Синтаксический анализ графов через умножение матриц

В рамках проекта лаборатории JetBrains

Автор: Рустам Шухратуллович Азимов, 646 группа
Научный руководитель: к.ф.-м.н., доцент С.В. Григорьев
Рецензент: преподаватель, научный координатор Академии Або и Центра Компьютерных Наук TUCS М.Л. Бараш

Санкт-Петербургский Государственный Университет
Кафедра системного программирования

24 мая 2018г.

Синтаксический анализ графов

- Вход:
  - Ориентированный граф D = (V,E) с метками на ребрах из алфавита Σ
  - Формальная грамматика (запрос к графу) $G = (\Sigma, N, P)$ над тем же алфавитом
- Выход для реляционной семантики запросов:
  - Множество всех троек (A, m, n), где существует путь из вершины m в вершину n, метки на ребрах которого образуют строку, выводимую из нетерминала A
- Выход для single-path семантики запросов:
  - Дополнительно предоставить один такой путь для каждой тройки (A, m, n)

Пример

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

Рис.: Пример входной КС-грамматики

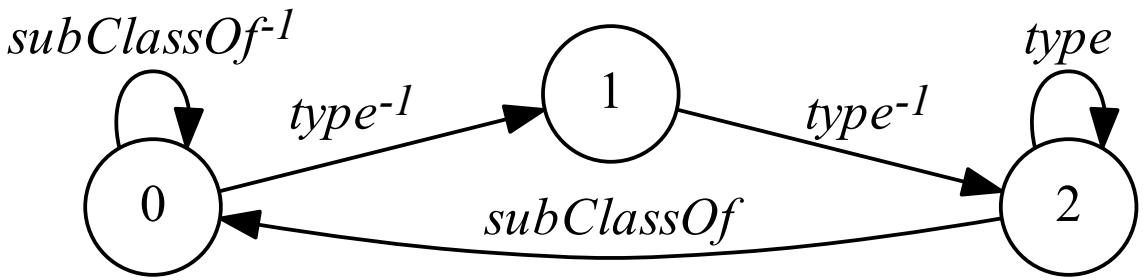


Рис.: Пример входного графа

Применимость

- Запросы к графовым базам данных
- Анализ RDF файлов
- Биоинформатика
- Статический анализ программ

Существующие алгоритмы

- Реляционная семантика запросов для KC-грамматик
  - Основанный на CYK (J. Hellings, 2014)
  - Тот же алгоритм, реализованный для анализа RDF файлов (X. Zhang, Z. Feng, X. Wang et al., 2016)
- Single-path семантика запросов для KC-грамматик
  - Основанный на методе динамического программирования (J. Hellings, 2015)
  - Основанный на GLL (Григорьев Семен, Рогозина Анастасия, 2016)
- Реляционная семантика запросов для конъюнктивных грамматик
  - Приближенное решение для статического анализа программ (Q. Zhang, Z. Su, 2017)

Проблемы

- Низкая производительность на больших графах
- Существующие алгоритмы не позволяют эффективно применить такие техники, как вычисление на графическом процессоре, параллельное вычисление
- Возможность создания матричного алгоритма синтаксического анализа графов является открытой проблемой
- Для конъюнктивных грамматик: не существует алгоритма, работающего с произвольной конъюнктивной грамматикой (есть алгоритм, работающий с линейными конъюнктивными грамматиками)

Постановка задачи

Цель: Разработать матричный алгоритм синтаксического анализа графов

Задачи:

- Разработать матричный алгоритм синтаксического анализа графов для KC-грамматик и реляционной семантики запросов
- Разработать матричный алгоритм синтаксического анализа графов для KC-грамматик и single-path семантики запросов
- Разработать матричный алгоритм синтаксического анализа графов для произвольных конъюнктивных грамматик и реляционной семантики запросов
- Показать практическую применимость предложенных алгоритмов на общепринятом наборе данных

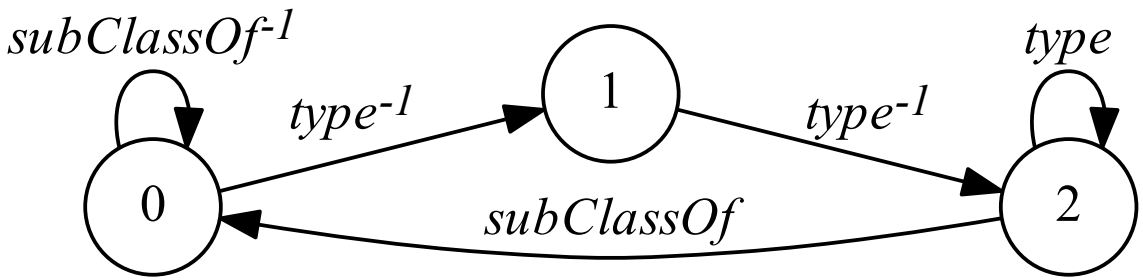
Ход работы

- Предложены алгоритмы синтаксического анализа графов, вычисляющие матричное транзитивное замыкание
- Доказана корректность предложенных алгоритмов
- Алгоритмы реализованы с использованием комбинаций таких оптимизаций, как:
  - Разреженное представление матриц
  - Умножение матриц на графическом процессоре
  - Параллельное умножение матриц
- Проведена апробация на общепринятом наборе RDF файлов

