

РЕГУЛЯРИЗАЦИЯ КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК НА ОСНОВЕ ЭКВИВАЛЕНТНЫХ ПРЕОБРАЗОВАНИЙ СИНТАКСИЧЕСКИХ ГРАФ-СХЕМ

Федорченко Людмила Николаевна

Специальность 05.13.11 – Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Научный руководитель: доктор физико-математических наук
профессор

Баранов Сергей Николаевич

Актуальность темы исследования

- сложность разработки трансляторов растёт из-за увеличения ассортимента вычислительных устройств и требований к функционированию ПО
- сокращение сроков разработки как следствие частого выпуска обновлённых моделей, адаптируемых к модернизируемым вычислительным системам и конкурентности в рыночной экономике
- огромное разнообразие языков (более 4 тыс. новых), применяемых не только в общесистемном ПО, но и во многих предметных областях
- неоднородность спецификаций реализуемых языков (БНФ, двухуровневые грамматики)

использование формальных методов на всех этапах разработки позволяет автоматизировать процесс создания и сопровождения транслирующих систем

- применение автоматизации в предварительной обработке синтаксиса реализуемых языков обеспечивает существенное сокращение трудозатрат и времени разработки фазы синтаксического анализа, которая занимает до 25% ресурсов при создании транслятора

Актуальность темы исследования(2)

Обусловлена ростом сложности и объёма разработки транслирующих систем при сокращении сроков их разработки, что требует создания и применения новых подходов к автоматизации этапа предварительной обработки (препроцессинг) и преобразований грамматик с целью приведения их в соответствие с применяемым методом распознавания и анализа реализуемого языка.

При разработке транслятора необходимо решать задачи **упрощения** грамматики и задачу разрешения **конфликтных ситуаций** в процессе построения анализатора языка, связанные с языковой **неоднозначностью и недетерминированностью** магазинного автомата

Фундаментальные работы таких авторов как:

Н.Хомский – иерархия грамматик

С.К.Клини – регулярные множества и регулярные выражения

Г.С. Цейтин, А.П.Ершов и С.С.Лавров – технология программирования

Д.Кнут, Ахо А.В., Ульман Д.– теория языков и автоматов

составляют основу современной теории построения анализаторов языка, которую необходимо расширять в соответствии с требованиями практики во всех областях применения – наличием инструментария и методов для быстрой настройки синтаксиса реализуемого языка в условиях жестких временных ограничений.

Цель работы и задачи исследования

Цель – разработка методики и алгоритмов эквивалентных преобразований трансляционных грамматик при обработке языков в системах построения трансляторов.

Задачи :

- анализ существующих подходов и методов синтаксического разбора при построении анализатора языка.
- построение эффективной модели определения и распознавания языка. Определение свойств и ограничений на входную грамматику для автоматического синтеза распознавателя языка.
- разработка алгоритма регуляризации грамматики языка на основе эквивалентных преобразований синтаксической граф-схемы.
- построение экспериментальной инструментальной программной системы, реализующей разработанные алгоритмы

Положения, выносимые на защиту

- Алгоритм регуляризации трансляционных контекстно-свободных грамматик с помощью их эквивалентных преобразований. Показана его корректность и исследована его сложность
- Способ задания модели реализуемого языка с использованием формального понятия синтаксической граф-схемы грамматики.
- Математическая модель и реализующие её алгоритмы построения распознавателя языка для подкласса КСР-грамматик, занимающего промежуточное место между регулярными и КС-грамматиками в иерархии классов (по Хомскому)
- Экспериментальная реализация инструментального средства SynGT (Syntax Graph Transformations), позволяющая проверить теоретические результаты работы. Основные алгоритмы эквивалентных преобразований протестированы на наборе тестов и примеров, часть которых представлена в приложении

Новые результаты

Научной новизной обладают следующие результаты выполненных исследований:

- Впервые сформулирована теорема об адекватности распознающего автомата, построенного на основе синтаксической граф-схемы и подкласса *KCP*-грамматик
- Разработан алгоритм регуляризации трансляционной грамматики на основе синтаксической граф-схемы с целью получения абсолютно минимального по числу состояний синтаксического анализатора

Технологии разработки трансляторов

- CDL-технология С.Н.А. Костера (аффиксные грамматики)
- Технологии LL-, LR-разборы
- Lex/Yacc, Flex/Bison- семейство компиляторов GNU
- Antlr -(семейство для различных платформ)
- Eli- коллективная разработка нескольких университетов (США)
- Форт-технология - С.Н.Баранов, Г. Ноздрунов
- SYNTAX- ЛГУ (Мартыненко Б.К. Лукичёв А.)

