TITLE

A1

Saint Petersburg State University
St. Petersburg, Russia

A2

ITMO University
St. Petersburg, Russia

Rustam Azimov
rustam.azimov19021995@gmail.com
Saint Petersburg State University
St. Petersburg, Russia
JetBrains Research
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

ABSTRACT

CCS CONCEPTS

• **Information systems** → **Query languages for non-relational engines**; • **Theory of computation** → **Grammars and context-free languages**; *Parallel computing models*; • **Computing methodologies** → **Massively parallel algorithms**; • **Computer systems organization** → *Single instruction, multiple data.*

ACM Reference Format:
A1, A2, Rustam Azimov, and Semyon Grigorev. 2021. TITLE. In . ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

CFPQ as a separated algorithms.

Integration with graph DB.

Integration with query languages. The problem. We cannon separate regular and context-free queryes.

Contribution

(1) New algorithm. Correctness and time complexity.
(2) The way to optimize queries.
(3) Evaluation.

SIGMOD'21, ,

ACM ISBN XXX-X-XXXXX-XXX-X…$15.00

2 CONTEXT-FREE PATH QUERYING BY KRONECKER PRODUCT

2.1 The algorithm

In this section, we introduce the algorithm for the computation of context-free reachability in a graph $\mathcal{G}$. The algorithm determines the existence of a path, which forms a sentence of the language defined by the input RSM R, between each pair of vertices in the graph $\mathcal{G}$. The algorithm is based on the generalization of the FSM intersection for an RSM, and an input graph. Since a graph can be interpreted as a FSM, in which transitions correspond to the labeled edges between vertices of the graph, and an RSM is composed of a set of FSMs, the intersection of such machines can be computed using the classical algorithm for FSM intersection, presented in [1].

The intersection can be computed as a Kronecker product of the corresponding adjacency matrices for an RSM and a graph. Since we are only determining the reachability of vertices, it is enough to represent intersection result as a Boolean matrix. It simplifies the algorithm implementation and allows one to express it in terms of basic matrix operations.

Listing 1 shows main steps of the algorithm. The algorithm accepts context-free grammar $G = (\Sigma, N, P)$ and graph $\mathcal{G} = (V, E, L)$ as an input. An RSM R is created from the grammar G. Note, that R must have no ε-transitions. M_1 and M_2 are the adjacency matrices for the machine R and the graph $\mathcal{G}$ correspondingly.

Then for each vertex i of the graph $\mathcal{G}$, the algorithm adds loops with non-terminals, which allows deriving ε-word. Here the following rule is implied: each vertex of the graph is reachable by itself through an ε-transition. Since the machine R does not have any ε-transitions, the ε-word could

be derived only if a state s in the box B of the R is both initial and final. This data is queried by the $getNonterminals()$ function for each state s.

The algorithm terminates when the matrix M_2 stops changing. Kronecker product of matrices M_1 and M_2 is evaluated for each iteration. The result is stored in M_3 as a Boolean matrix. For the given M_3 a C_3 matrix is evaluated by the $transitiveClosure()$ function call. The M_3 could be interpreted as an adjacency matrix for an directed graph with no labels, used to evaluate transitive closure in terms of classical graph definition of this operation. Then the algorithm iterates over cells of the C_3. For the pair of indices (i, j), it computes s and f — the initial and final states in the recursive automata R which relate to the concrete $C_3[i, j]$ of the closure matrix. If the given s and f belong to the same box B of R, $s = q_B^0$, and $f \in F_B$, then $getNonterminals()$ returns the respective non-terminal. If the the condition holds then the algorithm adds the computed non-terminals to the respective cell of the adjacency matrix M_2 of the graph.

The functions $getStates$ and $getCoordinates$ (see listing 2) are used to map indices between Kronecker product arguments and the result matrix. The Implementation appeals to the blocked structure of the matrix C_3, where each block corresponds to some automata and graph edge.

