Formal Languages Theory is Not Only a Parsing

Semyon Grigorev

JetBrains Research, Programming Languages and Tools Lab

13.04.2018

Paths in graphs

- Graph analysis
  - Graph database querying
  - Network analysis (social networks, Internet, etc)
- Static code analysis
  - Alias analysis
  - Taint analysis
  - Types-related problems
  - Static analysis of string-embedded languages
- ...

Language constrained path querying

Language-constrained path querying, language reachability

- Σ is a set of terminals
- $L(\Sigma)$ is a language over Σ
- $G = (V, E, D)$ is a directed graph, $E \subseteq V \times D \times V$, $D \subseteq \Sigma$
- $p = v_0 \xrightarrow{l_0} v_1 \xrightarrow{l_1} \cdots v_{n-1} \xrightarrow{l_{n-1}} v_n$ is a path in G
- $w(p) = w(v_0 \xrightarrow{l_0} v_1 \xrightarrow{l_1} \cdots v_{n-1} \xrightarrow{l_{n-1}} v_n) = l_0 l_1 \cdots l_{n-1}$
- $R = \{p \mid w(p) \in L(\Sigma)\}$
    - **Problem:** R can be an infinite in some cases
- Task may be formulated in other way:
  $Q = \{(v_0, v_n) \mid \exists p = v_0 \xrightarrow{l_0} \cdots \xrightarrow{l_{n-1}} v_n \ (w(p) \in L(\Sigma))\}$

Regular constarints

- $L(\Sigma)$ is a regular language
  - Graph databases query languages (SPARQL, Cypher, PGQL)
  - OpenCypher: `https://goo.gl/5h5a8P`

Context-free constraints

- $L(\Sigma)$ is a context-free language
- Graph databases and semantic networks
  - *Sevon P., Eronen L.* "Subgraph queries by context-free grammars." 2008
  - *Zhang X. et al.* "Context-free path queries on RDF graphs." 2016
  - *Hellings J.* "Conjunctive context-free path queries." 2014
- Static code analysis
  - *Thomas Reps et al.* "Precise interprocedural dataflow analysis via graph reachability." 1995
  - *Qirun Zhang et al.* "Efficient subcubic alias analysis for C." 2014
  - *Dacong Yan et al.* "Demand-driven context-sensitive alias analysis for Java." 2011
  - *Jakob Rehof and Manuel Fahndrich.* "Type-base flow analysis: from polymorphic subtyping to CFL-reachability." 2001

Context-free constarints

- *Kai Wang et. al.* Graspan: A Single-machine Disk-based Graph System for Interprocedural Static Analyses of Large-scale Systems Code. 2017
  - " We have identified a total of 1127 unnecessary NULL tests in Linux, 149 in PostgreSQL, 32 in httpd."
  - "Our analyses reported 108 new NULL pointer dereference bugs in Linux, among which 23 are false positives"
  - "For PostgreSQL and httpd, we detected 33 and 14 new NULL pointer bugs; our manual validation did not find any false positives among them."

Linear-conjunctive constraints

- $L(\Sigma)$ is a linear-conjunctive language
  - Interleaving of balcned brackets:
    $L_1 = \{a^n b^n | n \geq 0\}; L_2 = \{c^m d^m | m \geq 0\}; L_3 = L_1 \odot L_2 = \{ab; acbcdd; cdab; \dots\}$
- *Qirun Zhang and Zhendong Su.* Context-sensitive data-dependence analysis via linear conjunctive language reachability. 2017

Возможные направления

- Theoretical open problem
  - Is there exists an algorithm with time complexity $O(|V|^{3-\varepsilon}), \varepsilon > 0$
- В данной области применимы решения из “классического” синтаксического анализа
  - Алгоритмы: CYK, (Generalized) LL, (Generalized) LR, Эрли, ...
  - Техники: комбинаторы, генераторы парсеров, ...
  - Оптимизации: использование GPGPU, специальные структуры данных (сжатое представление леса разбора, структурированный в виде графа стек), ...
- Из-за существенно бОльших объёмов данных требуются специальные оптимизации (распределённые вычисления, параллельные вычисления, ...)

