

TITLE OF THE DOCUMENT

MASTERS THESIS

VON

FIRSTNAME LASTNAME

BORN ON THE 29TH OF FEBRUARY 1974 IN TOWN

20 АПРЕЛЯ 2009 Г.

SUPERVISORS:

PROF. V. NACHNAME

DR. V. NACHNAME

TOWN UNIVERSITY

FACULTY OF ...

INSTITUTE OF ...

Содержание

1	Введение.	2
2	Обзор.	2
3	Реализация.	3
3.1	Внутреннее представление.	3
3.2	Рекурсивно восходящий алгоритм.	4
3.3	Расширенные контекстно свободные грамматики.	5
3.3.1	Построение ДКА	5
3.3.2	Построение множества ситуаций	6
3.3.3	Вычисление GOTO	7
3.4	Построение деревьев вывода.	7
3.5	Заключение.	9

1 Введение.

Задачи автоматизированного реинжиниринга программ выдвигают особые требования к генераторам синтаксических анализаторов.

Для устаревшего языка сложно (а зачастую и невозможно) задать однозначную контекстно-свободную грамматику. Необходимо существенно преобразовать его спецификацию, которая приводится в документации, чтобы получить такую грамматику, но при этом она перестает быть сопровождаемой [1]. Поэтому устаревший язык обычно задается с помощью неоднозначной контекстно-свободной грамматики.

При поддержке нескольких диалектов языка необходима возможность лёгкой трансформации грамматики. Однако, зачастую, изменение одного правила приводит к появлению десятков конфликтов в грамматике [1], которые необходимо разрешать вручную. Это требует большого количества времени.

Как вариант решения этих задач в статье [1] предлагается использовать GLR грамматики и соответствующие инструменты построения анализаторов. Действительно, GLR-алгоритм разрешает неоднозначности в грамматике на уровне концепции. По этому задание спецификации трансляции становится проще, требует меньше времени. Получившийся код компактнее и сопровождаемое.

Главным достоинством GLR-алгоритма является обработка неоднозначных грамматик. Анализатор, построенный с помощью данного алгоритма, в результате разбора строит не единственное дерево, а несколько деревьев - лес, который можно сократить, используя специальные фильтры, а можно, при задании в одной спецификации нескольких диалектов, вернуть весь лес для дальнейшего выбора нужного дерева/диалекта.

Стоит отметить, что по производительности такой анализатор, являясь некоторой "надстройкой" над LR-анализатором, незначительно ему уступает. На сегодняшний день в соотношении производительность/класс разбираемых языков GLR-алгоритм выглядит наиболее предпочтительно.

2 Обзор.

Стоит задача разработки анализатора для произвольной контекстно свободных грамматик на платформе .NET и реализации его на функциональном языке программирования F#. Необходимость такой разработки была обоснована выше.

Автоматизированный реинжиниринг программ выдвигает особые требования к инструменту анализа и языку спецификаций трансляции [3]. Используемые в разрабатываемом инструменте грамматики YARD удовлетворяют основным таким требованиям. Поэтому наибольший интерес сейчас представляет анализ внутренних алгоритмов анализа в существующих инструментах.

Предпочтительным алгоритмам анализа является алгоритм разбора произвольной контекстно свободной грамматики. Поэтому нужно рассмотреть инструменты, основанные на этом алгоритме. Важен способ реализации алгоритма, так как существуют несколько альтернатив: алгоритм Томиты (GLR-алгоритм), алгоритм Эрли (Early), рекурсивно-восходящий алгоритм. Так-же следует обратить преобразования грамматики, необходимые для построения анализатора.

В настоящее время существуют следующие инструменты основанные на GLR-алгоритме.

- ASF+SDF [10] (Algebraic Specification Formalism + Syntax Definition Formalism) - генератор с широкими возможностями, но достаточно сложным входным языком. Является SGLR-инструментом (Scannerless, Generalized-LR).
- Bison [11] - развитие инструмента YACC. Все грамматики, созданные для оригинального YACC, будут работать и в Bison. Является одним из самых популярных и совершенных

"потомков"YACC. При включении соответствующей опции использует GLR-алгоритм (по умолчанию LALR).

