Parsing techniques for graph analysis

Ekaterina Verbitskaia

JetBrains Research, Programming Languages and Tools Lab
Saint Petersburg University

Oktober 22, 2017

Language-constrained paths filtering

- Can this automata generates SQL queries?
- Is there anybody one level above me in this hierarchy?

Language-constrained paths filtering: more formal

- $\mathbb{G} = (\Sigma, N, P)$ — context-free grammar
- $G = (V, E, L)$ — directed graph, $E \subseteq V \times L \times V$, $L \subseteq \Sigma$
- $p = (v_0, l_0, v_1), \cdots, (v_{n-1}, l_{n-1}, v_n)$ — path in G
- $\omega(p) = \omega((v_0, l_0, v_1), \cdots, (v_{n-1}, l_{n-1}, v_n)) = l_0 l_1 \cdots l_{n-1}$
- $R = \{p | \exists N_i \in N(\omega(p) \in L(\mathbb{G}, N_i))\}$

Applications

- Graph analysis
  - Graph database querying
  - Network graph analysis
- Code analysis
  - Static analysis CFL(linear conjunctive) reachability: alias analysis, points-to analysis, etc
  - Dynamically generated strings analysis
  - Multiple input parsing
- ...

Existing solutions

- Do not use power of advanced parsing techniques
  - Mostly based on CYK
    (Xiaowang Zhang, et al. “Context-free path queries on RDF graphs.”; Jelle Hellings. “Conjunctive context-free path queries.”)
  - Do not provide useful structural representation of result
- Have restrictions on input
  - Problems with cicles in the input graph
    (Petteri Sevon, Lauri Eronen. “Subgraph queries by context-free grammars.”)

Open problems

- Effective algorithm development
- Result representation for debugging, further processing
- GPGPU utilization

Bar-Hillel theorem

- Context-free languages are closed under intersection with regular languages
- Parsing algorithms are constructive proof of Bar-Hille theorem for one simple case ...
- ....so, calssical parsing can be generalized for arbitrary regular language processing

Example

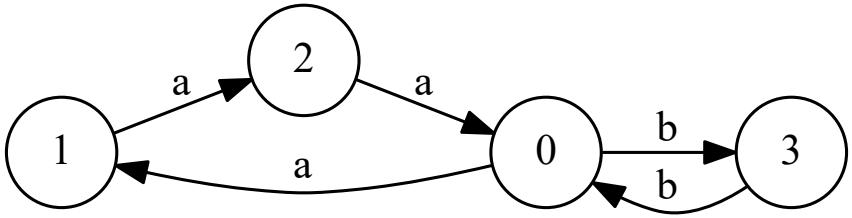


Figure: Input graph

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

Figure: Query: grammar for language $L = \{a^n b^n; n \geq 1\}$ with additional marker for the middle of a path

