CuSprs	CYK
atm-prim	291	685	3	2	1	269	515285
biomed	341	711	3	5	1	283	420604
pizza	671	2604	6	8	1	292	3233587
wine	733	2450	7	6	1	294	4075319

- 2019 (GPU) is 10^6 times faster than 2016 (CYK) on real-world data
  - Reasonable time even for CPU based implementations
- We should find bigger RDFs
- We should find other real-world cases for CFPQ
  - Both graphs and queries

We Should Do More Research on the Algorithms Scaling

	Graph	Scipy	M4RI	GPU	CuSprs
Sparse graphs are generated by GTgraph Query: $S \to a\ S\ b \mid a\ b$	G10k-0.001	37.286	2.395	0.215	35.937
	G10k-0.1	601.182	1.050	0.114	395.393
	G40k-0.001	-	97.841	8.393	-
	G80k-0.001	-	1142.959	65.886	-
Graph is a cycle Query: $S \to S\ S \mid a$	25000	-	33.236	5.314	-
	50000	-	360.035	44.611	-
	80000	-	1292.817	190.343	-

- We can handle graphs with 80k vertices in reasonable time by using GPGPU
  - Technical bound: GPGPU RAM does not fit bigger graphs
- We should evaluate multi-GPU systems
- We should evaluate destributed solutions
- We should implement a sparse boolean matrices library for GPGPU

Contact us

Our team:

- Semyon Grigorev: s.v.grigoriev@spbu.ru
- Nikita Mishin: mishinnikitam@gmail.com
- Iaroslav Sokolov: sokolov.yas@gmail.com
- Egor Spirin: egor@spirin.tech
- Egor Nemchinov: nemchegor@gmail.com
- Vladimir Kutuev: vladimir.kutuev@gmail.com
- Sergey Gorbatyuk: sergeygorbatyuk171@gmail.com

Both dataset and implementations are available on GitHub:
`https://github.com/SokolovYaroslav/CFPQ-on-GPGPU`

References

[1] Rustam Azimov and Semyon Grigorev. 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, pages 5:1–5:10, 2018.

[2] Nikita Mishin, Iaroslav Sokolov, Egor Spirin, Vladimir Kutuev, Egor Nemchinov, Sergey Gorbatyuk, and Semyon Grigorev. 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, pages 12:1–12:5, New York, NY, USA, 2019. ACM.

[3] Xiaowang Zhang, Zhiyong Feng, Xin Wang, Guozheng Rao, and Wenrui Wu. Context-free path queries on rdf graphs. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016*, pages 632–648, Cham, 2016. Springer International Publishing.

Acknowledgments

The research is supported by the JetBrains Research grant and the Russian Science Foundation grant 18-11-00100.