- Elkhound [12] - позиционируется как быстрый и удобный GLR-инструмент, созданный в университете Беркли (США), тем не менее обладает достаточно "бедным"входным языком (например, он не поддерживает конструкций расширенной формы Бэкуса-Наура).

В работе [3] проведён подробный анализ этих инструментов. Подводя итог, можно сказать, что на текущий момент нет инструмента, полностью удовлетворяющего требованиям автоматизированного реинжиниринга программ.

Так же нас будет интересовать такой инструмент как Jade, являющийся генератором рекурсивно-восходящих парсеров.

Jade это генератор рекурсивно-восходящих LALR(1) парсеров с целевым языком C. Его подробное описание приводится в статье [6]. При реализации данного инструмента появилась проблема объёма кода целевого парсера. Так как при построении детерминированного парсера необходимо генерировать процедуры для каждого состояния, то объём кода быстро растёт, с ростом количества правил в грамматике. Так для языка Java, по расчётам, приведённым в статье[6], объём кода составляет примерно 4 мегабайта. В Jade эта проблема решается путём создания глобальной структуры(массива состояний), где хранится информация, позволяющая переиспользовать процедуры. (Подробнее об этом можно прочесть в статье [6]).

3 Реализация.

Платформой для создания инструмента выбран .NET . Основным языком реализации является F# [2].

Общая схема реализуемого алгоритма такова:

- На вход получаем грамматику в виде списка правил. Одно правило - специальная структуры данных, описанная ниже.
- Нумеруем правила по порядку.
- Построение множества LR-ситуаций
- Вычисление функции GOTO
- Построение деревьев вывода.

3.1 Внутреннее представление.

За основу внутреннего представления грамматики взято представление инструмента YARD. Это представление и ,соответственно, входной язык инструмента, содержит такие конструкции, как

- Конструкции расширенной формы Бэкуса-Наура;
- Макроправила (параметризация одних правил другими);
- Сгруппированные альтернативы;
- Предикаты;

3.2 Рекурсивно восходящий алгоритм.

Существует альтернатива табличным анализаторам - рекурсивно-восходящие анализаторы. Идея состоит в том, чтобы не использовать стек явно, а заменить его стеком вызова рекурсивных функций и эмулировать поведения автомата вызовом функций. Для этого можно построить функцию соответствующую каждому состоянию, как это делается в статье [8]. В этом случае можно получить детерминированный LR анализ.

Таким образом - рекурсивно восходящий алгоритм это аналог рекурсивного спуска, но для LR грамматики. Однако, как показано в статье [6], при такой реализации возникают проблемы с быстрым ростом объёма кода целевых функций.

В статье [9] показан подход к решению этой задачи несколько с другой стороны. Показано, что можно построить функции, оперирующие уже не одним состоянием, а множеством состояний, и свести ,таким образом, количество необходимых функций к двум взаимно рекурсивным:

- $parse\ q\ i = \{(A \rightarrow a., i) | A \rightarrow a. \in q\} \cup$
 $\{(A \rightarrow a.b, k) | i = xj, (A \rightarrow a.b, k) \in climb\ q\ x\ j\} \cup$
 $\{(A \rightarrow a.b, k) | B \rightarrow e, (A \rightarrow a.b, k) \in climb\ q\ B\ j\}$
- $climb\ q\ X\ i = \{(A \rightarrow a.Xb, k) | (A \rightarrow a.Xb, k) \in parse(goto\ q\ X)i, a \neq e, A \rightarrow a.Xb \in q\} \cup$
 $\{(A \rightarrow a.b, l) | (C \rightarrow X.c, j) \in parse(goto\ q\ X)i, (A \rightarrow a.b, l) \in climb\ q\ C\ j\}$

Определим также функцию:

- $goto\ q\ X = \{A \rightarrow a.Xb | A \rightarrow a.Xb \in q^*\},$
 где
- $q^* = q \cup \{B \rightarrow .c | A \rightarrow a.Bb \in q^*\} \cup \{x \rightarrow .x | A \rightarrow a.xb \in q^*\}$

Пользуясь таким определением можно построить функции из статьи[8], однако сейчас интереснее их реализация, согласно определению.

