Parsing Techniques for Contex-Free Path Querying

Semyon Grigorev

JetBrains Research, Programming Languages and Tools Lab
Saint Petersburg University

April 05, 2019

Formal language constrained path querying

- Finite directed edge-laballed graph $\mathcal{G} = (V, E, L)$
- The path is a world over L:
  $\omega(p) = \omega(v_0 \xrightarrow{l_0} v_1 \xrightarrow{l_1} \ldots \xrightarrow{l_{n-1}} v_n) = l_0 \cdot l_1 \cdot \ldots \cdot l_{n-1}$
- The language $\mathcal{L}$ (over L)

Formal language constrained path querying

- Finite directed edge-laballed graph $\mathcal{G} = (V, E, L)$
- The path is a world over L:

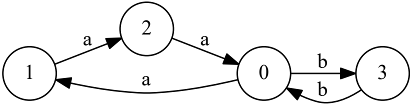
  $\omega(p) = \omega(v_0 \xrightarrow{l_0} v_1 \xrightarrow{l_1} \ldots \xrightarrow{l_{n-1}} v_n) = l_0 \cdot l_1 \cdot \ldots \cdot l_{n-1}$
- The language $\mathcal{L}$ (over L)

- Reachability problem: $Q = \{(v_i, v_j) \mid \exists p = v_i \ldots v_j, \omega(p) \in \mathcal{L}\}$
- Path querying problem: $Q = \{p \mid \omega(p) \in \mathcal{L}\}$
  - Single path, all paths, shortest path ...

Context-Free path querying

- $\mathcal{L}$ is a context-free language
- $G_{\mathcal{L}} = (N, \Sigma, R, S)$
- Reachability problem: $Q = \{(v_i, v_j) \mid \exists p = v_i \dots v_j, S \xrightarrow[G_L]{*} \omega(p)\}$
- Path querying problem: $Q = \{p \mid \omega(p) \in \mathcal{L}\}$

Example of CFPQ



Input graph

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

Query: language $\{a^n b^n \mid n > 0\}$

Paths:

$2 \xrightarrow{a} 0 \xrightarrow{b} 3$

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

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

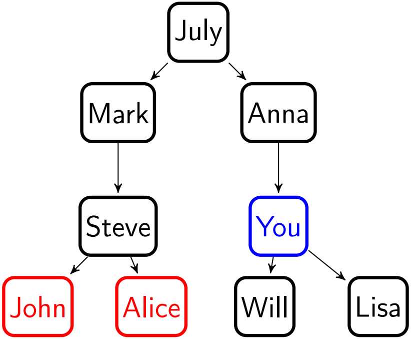
$p_2 = 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3 \xrightarrow{b} 0 \xrightarrow{b} 3 \xrightarrow{b} 0$

...

Applications

- Graph data bases querying
  Yann ...
- Static code analysis
  Reps CFL reachability
- CFL editing distance/Error recovery
  Aho

Graph data bases querying



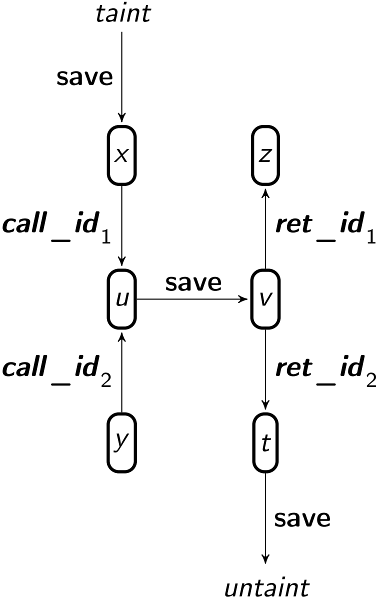
Find your cousins once removed

$$S \to H \downarrow$$
$$H \to \varepsilon \mid \uparrow H \downarrow$$

Same generation query, similarity query.

Static code analysis

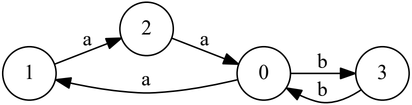
```
int id(int u)
{
  v = u;
  return v;
}
int main()
{
  //taint
  int x;
  int z, y;
  //untaint
  int t;
  z = id(x);
  t = id(y);
}
```



Error recovery

- !!!!

Structural representation of result



Input graph

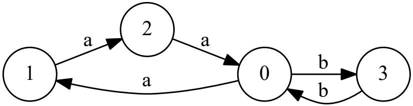
$0: \quad S \rightarrow a \ S \ b$

$1: \quad S \rightarrow Middle$

$2: \quad Middle \rightarrow a \ b$

Grammar

Structural representation of result



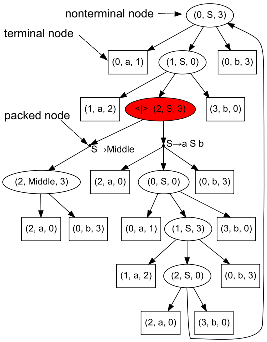
Input graph

$0: \quad S \to a\ S\ b$

$1: \quad S \to Middle$

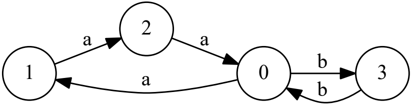
$2: \quad Middle \to a\ b$

Grammar



Query result (SPPF)