The algorithm returns the updated matrix M_2 which contains the initial graph $\mathcal{G}$ data as well as non-terminals from N. If a cell $M_2[i, j]$ for any valid indices i and j contains symbol $S \in N$, then vertex j is reachable from vertex i in grammar G for non-terminal S.

Listing 1 Kronecker product based CFPQ

```
1:  function contextFreePathQuerying(G, 𝒢)
2:      R ← Recursive automata for G
3:      M1 ← Adjacency matrix for R
4:      M2 ← Adjacency matrix for 𝒢
5:      for s ∈ 0..dim(M1) − 1 do
6:          for i ∈ 0..dim(M2) − 1 do
7:              M2[i, i] ← M2[i, i] ∪ getNonterminals(R, s, s)
8:      while Matrix M2 is changing do
9:          M3 ← M1 ⊗ M2                          ▷ Evaluate Kroncker product
10:         C3 ← transitiveClosure(M3)
11:         n ← dim(M3)                           ▷ Matrix M3 size = n × n
12:         for (i, j) ∈ [0..n − 1] × [0..n − 1] do
13:             if C3[i, j] then
14:                 s, f ← getStates(C3, i, j)
15:                 if getNonterminals(R, s, f) ≠ ∅ then
16:                     x, y ← getCoordinates(C3, i, j)
17:                     M2[x, y] ← M2[x, y] ∪ getNonterminals(R, s, f)
18:     return M2
```

Listing 2 Help functions for Kronecker product based CFPQ

```
1:  function getStates(C, i, j)
2:      r ← dim(M1)                               ▷ M1 is adjacency matrix for automata R
3:      return ⌊i/r⌋, ⌊j/r⌋
4:  function getCoordinates(C, i, j)
5:      n ← dim(M2)                               ▷ M2 is adjacency matrix for graph 𝒢
6:      return i mod n, j mod n
```

Lemma 2.1. *Let $\mathcal{G} = (V, E, L)$ be a graph and $G = (\Sigma, N, P)$ be a grammar. Let $\mathcal{G}_k = (V, E_k, L \cup N)$ be graph and M_k its adjacency matrix of the execution some iteration $k \geq 0$ of the algorithm **??**. Then for each edge $e = (m, S, n) \in E_k$, where $S \in N$, the following statement holds: $\exists m\pi n : S \rightarrow_G l(\pi)$.*

Proof. (Proof by induction)

Basis: For $k = 0$ and the statement of the lemma holds, since $M_0 = M$, M where is adjacency matrix of the graph G. Non-terminals, which allow to derive ε-word, are also added at algorithm preprocessing step, since each vertex of the graph is reachable by itself through an ε-transition.

Inductive step: Assume that the statement of the lemma holds for any $k \leq (p - 1)$ and show that it also holds for $k = p$, where $p \geq 1$.

For the algorithm iteration p the Kronecker product K_p and transitive closure C_p are evaluated as described in the algorithm. By the properties of this operations, some edge $e = ((s, m), (f, n))$ exists in the directed graph, represented by adjacency matrix C_p, if and only if $\exists s\pi' f$ in the RSM graph, represented by matrix M_r, and $\exists m\pi n$ in graph, represented by M_{p-1}. Concatenated symbols along the path π' form some derivation string v, composed from terminals and non-terminals, where $v \rightarrow_G l(\pi)$ by the inductive assumption.

The new edge $e = (m, S, n)$ will be added to the E_p only if s and f are initial and final states of some box B of the RSM corresponding to the non-terminal S_B. In this case, the grammar G has the derivation rule $S_B \rightarrow_G v$, by the inductive assumption $v \rightarrow_G l(\pi)$. Therefore, $S_B \rightarrow_G l(\pi)$ and this completes the proof of the lemma.