Важно, что при решении нашей задачи нет необходимости генерации функции для каждого состояния. Так как цель - получение недетерминированного разбора, то можно реализовать всего две функции: *parse* и *climb*, определённые выше. Это позволит решить проблему объёма кода, возникшую в Jade. Объём кода резко сократится и, более того, будет получен недетерминированный вариант разбора.

Используя рекурсивно-восходящий алгоритм можно получить эквивалент алгоритму Томиты. При использовать описанных выше функции автоматически получается ветвление в момент возникновения неоднозначности (разделение стека в алгоритме Томиты). Это получается благодаря тому, что они принимают на вход состояние и возвращают множество всех возможных состояний.

Однако возникает проблема - сложность алгоритма, основанного на функциях *parse* и *climb* экспоненциальная. Для борьбы с этим в статье [7] предлагается дополнить функции механизмом запоминания результатов предыдущих вызовов, чтобы при очередном вызове , в случае, если эта функция уже вызывалась с такими параметрами, то результат возвращался без повторного вычисления. Доказывается [7], что в при такой реализации оценки по времени будут порядка $O(n^3)$.

Запоминание результатов предыдущих вычислений и будет аналогом объединения состояний и получения структурированного в виде графа стека (graph-structured stack) в алгоритме Томиты.

Технически этого можно добиться реализовав функцию *memoize*. На языке F# её можно реализовать так:

```

let memoize (f: 'a ->'b) =
    let t = new System.Collections.Generic.Dictionary<'a,'b>()
    in
    fun a ->
        if t.ContainsKey(a)
        then t.[a]
        else
            let res = f a
            in
            t.Add(a,res);
            res

```

В дальнейшем в качестве функции f будет выступать функция `parse` и `climb`.

Дополнительного ускорения можно добиться заранее вычислив функцию `goto`. Действительно, построение замыкания (closure), необходимое для вычисления функции `goto` дорогая операция, а вызов `goto` происходит при каждом вызове функции `climb`. Поэтому можно вычислить `goto` на этапе построения анализатора, и в целевой программе (в функции `climb`) свести вызов (`goto q X`) к поиску по ключу.

3.3 Расширенные контекстно свободные грамматики.

На практике оказывается удобно пользоваться расширенной контекстно свободной или EBNF грамматикой. То-есть такой грамматикой, у которой в правых частях правил могут использоваться конструкции регулярных выражений. Необходимость использования EBNF в инструментах генерации анализаторов обоснована в работе [3].

В качестве примера рассмотрим грамматику арифметических выражений без приоритетов :

- 1) $E \rightarrow E' + E'$
- 2) $E \rightarrow E' * E'$
- 3) $E \rightarrow '(E)'$
- 4) $E \rightarrow 'a'$

В данном случае удобно объединить правила 1) и 2) в одно:

$$E \rightarrow E('+' \mid '*')E'$$

Далее:

$$E \rightarrow (E('+' \mid '*')E) \mid ('(E)') \mid 'a' \quad (1).$$

3.3.1 Построение ДКА

Для того, чтобы построить табличный LR анализатор необходимо преобразовать расширенную контекстно свободную грамматику. Нужно избавиться от конструкции BNF в правых частях. Однако можно попробовать обойтись без преобразований грамматики.

Существуют способы поддержки расширенной контекстно свободной грамматики на уровне анализатора, без преобразований входной грамматики [7]. Для этого нужно переопределить функцию `goto`. Можно заменить позицию точки в правиле на номер состояния конечного автомата, построенного по соответствующему регулярному выражению. Заметим, что определённая выше функция `goto` может быть описана в терминах конечного автомата. В этом случае регулярное выражение и ,следовательно ,построенный по нему автомат очень прост.