Example

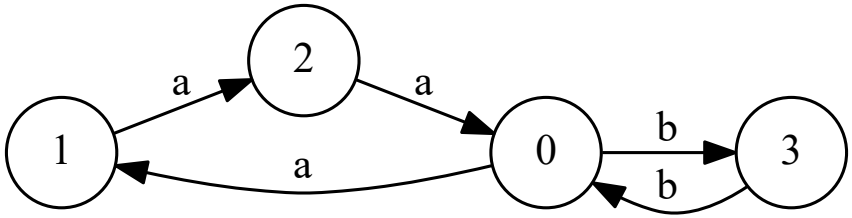
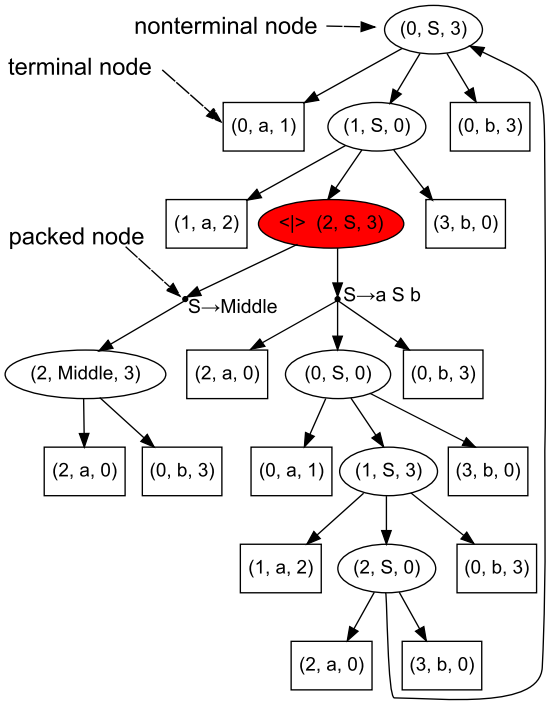
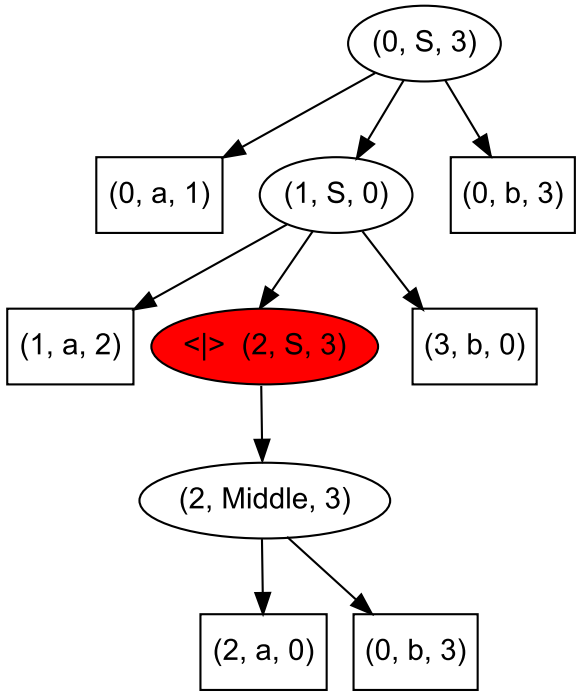


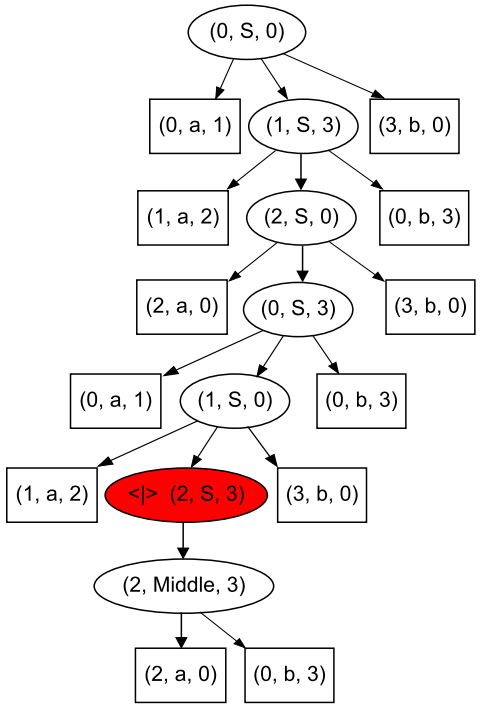
Figure: Input graph



SPPF



Tree



Tree

Our solutions

- Relaxed parsing of dynamically generated SQL-queries
  - Based on RNGLR parsing algorithm (Elizabeth Scott, Adrian Johnstone)
- Context-free path querying with structural representation of result
  - Based on GLL parsing algorithm (Elizabeth Scott, Adrian Johnstone)
- Combinators for context-free path querying
  - Based on the Meerkat: a general parser combinator library for Scala (Ali Afroozeh, Anastasia Izmaylova)
- Context-free path querying by matrix multiplication
  - Inspired by Valiant and Okhotin

Future work

- Other grammars and language classes intersection
  - Context-free grammars intersection: Mark-Jan Nederhof, "The language intersection problem for non-recursive context-free grammars"
  - Approximated intersection of regular and conjunctive/boolean languages
  - ...
- Mechanization in Coq
  - Bar-Hillel theorem
  - GLL-based algorithms
  - ...
- New areas for application

Information

- Ekaterina Verbitskaia: kajigor@gmail.com
- Semyon Grigorev: semen.grigorev@jetbrains.com

- YaccConstructor: https://github.com/YaccConstructor