Методика эквивалентных преобразований

- ✚ Формальной моделью методики, описанной в диссертации (глава 2) и реализованной в экспериментальном программном средстве SynGT являются регулярные выражения – как средство описания языка, и конечный автомат – в качестве адекватного средства его обработки.
- ✚ Теоретическим обоснованием этой модели являются теорема С.Клини и её следствия о том, что класс регулярных множеств является *минимальным* классом, содержащим все конечные множества, *замкнутым* относительно операций объединения, конкатенации и замыкания;
- ✚ Регулярные множества распознаются конечными автоматами и представляются регулярными выражениями. Использование этой модели в обработке данных возможно лишь в применении к регулярным языкам.

Обобщение регулярной модели

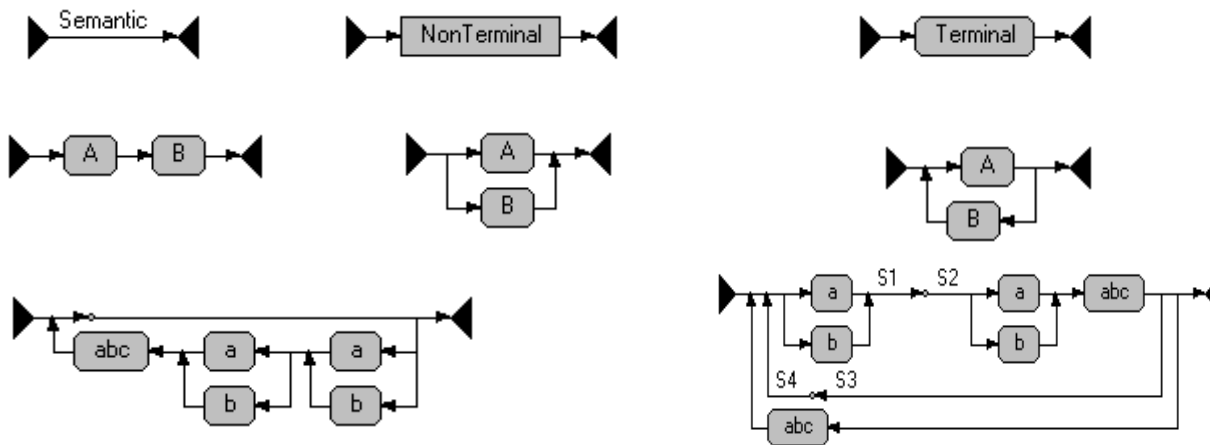
- Вводится *итерация с разделителем* или *обобщённая итерация #* (или итерация Цейтина) не расширяет множество регулярных слов и может быть определена через традиционную (одноместную) операцию Клини (*) как $P\#Q = P(Q,P)^*$, но частично решает задачу минимизации регулярного выражения по числу вхождений символов из объединённого алфавита СИМВОЛОВ
- *КСР-грамматика*- обобщение КС-грамматики. Класс языков не расширяет, но снимает излишнюю структурированность языка
- *Синтаксическая граф-схема*- графический аналог КСР-грамматики, стартовый объект для синтеза анализатора языка, в терминах вершин которой удобно формулировать свойство детерминированности языка.

Синтаксическая граф-схема (СГС)

Регулярное выражение	→	Конечный автомат
КС-грамматика	→	Диаграммы Вирта
КСР-грамматика	→	Граф-схема

Синтаксическая граф-схема (СГС)

- Создается рекурсивно из графов для элементарных регулярных выражений и бинарных операций



"#(('a';'b'),('a';'b'),'abc'). (('a';'b'),S1,"",S2,('a';'b'),'abc')#(S3,"",S4)#'abc'.

Модель распознавателя КС-языков

- ✚ **В работе** изложен метод построения синтаксических распознавателей по контекстно-свободным грамматикам в регулярной форме (КСР-грамматикам),
- ✚ Введены следующие понятия:
 - ✚ синтаксическая граф - схема (СГС),
 - ✚ отношение достижимости в граф - схеме
 - ✚ маршруты в СГС
 - ✚ язык, порождаемый СГС
 - ✚ состояние распознавателя в СГС
- ✚ Рассмотрены свойства синтаксической граф-схемы для построения детерминированного распознавателя
- ✚ Доказана теорема об адекватности языка, порождаемого СГС, распознавателю, построенному по СГС

Инструментальная система SynGT

- **Основные модули:**

- *текстовый редактор*
 - проверяет корректность регулярных выражений
 - формирует well-formed грамматику (удаление лишних правил, редактирование словарей символов и др.)
- *графический редактор*
 - осуществляет пользовательский интерфейс для обработки грамматики
- *эквивалентные синтаксические преобразования*
 - Формирует грамматику для синтеза детерминированного процессора
- *генератор тестов*

SynGT- базовые функции

Алгоритм регуляризации предполагает выполнение следующих базовых эквивалентных преобразований над регулярными выражениями и КСР-правилами:

- подстановка вместо нетерминала его порождения
- удаление вхождения леворекурсивного или праворекурсивного нетерминала в правой части правила
- объединение общих префиксов в графе для нетерминала
- удаление повторяющихся альтернатив для нетерминала
- удаление крайних рекурсий для самовложенных нетерминалов
- сведение регулярного подвыражения в новый нетерминал в графе для и создание нового правила в грамматике
- удаление лишних правил