Для построения конечного автомата по регулярному выражению воспользуемся алгоритмом Томпсона[5]. Результатом работы данного алгоритма является недетерминированный конечный автомат (НКА). Так как правая часть правила (1) - регулярное выражение, то применим к ней алгоритм Томпсона. Получим следующий НКА: **здесь должен быть граф**

Каждое состояние и каждый переход КА должны преобразоваться в состояние и переход LR автомата соответственно. Заметим, что в полученном НКА много ε -переходов. Например, каждая альтернатива вносит два дополнительных состояния и 4 ε -перехода. Чтобы уменьшить количество переходов в результирующем LR-автомате можно преобразовать НКА в детерминированный конечный автомат (ДКА). Для этого применим стандартный алгоритм преобразования НКА в ДКА[5]. Полученный в результате преобразований ДКА будет выглядеть следующим образом: **здесь должен быть граф**

На практике удобно представить результирующий ДКА в виде тройки (S, F, R) , где S – начальное состояние ДКА, F – множество конечных состояний и R – множество функций перехода (правил), задающее ДКА.

3.3.2 Построение множества ситуаций

Перейдём к построению LR ситуаций. Как было сказано выше, основная идея состоит в том, что точку из классического определения ситуации нужно заменить состоянием ДКА, построенного по правой части правила, как описано выше. Поясним. Предположим, что правило грамматики не содержит конструкций BNF. Например:

$$E \rightarrow E' + E$$

Множество ситуаций для данного правила:

- 1) $E \rightarrow .E' + E$
- 2) $E \rightarrow E'. + E$
- 3) $E \rightarrow E' + .E$
- 4) $E \rightarrow E' + E.$

Теперь построим ДКА в соответствии с нашим алгоритмом. Результат показан на рисунке ниже.



Рис. 1:

У построенного ДКА четыре состояния: S_1 , S_2 , S_3 , S_4 . Сопоставим состоянию S_1 ситуацию $E \rightarrow .E' + E$, S_2 ситуацию $E \rightarrow E'. + E$ и так далее. В итоге получим соответствие:

- $$\begin{aligned} S_1 &- E \rightarrow .E' + E \\ S_2 &- E \rightarrow E'. + E \\ S_3 &- E \rightarrow E' + .E \\ S_4 &- E \rightarrow E' + E. \end{aligned}$$

Таким образом, каждой LR-ситуации мы смогли сопоставить состояние ДКА, построенного по правой части правила. Заметим, что ситуация характеризуется правилом, для которого она строится и положением точки в правой части этого правила. Теперь вернёмся к предыдущему примеру и обобщим ситуацию.

В правой части правила – регулярное выражение. Выше мы построили по нему ДКА. Теперь построим для каждого состояния ДКА LR-ситуацию.

Для этого сперва опишем структуру данных, в которой будем сохранять одну ситуацию. Для дальнейшей работы нужна следующая информация и ситуации:

- Номер правила.
- Левая часть правила (нетерминал).
- Символ принимаемый ДКА в данном состоянии.

- Номер состояния ДКА, в которое он перейдет приняв данный имвол.
- Номер начального состояния ДКА.
- Номера конечных состояний ДКА.

Заметим, что эта информация состоит из двух частей: информации о ДКА в целом и правило перехода. Это ляжет в основу функции построения множества ситуаций по ДКА.

```
create_items (s,f,{rule | rule = (cur_num,symb,next_num)}) =
  {(prod_number,prod_name,item_number,symbol,next_num,start_state,finale_state) |
   start_state = s , finale_state = f , }
```

3.3.3 Вычисление GOTO

Вычисление функции GOTO - очень дорогая операция. При анализе вызов происходит при каждом вызове функции parse. По этому вычисление GOTO каждый раз сильно ухудшит производительность. Самое простое решение этой проблемы - вычислить значения функции заранее, на стадии генерации данных.

На данном этапе проще всего вычислить GOTO для всех возможных пар (состояние, символ) и сконструировать коллекцию, в которой, вместо вычислений, будет производиться поиск при анализе. В качестве ключей можно взять, например, значение hash-функции от значения параметров, для которых вычислент соответствующее значение.