Our experiments

- Generalized LL for CFPQ (GLL)
  - Based on Generalized LL: *Scott E., Johnstone A.* "GLL parsing"
  - Time complexity of !!!: $O\left(|V|^3 * \max_{v \in V}\left(deg^+(v)\right)\right)$
  - *Semyon Grigorev and Anastasiya Ragozina.* "Context-free path querying with structural representation of result." 2017
- GPGPU utilization for CFPQ (GPGPU)
  - Based on *Valiant L.* "General context-free recognition in less than cubic time." 1974
  - Временная сложность предложенного алгоритма: $O(|V|^2|N|^3(BMM(|V|) + BMU(|V|)))$
    - $BMM(n)$ — время, необходимое для умножения сложения булевых матриц $n \times n$
    - $BMU(n)$ — время, необходимое для поэлементного сложения булевых матриц $n \times n$
  - *Rustam Azimov, Semyon Grigorev.* "Context-Free Path Querying by Matrix Multiplication." 2017
- Parser-Combinators for Context-Free Path Querying (in Scala)

Performance comparison setup

$$
\begin{aligned}
0: &\quad \mathbf{S} \rightarrow \mathit{subClassOf}^{-1}\ \mathbf{S}\ \mathit{subClassOf} \\
1: &\quad \mathbf{S} \rightarrow \mathit{type}^{-1}\ \mathbf{S}\ \mathit{type} \\
2: &\quad \mathbf{S} \rightarrow \mathit{subClassOf}^{-1}\ \mathit{subClassOf} \\
3: &\quad \mathbf{S} \rightarrow \mathit{type}^{-1}\ \mathit{type}
\end{aligned}
$$

Query 1

$$
\begin{aligned}
0: &\quad \mathbf{S} \rightarrow \mathbf{B}\ \mathit{subClassOf} \\
1: &\quad \mathbf{S} \rightarrow \mathit{subClassOf} \\
2: &\quad \mathbf{B} \rightarrow \mathit{subClassOf}^{-1}\ \mathbf{B}\ \mathit{subClassOf} \\
3: &\quad \mathbf{B} \rightarrow \mathit{subClassOf}^{-1}\ \mathit{subClassOf}
\end{aligned}
$$

Query 2

Performance comparison result

№	#V	#E	Query 1 (ms)			Query 2 (ms)	
			CYK[1]	GLL	GPGPU	GLL	GPGPU
1	144	323	1044	10	12	1	1
2	129	351	6091	19	13	1	0
3	131	397	13971	24	30	1	10
4	179	413	20981	25	15	11	9
5	337	834	82081	89	32	3	6
6	291	685	515285	255	22	66	2
7	341	711	420604	261	20	45	24
8	671	2604	3233587	697	24	29	23
9	733	2450	4075319	819	54	8	6
10	6224	11840	–	1926	82	167	38
11	5864	19600	–	6246	185	46	21
12	5368	20832	–	7014	127	393	40

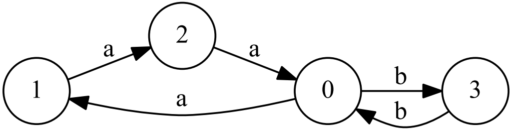
[1] Zhang, et al. "Context-free path queries on RDF graphs."

Info

- E-mail: `semen.grigorev@jetbrains.com`
- GitHub-community YaccConstructor: `https://github.com/YaccConstructor`

Example

Input graph



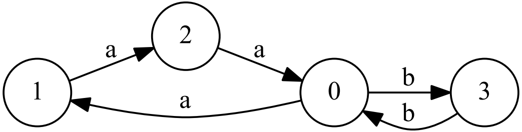
query is a grammar G which specifies the language $L = \{a^n b^n \mid n \geq 1\}$

$$
\begin{aligned}
0: &\quad S \rightarrow a\ S\ b \\
1: &\quad S \rightarrow Middle \\
2: &\quad Middle \rightarrow a\ b
\end{aligned}
$$

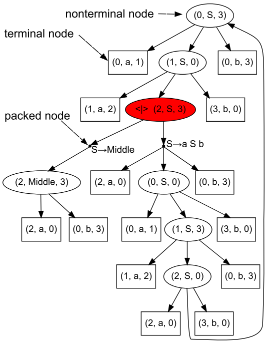
Query result is an infinite set of paths

- $p_1 = 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3$
- $p_2 = 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3 \xrightarrow{b} 0$
- ...

