**Министерство науки и высшего образования Российской Федерации**

**федеральное государственное автономное образовательное учреждение высшего образования**

**“НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО”**

**Факультет** Программной инженерии и компьютерной техники

**Образовательная программа** Системное и прикладное программное обеспечение

**Направление подготовки (специальность)** 09.04.04 Программная инженерия\_\_\_\_\_\_\_\_\_

**Отчет**

**о научно-исследовательской работе**

Тема задания: «Выработка методов к анализу мультиязыковых текстов программ»

Обучающийся: Орловский М.Ю P4116

(Фамилия И.О.) (номер группы)

Руководитель практики от университета: Маркина Т.А, доцент факультета ПИиКТ

Санкт-Петербург

2023 г.

# Эпоха

## Исследовать возможность обобщения предлагаемого метода на сочетание различных языков программирования.

В ходе подготовки статьи на конференцию «КМУ XII» была проведена небольшая ревизия метода анализа, для его большей формализации. Было выяснено, что структуру семантического узла можно обобщить, если немного изменить свойства его компонентов. В данный момент, структура узла представляет информацию, представленную в таблице 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Таблица 1 – Обновленная структура семантического узла | | |  |
| Имя поля | Тип поля | Семантика | Что обновлено |
| Узел | Идентификатор AST узла | Указывает позицию анализируемой синтаксической конструкции | Было решено, что будет разумнее рассматривать позицию как конкретный AST узел из исходного дерева кода |
| Атрибуты | Гетерогенный список | Содержит список атрибутов узла | В отличие от предыдущего варианта хранения атрибутов (строковый ключ – строковое значение), гетерогенный список является более гибким механизмом, который проще обрабатывать программно |
| Вид | Перечисление | Является способом типизации конкретного семантического узла, исходя из ЯП, который он описывает | Решено было удалить этот атрибут, т. к. необходимая информация уже содержится в поле Узел |
| Семантика | Перечисление | Используется для отражения семантики, представляемой узлом | Решено было удалить этот атрибут, по причинам описанным ниже |

Решение об удалении поля «Семантика» связано в первую очередь с распределением сложности анализа. Предполагается, что у узлов будет единая семантика связи (семантика ребра в сети) – это семантика «Зависимость». Такая семантика была выбрана по нескольким причинам:

1. Многие семантические зависимости узлов являются бесполезными на этапе анализа – по ним сложно делать какие-либо выводы о структуре сети в связи с неупорядоченностью узлов при их связывании;
2. Семантические зависимости разного рода усложняют анализатор, делая его более специфичным и снижая гибкость;
3. Предполагается, что все семантические отношения между узлами можно будет вывести из готовой сети, при практическом применении этой сети в конкретном инструменте.

Таким образом, предполагается, что данная структура узла будет более гибкой в отношении способов хранения информации о семантике конкретных синтаксических конструкций.

Что касается алгоритма связывания, то он не претерпел явных изменений. Более формальное описание алгоритма следует в другой главе, но на высоком уровне алгоритм производит связывание узлов исходя из их типа и набора атрибутов. Алгоритм при этом имеет сложность O(n2).

## Исследовать наиболее часто используемые парадигмы программирования (процедурное, ОО, декларативное) в контексте мультиязыкового анализа

Прежде чем рассматривать мультиязыковой анализ в контексте обрабатываемых в нем языков, было бы полезно обобщить и категорировать эти языки. В целях работы предлагается использовать общепринятое разделение языков по парадигмам [1]. Данное разделение представлено в таблице 2.

|  |  |  |
| --- | --- | --- |
| Таблица 2 – Разделение языков по парадигмам | | |
| Парадигма | Основные особенности | Представители |
| Императивная | Переменные как способ манипуляции памятью; Мутабельность как основной способ проведения вычислений; Описание вычислений как конкретного набора инструкций | C, Pascal, Java, Golang, Bash, JavaScript и многие другие |
| Декларативная | Описание результата и его свойств, как способ его получения; Описание вычислений как высокоуровневых объявлений; | ML, Lisp, Haskell, SQL, Make, HTML и также многие другие |