Основа алгоритма - стандартный алгоритм вычисления GOTO при LR анализе, подробное описание которого можно найти в ***ссылка***. Небольшое изменение лишь в том, что ситуация имеет специальный вид.

3.4 Построение деревьев вывода.

Для практических целей одна только информация о том, принадлежит входная цепочка данному языку или нет, не очень полезна. Гораздо более ценный результат - дерево вывода цепочки в данной грамматике. Так как обсуждается недетерминированный алгоритм, то в нашем случае речь будет идти о множестве деревьев (лесе) вывода. Для краткости будем называть его лесом вывода.

Рекурсивно-восходящий алгоритм анализа, описанный выше, будет основой следующего шага. Теперь необходимо получить деревья вывода.

Строить дерево вывода можно в два этапа. Сперва, в результате, анализа цепочки, получить порядок применения правил, а затем построить по нему дерево вывода. Мы же будем строить дерево сразу во время анализа. Для этого потребуется изменить функции `parse` и `climb`, описанные ранее.

Вначале рассмотрим детерминированный случай, а затем перейдем к недетерминированному.

При детерминированном разборе, если цепочка порождается входной грамматикой, должно быть получено единственное дерево вывода.

$$parseq(u : A \rightarrow a.b, u = vx, b \Rightarrow v) \rightarrow (A \rightarrow a.b)(x)(b)$$

$$climb(q)(X)(u : A \rightarrow a.Xb, u = vx, b \Rightarrow v)(X) \rightarrow (A \rightarrow a.Xb)(x)(Xb)$$

На данном этапе структуру дерева можно сделать минимально простой. В вершине можно хранить только имя соответствующего нетерминала или терминала и информацию о сыновьях. Для представления дерева опишем соответствующий алгебраический тип: `tree = Node of string*tree*tree|Leaf of string`.

Теперь надо изменить эти функции так, чтобы они строили дерево вывода. Для этого добавим дополнительный параметр....

$$parseq(u : A \rightarrow a.b, u = vx, b \Rightarrow v) \rightarrow (A \rightarrow a.b) (x) \text{ (узлы СД для b)}$$

$climb(q)(X)(u : A \rightarrow aX.b, u = vx, b \Rightarrow v)(X) \rightarrow (A \rightarrow a.Xb) (x)$ (узлы СД для Xb)

$A \rightarrow R$ – правило грамматики: A – нетерминал, R – ДКА, построенный по регулярному выражению $(A \rightarrow R, i)$ – LR-ситуация, i – состояние ДКА $is_final(R, i)$ – проверка, что i – конечное состояние R

Дополним эту грамматику и пронумеруем правила. Введём стартовый нетерминал S и соответственно правило $S \rightarrow E$ и получим следующую грамматику:

- 1) $S \rightarrow E$
- 2) $E \rightarrow E + E$
- 3) $E \rightarrow E * E$
- 4) $E \rightarrow (E)$
- 5) $E \rightarrow a$

Данная грамматика порождает арифметические выражения с двумя бинарными операциями и скобками.

Очевидно, что такая грамматика неоднозначна. Существуют цепочки, которые выводимы несколькими способами. Например: возьмём цепочку $a+a+a$. Она имеет два левосторонних вывода:

$S \xrightarrow{1} E \xrightarrow{2} E + E \xrightarrow{5} a + E \xrightarrow{2} a + E + E \xrightarrow{5} a + a + E \xrightarrow{5} a + a + a$

и

$S \xrightarrow{1} E \xrightarrow{2} E + E \xrightarrow{2} E + E + E \xrightarrow{5} a + E + E \xrightarrow{5} a + a + E \xrightarrow{5} a + a + a$