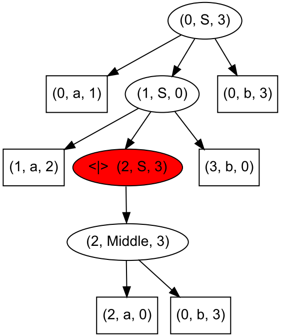
Structural representation of query result



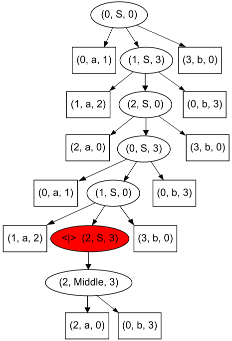
Input graph



Query result (SPPF)

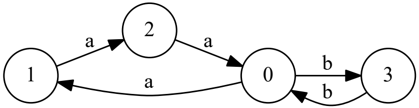


Tree for path p_1



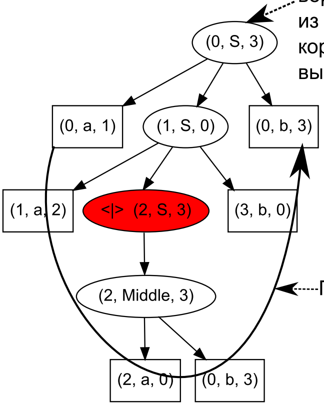
Tree for path p_2

Paths extraction



Path: $0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3$

Существует путь p из вершины 0 в вершину 3, такой w(p) что выводима из S в грамматике G. Поддерево с корнем в данной вершине - дерево вывода w(p).



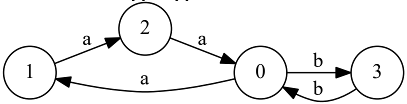
Путь - крона дерева

Key idea

Context-free languages are closed under intersection with regular languages

Key idea

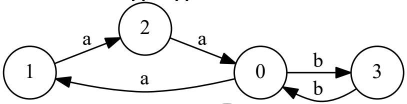
Context-free languages are closed under intersection with regular languages

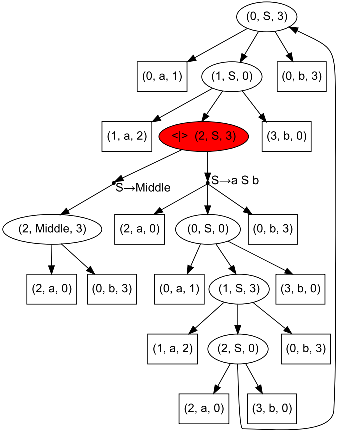


$$0: \quad S \to a\ S\ b$$
$$1: \quad S \to Middle$$
$$2: \quad Middle \to a\ b$$

Key idea

Context-free languages are closed under intersection with regular languages





$$
\begin{aligned}
0 :&\quad S \rightarrow a\ S\ b \\
1 :&\quad S \rightarrow Middle \\
2 :&\quad Middle \rightarrow a\ b
\end{aligned}
$$

$$
\begin{aligned}
(0, S, 3) &\rightarrow (0, a, 1)\ (1, S, 0)\ (0, b, 3) \\
(1, S, 0) &\rightarrow (1, a, 2)\ (2, S, 3)\ (3, b, 0) \\
(2, S, 3) &\rightarrow (2, a, 0)\ (0, S, 0)\ (0, b, 3) \\
(2, S, 3) &\rightarrow (2, Middle, 3) \\
(0, S, 0) &\rightarrow (0, a, 1)\ (1, S, 3)\ (3, b, 0) \\
(1, S, 3) &\rightarrow (1, a, 2)\ (2, S, 0)\ (0, b, 3) \\
(2, S, 0) &\rightarrow (2, a, 0)\ (0, S, 3)\ (3, b, 0) \\
(0, Middle, 3) &\rightarrow (2, a, 0)\ (0, b, 3)
\end{aligned}
$$