Пример работы алгоритма для КС-грамматик и реляционной семантики запросов

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

Рис.: Пример входной грамматики



Пример: Грамматика в нормальной форме

$$
\begin{array}{lrcl}
0: & S & \to & S_1\ S_5 \\
1: & S & \to & S_3\ S_6 \\
2: & S & \to & S_1\ S_2 \\
3: & S & \to & S_3\ S_4 \\
4: & S_5 & \to & S\ S_2 \\
5: & S_6 & \to & S\ S_4 \\
6: & S_1 & \to & subClassOf^{-1} \\
7: & S_2 & \to & subClassOf \\
8: & S_3 & \to & type^{-1} \\
9: & S_4 & \to & type
\end{array}
$$

Рис.: Входная грамматика в нормальной форме Хомского

Пример: Начальная матрица и первая итерация

$$T_0 = \begin{pmatrix} \{S_1\} & \{S_3\} & \varnothing \\ \varnothing & \varnothing & \{S_3\} \\ \{S_2\} & \varnothing & \{S_4\} \end{pmatrix}$$

Рис.: Начальная матрица

$$T_0 \cdot T_0 = \begin{pmatrix} \varnothing & \varnothing & \varnothing \\ \varnothing & \varnothing & \{S\} \\ \varnothing & \varnothing & \varnothing \end{pmatrix}$$

$$T_1 = T_0 \cup (T_0 \cdot T_0) = \begin{pmatrix} \{S_1\} & \{S_3\} & \varnothing \\ \varnothing & \varnothing & \{S_3, S\} \\ \{S_2\} & \varnothing & \{S_4\} \end{pmatrix}$$

Пример: Остальные итерации

$$T_2 = \begin{pmatrix} \{S_1\} & \{S_3\} & \varnothing \\ \{S_5\} & \varnothing & \{S_3, S, S_6\} \\ \{S_2\} & \varnothing & \{S_4\} \end{pmatrix}$$

$$T_3 = \begin{pmatrix} \{S_1\} & \{S_3\} & \{S\} \\ \{S_5\} & \varnothing & \{S_3, S, S_6\} \\ \{S_2\} & \varnothing & \{S_4\} \end{pmatrix}$$

$$T_4 = \begin{pmatrix} \{S_1, S_5\} & \{S_3\} & \{S, S_6\} \\ \{S_5\} & \varnothing & \{S_3, S, S_6\} \\ \{S_2\} & \varnothing & \{S_4\} \end{pmatrix}$$

$$T_5 = \begin{pmatrix} \{S_1, S_5, S\} & \{S_3\} & \{S, S_6\} \\ \{S_5\} & \varnothing & \{S_3, S, S_6\} \\ \{S_2\} & \varnothing & \{S_4\} \end{pmatrix}$$

Пример: Результирующие отношения по матрице $T_6 = T_5$

$$
\begin{aligned}
R_S &= \{(0,0),(0,2),(1,2)\},\\
R_{S_1} &= \{(0,0)\},\\
R_{S_2} &= \{(2,0)\},\\
R_{S_3} &= \{(0,1),(1,2)\},\\
R_{S_4} &= \{(2,2)\},\\
R_{S_5} &= \{(0,0),(1,0)\},\\
R_{S_6} &= \{(0,2),(1,2)\}.
\end{aligned}
$$

Рис.: Результирующие КС-отношения

Апробация: Запрос

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

Рис.: Грамматика для запроса

Апробация: результаты

Ontology	edgs	result	GLL	dGPU	sCPU	sGPU
skos	252	810	10	56	14	12
generations	273	2164	19	62	20	13
travel	277	2499	24	69	22	30
univ-bench	293	2540	25	81	25	15
atom-primitive	425	15454	255	190	92	22
biomedical	459	15156	261	266	113	20
foaf	631	4118	39	154	48	9
people-pets	640	9472	89	392	142	32
funding	1086	17634	212	1410	447	36
wine	1839	66572	819	2047	797	54
pizza	1980	56195	697	1104	430	24
g_1	8688	141072	1926	—	26957	82
g_2	14712	532576	6246	—	46809	185
g_3	15840	449560	7014	—	24967	127

Результаты

- Разработан матричный алгоритм синтаксического анализа графов для KC-грамматик и реляционной семантики запросов
- Разработан матричный алгоритм синтаксического анализа графов для KC-грамматик и single-path семантики запросов
- Разработан матричный алгоритм синтаксического анализа графов для произвольных конъюнктивных грамматик и реляционной семантики запросов
- Показана практическая применимость предложенных алгоритмов на общепринятом наборе данных