Абстрактный лексический анализ.

В рамках проекта лаборатории JetBrains "Синтаксический и семантический анализ встроенных языков"

Автор: студентка 344 группы Вербицкая Екатерина
Научный руководитель: Григорьев Семён

2013 год

Встроенные языки

- Встроенный SQL
  - Динамический SQL
- HTML, XML
- Текстовые встроенные DSL

```
IF @X = @Y
    SET @TABLE = '#table1'
ELSE
    SET @TABLE = 'table2'
SET @S = 'SELECT x FROM' + @TABLE + ' WHERE ISNULL(n,0) > 1'
EXECUTE (@S)
```

Проблема

- Код на встроенных языках — безжизненные строки, не подлежащие анализу стандартными средствами.
- Они не проверяются на наличие ошибок статически.
  - Ошибки "всплывут" на этапе выполнения.
- Нужны средства для проверки корректности.
  - Синтаксическая корректность.
  - Проверка типов.
- Полезными будут средства, упрощающие разработку:
  - Автодополнение.
  - Рефакторинги.

Цели

- Создание платформы для работы со встроенными языками
  - Выделение строк
  - **Лексический анализ**
  - Синтаксический анализ
  - Семантический анализ

Задачи

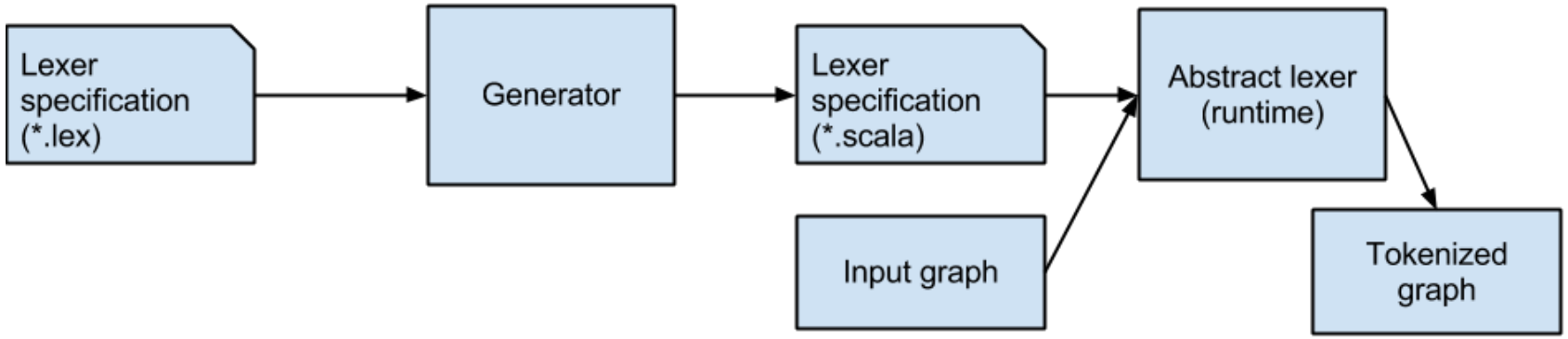
- Создание генератора абстрактных лексических анализаторов.
  - Продумать архитектуру так, чтобы это было абстрактное решение в смысле переиспользования в составе платформы.
- Апробация на примере языка T-SQL (встроенный SQL).

Существующие инструменты

- Alvor
  - Плагин для eclipse, статически анализирующий встроенный в Java SQL код.
- Java string analyzer
  - Инструмент для Java-программ, предназначенный для анализа строковых выражений. Для каждого строкового выражения генерируется автомат, являющий собой аппроксимацию значений, которые могут быть получены на этапе исполнения.

Архитектура решения

Абстрактный лексер генерируется на основе спецификации грамматики, за счет чего достигается универсальность платформы.



Генератор лексических анализаторов

- Построен на основе JFlex.
- Реализована генерация управляющих данных (описание конечного автомата, пользовательский код) в язык программирования Scala.
- Функция токенизации не генерируется, а реализована статически и параметризуется сгенерированными данными.

Процесс токенизации

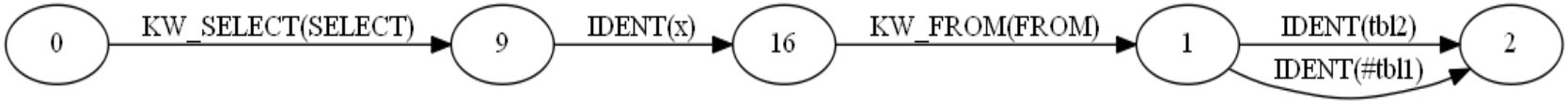
- Входной граф является компактным представлением множества значений динамически формируемого выражения.
- Входной граф преобразуется, чтобы на каждом ребре присутствовал лишь один символ.
- В основе — finite state transducer.
- Результатом является граф, собранный из токенов, встречающихся в строках, содержащих встроенный код.

Пример работы

```
IF @X = @Y
  SET @TABLE = '#tbl1'
ELSE
  SET @TABLE = 'tbl2'
SET @S = 'SELECT x FROM ' + @TABLE
EXECUTE (@S)
```

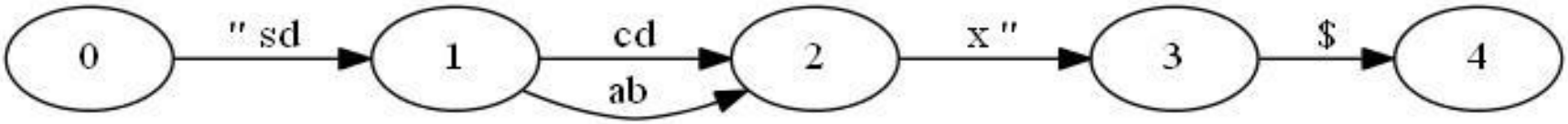
Входной граф:

Результат работы:

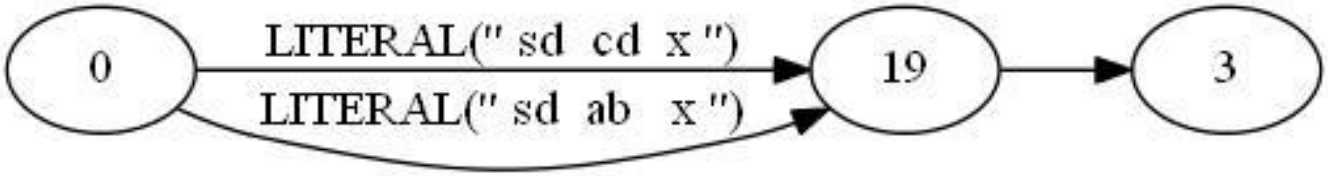


Пример работы: "рваные" литералы

Входной граф:



Результат работы:



Результаты

- Реализован генератор абстрактных лексических анализаторов.
- Проведена апробация на примере T-SQL.
- Работа представлялась на конференции СПИСОК-2013 и получила рекомендацию к публикации. Тезисы опубликованы в сборнике материалов конференции.