



Синтаксический анализ графов и задача генерации строк с ограничениями

Докладчик: Рустам Азимов

Лаборатория языковых инструментов JetBrains
Санкт-Петербургский государственный университет
Математико-механический факультет

4 апреля 2017г.

- **Входной граф:**
 - ▶ $D = (V, E)$ — помеченный граф с метками на ребрах из Σ
- **Запрос к входному графу:**
 - ▶ C — формальная грамматика, порождающая язык $L(C) \subseteq \Sigma^*$
- **Результат запроса:**
 - ▶ Некоторая информация о путях в графе, метки на ребрах которых образуют строку $w \in L(C)$
- **Пример области применения:**
 - ▶ Статический анализ динамически формируемого кода (динамические SQL-запросы, генераторы Web-страниц)
 - ▶ Графом в данном случае является регулярная аппроксимация множества возможных значений динамически формируемых строк

- Регулярные
- Контекстно-свободные:
 - ▶ $C = (N, \Sigma, P)$
 - ▶ Для стартового нетерминала $a \in N$ порождается язык $L(C, a)$
- Конъюнктивные:
 - ▶ $C = (N, \Sigma, P)$
 - ▶ Правила грамматики имеют вид $a \rightarrow \alpha_1 \& \dots \& \alpha_n$, где $a \in N$, $\alpha_i \in (\Sigma \cup N)^*$, $n \geq 1$
 - ▶ Для стартового нетерминала $a \in N$ порождается язык $L(C, a)$

- **Реляционная:**

- ▶ Для всех $a \in N$ вычислить $\{(m, n) \mid L(C, a) \cap L(D, m, n) \neq \emptyset\}$

- **All-path:**

- ▶ Для всех $a \in N$ и $m, n \in V$ предъявить все пути из вершины m в n , такие что метки на ребрах этих путей образуют строку $w \in L(C, a)$

- **Single-path:**

- ▶ Для всех $a \in N$ и $m, n \in V$ предъявить какой-нибудь путь (если он существует) из вершины m в n , такие что метки на ребрах этих путей образуют строку $w \in L(C, a)$

Пример

- Граф D — коллекция генеалогических деревьев
- Вершины графа — люди
- Ребра представляют отношение между родителями и детьми (*parentOf* или *childOf*)
- КС-грамматика C , порождающая язык $L(C) = \{parentOf^n childOf^n | n > 0\}$
- Пути, соответствующие языку $L(C)$, соединяют потомков общего предка из одного поколения
- Находить данные пути можно с помощью **генератора строк** КС-языка $L(C) \cap L(D)$

Существующие работы

- Алгоритм синтаксического анализа графов для КС-грамматик и реляционной семантики запросов (Hellings, 2014)
 - ▶ Основан на CYK алгоритме
- Алгоритм синтаксического анализа графов для КС-грамматики и All-path, Single-path семантик запросов (Hellings, 2016)
 - ▶ Строит некоторую КС-грамматику C_D , порождающую язык $L(C) \cap L(D)$
 - ▶ Грамматика C_D является компактным представлением ответа на запрос с **all-path** семантикой
 - ▶ Если язык $L(C_D) \neq \emptyset$, то в нем находится строка минимальной длины, которая и будет ответом на запрос с **single-path** семантикой
- Алгоритмы для задач генерации строк с ограничениями с использованием КС и конъюнктивных грамматик (Охотин, 2003)

Цель: Исследование связи между задачей генерации строк с ограничениями и задачами синтаксического анализа графов, использующих реляционную, *all-path* и *single-path* семантики запросов

Задачи:

- Исследовать связь между данными задачами для КС-грамматик
- Исследовать связь между данными задачами для конъюнктивных грамматик

Использование генератора строк с ограничениями

- Для использования генератора строк с ограничениями необходимо иметь:
 - ▶ Формально определенную входную грамматику
 - ▶ Возможность проверить непустоту языка L , порождаемого входной грамматикой
 - ▶ Оценку сверху на минимальную длину строки из языка L

- Входная грамматика — C_D , порождает КС-язык $L(D) \cap L(C)$
- Задача проверки пустоты КС-языка — разрешима
- Задача оценки сверху на минимальную длину строки из языка, порожденного входной грамматикой — разрешима

Связь с синтаксическим анализом графов для КС-грамматик

- Задача синтаксического анализа графов сводится к задаче генерации строк языка $L(C_D)$
- ***All-path* семантика**: построение грамматики C_D автоматически решает задачу синтаксического анализа графов
 - ▶ Генератор может применяться для последовательного вывода строк из языка $L(C_D)$
- **Реляционная семантика**: проверка пустоты языка $L(C_D)$ автоматически решает задачу синтаксического анализа графов
- ***Single-path* семантика**: алгоритмы синтаксического анализа графов и генерации строк эквивалентны:
 - ▶ Проверяется непустота языка $L(C_D)$
 - ▶ Находится строка минимальной длины из языка $L(C_D)$

- Входная грамматика строится с помощью явной операции $\&$
- Задача проверки пустоты конъюнктивного языка — неразрешима
- Задача оценки сверху на минимальную длину строки из языка, порожденного входной грамматикой — неразрешима

Связь с синтаксическим анализом графов для конъюнктивных грамматик

Из неразрешимости задачи проверки пустоты конъюнктивного языка следует неразрешимость задачи генерации строк с ограничениями

- Неразрешимы задачи синтаксического анализа графов для конъюнктивных грамматик
- ***All-path*** семантика: применение генератора приводит к перебору всех возможных строк, что не соответствует практическому смыслу задачи
- ***Single-path*** семантика: применение генератора возможно, только если заранее гарантируется непустота языка, порождаемого входной грамматикой

- Показана связь между задачей генерации строк с ограничениями и задачами синтаксического анализа графов, использующих реляционную, *all-path* и *single-path* семантики запросов:
 - ▶ Сведение задачи синтаксического анализа графов к задаче генерации строк для КС-грамматик
 - ▶ Неразрешимость задачи синтаксического анализа графов для конъюнктивных грамматик
- Полученные результаты могут быть использованы в дальнейших исследованиях рассматриваемых областей