□

Lemma 2.2. *Let $\mathcal{G} = (V, E, L)$ be a graph and $G = (\Sigma, N, P)$ be a grammar. Let $\mathcal{G}_k = (V, E_k, L \cup N)$ be graph and M_k its adjacency matrix of the execution some iteration $k \geq 1$ of the algorithm **??**. For any path $m\pi n$ in graph $\mathcal{G}$ with word $l = l(\pi)$ if exists the derivation tree of l for the grammar G and starting non-terminal S with the height $h \leq k$, then $\exists e = (m, S, n) : e \in E_k$.*

Proof. (Proof by induction)

Basis: Show that statement of the lemma holds for the $k = 1$. Matrix M and edges of the graph $\mathcal{G}$ contains only labels from L. Since the derivation tree of height $h = 1$ contains only one non-terminal S as a root and only symbols from $\Sigma \cup \varepsilon$ as leafs, for all paths, which form a word with derivation tree of the height $h = 1$, the corresponding nonterminals will be added to the M_1 via preprocessing step and first iteration of the algorithm. Thus, the lemma statement holds for the $k = 1$.

Inductive step: Assume that the statement of the lemma hold for any $k \leq (p-1)$ and show that it also holds for $k = p$, where $p \geq 2$.

For the algorithm iteration p the Kronecker product K_p and transitive closure C_p are evaluated as described in the algorithm. By the properties of this operations, some edge $e = ((s, m), (f, n))$ exists in the directed graph, represented by adjacency matrix C_p, if and only if $\exists s \pi_1 f$ in the RSM graph, represented by matrix M_{RSM}, and $\exists m \pi n$ in graph, represented by M_{p-1}.

For any path $m\pi n$, such that exist derivation tree of height $h < k$ for the word $l(\pi)$ with root non-terminal S, there exists edge $e = (m, S, n) : e \in E_k$ by inductive assumption.

Suppose, that exists derivation tree T of height $h = p$ with the root non-terminal S for the path $m\pi n$. The tree T is formed as $S \rightarrow a_1..a_d, d \geq 1$ where $\forall i \in [1..d]$ a_i is sub-tree of height $h_i \leq p - 1$ for the sub-path $m_i \pi_i n_i$. By inductive hypothesis, there exists path π_i for each derivation sub-tree, such that $m = m_1 \pi_1 m_2 .. m_d \pi_d m_{d+1} = n$ and concatenation of these paths forms $m\pi n$, and the root non-terminals of this sub-trees are included in the matrix M_{p-1}.

Therefore, vertices m_i $\forall i \in [1..d]$ form path in the graph, represented by matrix M_{p-1}, with complete set of labels. Thus, new edge between vertices m and n with the respective non-terminal S will be added to the matrix M_p and this completes the proof of the lemma.

□

Theorem 2.3. *Let $\mathcal{G} = (V, E, L)$ be a graph and $G = (\Sigma, N, P)$ be a grammar. Let $\mathcal{G}_R = (V, E_R, L)$ be a result graph for the execution of the algorithm* **??**. *The following statement holds: $e = (m, S, n) \in E_R$, where $S \in N$, if and only if $\exists m\pi n : S \rightarrow_G l(\pi)$.*

Proof. This theorem is a consequence of the Lemma 2.1 and Lemma 2.2.

□

Theorem 2.4. *Let $\mathcal{G} = (V, E, L)$ be a graph and $G = (\Sigma, N, P)$ be a grammar. The algorithm* **??** *terminates in finite number of steps.*

Proof. The main algorithm *while-loop* is executed while graph adjacency matrix M is changing. Since the algorithm only adds the edges with non-terminals from N, the maximum required number of iterations is $|N| \times |V| \times |V|$, where each component has finite size. This completes the proof of the theorem.

□

3 EVALUATION

Questions.

(1) Compare classical RPQ algorithms and our agorithm
(2) Compare other CFPQ algorithms and our algorithms
(3) Iveatigate effect of grammar optimization

3.1 RPQ

3.2 CFPQ

Comparison with matrix-based.

On query optimization.

4 CONCLUSION

REFERENCES

[1] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2006. *Introduction to Automata Theory, Languages, and Computation (3rd Edition).* Addison-Wesley Longman Publishing Co., Inc., USA.