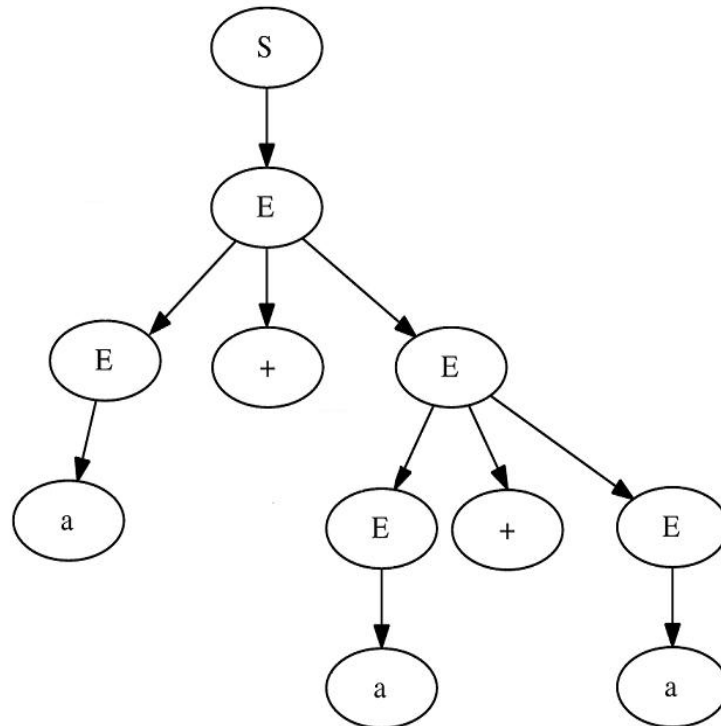
Над стрелкой – номер применяемого на данном шаге вывода правила.

Видим, что существует два вывода цепочки в данной грамматике:

$(1) \rightarrow (2) \rightarrow (5) \rightarrow (2) \rightarrow (5) \rightarrow (5)$

$(1) \rightarrow (2) \rightarrow (2) \rightarrow (5) \rightarrow (5) \rightarrow (5)$

Им соответствуют деревья вывода:



Соответ

Так как инструмент предназначен для работы с неоднозначными грамматиками, то в общем случае он должен строить лес вывода. То есть в случае если грамматика неоднозначна и существует несколько способов вывода некоторой входной цепочки, то должны быть построены все возможные деревья вывода.

4 Заключение.

На данный момент работа продолжается, но уже есть следующие результаты:

- бла-бла
- бла-бла

Задачи, требующие решения:

Список литературы

- [1] *Mark G.J. van den Brand, Alex Sellink, Chris Verhoef* Current Parsing Techniques in Software Renovation Considered Harmful.// IWPC '98: Proceedings of the 6th International Workshop on Program Comprehension. - IEEE Computer Society, Washington, 1998.
- [2] <http://www.research.microsoft.com/fsharp> (дистрибутивы и документация по языку F#)
- [3] *Чемоданов И.С.* Генератор синтаксических анализаторов для решения задач автоматизированного реинжиниринга программ.
- [4] *Dick Grune, Criel Jacobs* PARSING TECHNIQUES A Practical Guide
- [5] *Ахо А., Сети Р., Ульман Дж.* Компиляторы: принципы, технологии, инструменты. М.: Издательский дом <Вильямс> 2003. 768 с.
- [6] *Ronald Veldena* Jade, a recursive ascent LALR(1) parser generator. September 8, 1998
- [7] *Rene Leermakers* Non-deterministic Recursive Ascent Parsing. Philips Research Laboratories, P.O. Box 80.000, 5600 JA Eindhoven, The Netherlands.
- [8] *Larry Morell, David Middleton* RECURSIVE-ASCENT PARSING. Arkansas Tech University Russellville, Arkansas.
- [9] *Lex Augusteijn* Recursive Ascent Parsing (Re: Parsing techniques). lex@prl.philips.nl (Lex Augusteijn) Mon, 10 May 1993 07:03:39 GMT
- [10] <http://www.cwi.nl/projects/MetaEnv> (сайт разработчиков ASF+SDF)
- [11] <http://www.gnu.org/software/bison> (сайт разработчиков Bison)
- [12] <http://www.cs.berkeley.edu/smcpeak/elkhound> (сайт разработчиков Elkhound)