Набор базовых функций, реализованных в SynGT, служит основной цели – автоматизированное преобразование КС-грамматики с целью ее максимальной регуляризации, то есть превращение синтаксической граф-схемы грамматики в минимальный набор конечных автоматов для построения анализатора языка. Если КС-грамматика порождает регулярный язык, то синтаксическая граф-схема вырождается в один граф, представляющий эквивалентное регулярное выражение

Основные задачи, решаемые в системе SynGT

1. Предварительная обработка текстового представления КСР-грамматики

В результате должны иметь:

- правильность записи выражений: скобочной структуры, контроль всех операций в регулярных выражениях, наличие конечного символа;
- все входящие нетерминалы являются продуктивными, т.е. порождают непустые цепочки.

- ## 2. Преобразование текстового представления регулярного выражения в графическое (задача размещения на экране)
- ## 3. Собственно редактор граф-схемы (стандартные редак. функции)
- ## 4. Преобразование граф-схемы в текстовое регулярное выражение
- ## 5. Эквивалентные преобразования КСР-грамматики и синтаксической граф-схемы.

Метод регуляризации грамматик

- **Определение**

Регуляризация КС-грамматики - это процесс эквивалентного преобразования произвольной КС-грамматики G в КС-грамматику G_1 в регулярной форме (КСР-грамматику), причем такой, что из нее исключаются *самовставленные* нетерминальные символы.

- Если все нетерминалы КС-грамматики G являются несамовставленными, то их можно исключить с помощью эквивалентных преобразований. В этом случае грамматика G преобразуется в грамматику G_1 с одним правилом, правая часть которого является регулярным выражением над терминалами. Значением этого регулярного выражения является регулярный язык $L = L(G_1) = L(G)$.

Метод регуляризации грамматик(2)

Схема регуляризации:

- Выделение контекстно-свободной составляющей языка, в том случае, если синтаксис языка задан в виде двухуровневой грамматики (аффиксной или Ван Вейнгаардена) или какой-либо другой, отличной от КС-грамматики. Такая КС-составляющая будет входной грамматикой для системы эквивалентных преобразований SynGT.
- Эквивалентное преобразование входной грамматики в приведённую КСР-грамматику, когда удаляются е-правила, непродуктивные нетерминальные символы и циклические порождения.
- Исключение лево- (право) рекурсивных нетерминалов из правил грамматики.
- Выполнение операции глобальной подстановки вместо нетерминальных вершин их порождений.
- Минимизация регулярных выражений в правых частях правил. 17

Практическое применение метода

Применение разработанной методики и инструментария поддержки
проведено в следующих проектах:

- Разработка компилятора для языка ANSI Си в совместном с INRIA (Франция) и институтом А.М. Ляпунова (МГУ) проектах «ОСС- открытый компилятор языка Си», проект № 14
- Разработка «Компилятор языка Си для цифровых процессоров» в проекте, совместном с INRIA и институтом А.М.Ляпунова (МГУ)
- Модуль проверки циклов в логических правилах в Российско-французском проекте «ТурСА-теория типов в компьютерной алгебре»
- Синтаксический анализатор для языка Гезал в Будапештском институте Вычислительной техники Венгерской Академии Наук
- При выполнении НИР по созданию транслятора языка Алгол-68, выполнявшейся в Ленинградском государственном университете

Заключение

- Разработанная методика и алгоритмы эквивалентных преобразований трансляционной грамматики могут быть эффективно использованы в проектах различной направленности, связанных с реализацией языка для различных предметных областей, в том числе на всех уровнях в системах обеспечения информационной безопасности, например, при обработке и анализе правил политик безопасности, определении поведенческих сценариев, проверке их на непротиворечивость, разработке тестов и тестовых процедур, поиске областей уязвимости кода приложений, решении задач обработки больших по объёму спецификаций требований программного проекта, документы для которых могут содержать сотни и тысячи пунктов требований, в задаче преобразования внутренних форматов данных, например, RDF-формата, описывающего IDEF0-диаграмму, к формату данных, описывающему UML-диаграмму, и многие другие.
- Алгоритм регуляризации грамматики может применяться для создания миниатюрных языковых процессоров во встраиваемое ПО
- Достигнутый уровень автоматизации при использовании системы эквивалентных преобразований совместно с выбранной технологией автоматизации проектирования синтаксических анализаторов составляет от 50% до 100% в зависимости от специфики проекта. Снижение трудозатрат – примерно в 3-4 раза.

Направления дальнейших исследований

- Расширить класс допустимых трансляционных грамматик, добавлением атрибутов и аффиксов. Реализовать это расширение в системе SynGT
- Улучшить инструментальную систему, переписав алгоритмы в языке C++
- Расширить область использования данной методики и инструментального средства.