Structural representation of result



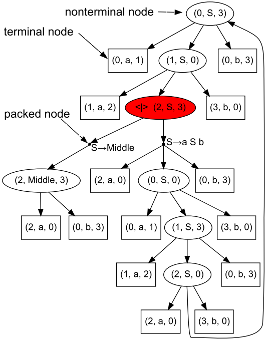
Input graph

$0: \quad S \to a\ S\ b$

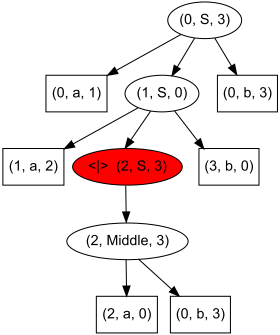
$1: \quad S \to Middle$

$2: \quad Middle \to a\ b$

Grammar

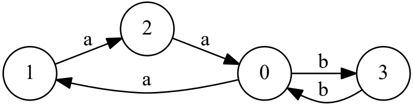


Query result (SPPF)



Tree for p_1

Structural representation of result



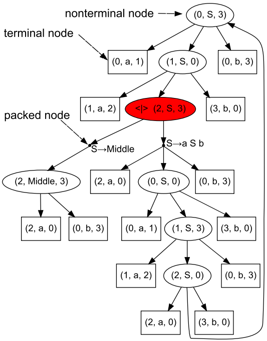
Input graph

0 : $S \to a\ S\ b$

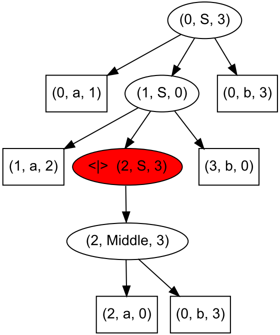
1 : $S \to Middle$

2 : $Middle \to a\ b$

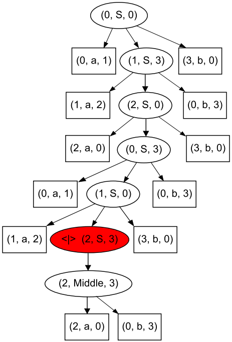
Grammar



Query result (SPPF)

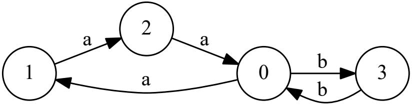


Tree for p_1

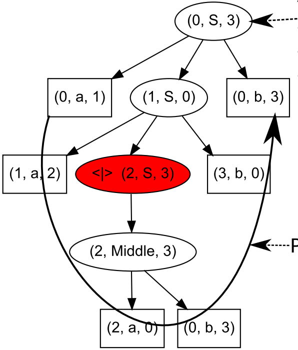


Tree for p_2

Paths extraction



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



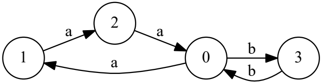
There exists a path p from vertex 0 to vetex 3 such that $\omega(p)$ is derivable from S. Subtree which is rooted by this node is a derivation tree of $\omega(p)$.

Path is a seqence of leafs

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

Bar-Hillel theorem

Context-free languages are closed under intersection with regular languages



Regular language

$0: \quad S \rightarrow a \; S \; b$

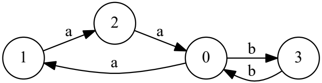
$1: \quad S \rightarrow Middle$

$2: \quad Middle \rightarrow a \; b$

Context-free language

Bar-Hillel theorem

Context-free languages are closed under intersection with regular languages



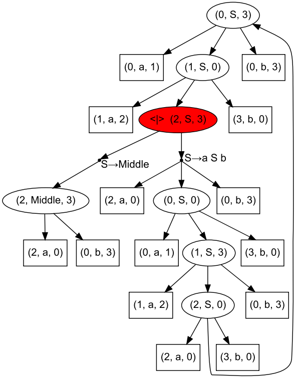
$0: \quad S \to a\ S\ b$

$1: \quad S \to Middle$

$2: \quad Middle \to a\ b$

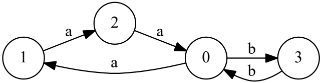
Regular language

Context-free language



Bar-Hillel theorem

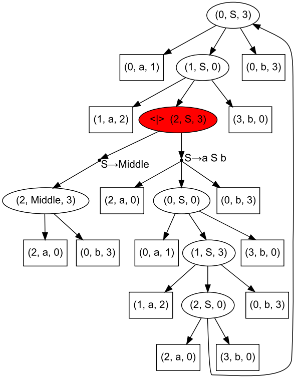
Context-free languages are closed under intersection with regular languages



Regular language

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

Context-free language



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

Directions for research

- Parallel and distributed parsing
- O(BMM) complexity
- Incremental parsing

Contact Information

- Semyon Grigorev:
    - s.v.grigoriev@spbu.ru
    - Semen.Grigorev@jetbrains.com