Несмотря на то, что такого разделения достаточно для категоризации языков, оно будет расширено и детализировано, так как в рамках данной работы предстоит рассматривать различные языки очень разной природы. Такое расширенное и уточненное разделение представлено в таблице 3.

|  |  |  |
| --- | --- | --- |
| Таблица 3 – Расширенное разделение языков по парадигмам | | |
| Парадигма | Основные особенности | Представители |
| Процедурная и структурная | Инструкции сгруппированы в процедуры/подпрограммы; переменные как способ взаимодействия с памятью | C, Pascal, Golang, Bash, и многие другие |
| Объектно-ориентированная | Данные и код инкапсулированы в единое представление – объект; Объекты общаются посредством передачи сообщений друг-другу | Java (отчасти), C# (отчасти), JavaScript (отчасти), C++ (отчасти), Smalltalk, Erlang, Elixir, Self и другие |
| Функциональная | Функции, как способ выстроить процесс вычисления; Иммутабельность как фундамент формальных гарантий в коде | ML, Rust (отчасти), Haskell, Erlang, Elixir, Lisp |
| Макро | Процесс вычислений проводится путем лексических подстановок строк для достижения конечной строки; встречаются варианты, оперирующие на AST | Препроцессор C, Make, макросистема Rust, различные DSL в Web фреймворках |

Исходя из данной классификации, можно разделить анализируемые языки по категориям, в каждой из которых можно будет сформировать свою онтологию, описывающую основные особенности языков данной категории и их основные компоненты. Стоит учесть, что данное разделение сделано в целях упрощения и обобщения анализа и может не соответствовать общепринятому.

## Рассмотреть различные подходы к представлению семантической информации

Основным способом представления информации (и самым универсальным) является семантическая сеть. В данной работе был избран именно такой подход, однако, семантические сети трудно структурировать и формализовывать. И хотя в дальнейшем будет использоваться именно этот подход к представлению семантической информации, полезно рассмотреть иные подходы, активно используемые в компиляторах, оптимизаторах и анализаторах. Такие подходы отражены в таблице 4.

|  |  |  |  |
| --- | --- | --- | --- |
| Таблица 4 – Представление информации об исходном коде | | |  |
| Название представления | Основные особенности | Плюсы (в рамках данной работы) | Минусы (в рамках данной работы) |
| AST | Представление иерархической структуры программы, посредством связывания её синтаксических конструкций в единое дерево | Легко анализируется, может быть использовано в анализах, где порядок анализа не важен; легко визуализируется; универсально для многих видов анализа | Сложно изменяется; не подходит для анализов, которые зависят от порядка; некоторые анализы неприменимы |
| CFG | Представление пути исполнения программы, моделируется связями (переходами) между узлами (операторами или выражениями) | Многие другие анализы легко реализуются через CFG; хорошо подходит для императивных языков | Плохо подходит для неимперативных языков, где путь исполнения программы задан неявно; сложнее реализовать чем AST |
| Семантическая сеть | Представление сущностей и их зависимостей в виде графа, узлы которого определяют сущности, а связи - отношения | Легко синтезируется; очень универсально; очень гибко и может быть использовано в различных анализах | Слабо структурировано; нет четкой схемы как связывать узлы между собой, по какому принципу выявлять сущности и отношения |
| Иерархия типов | Представление информации о связях сущностей в виде типов этих сущностей, формирующих определенную иерархию | Может быть простым вариантом представления большого объема информации в ОО и функциональных системах; имеет много формальных средств и методов для анализа | Сильно зависит от избранного языка; может не нести ценности, если язык имеет слабую типизацию; довольно сложна для анализа в общем случае |
| Онтология | Представление, основанное на формальных знаниях в конкретной предметной области | Однажды специфицированная онтология может быть использована в другом языке при наличии схожих семантик; универсальна и расширяема | Быстро становится сложной при расширении; обычно, предметные области (а в особенности ЯП) трудно формализуемы |

Стоит заметить, что существует еще большое множество иных представлений, которые слабо подходят для описания семантики языков программирования: фреймовые, логические и вероятностные модели, а также нейронные сети. Хоть они и представляют ценность для реализации различных анализов, в рамках данной работы они либо слишком сложны для применения, либо не обеспечивают необходимой точности.

Исходя из перечисленных в таблице 4 подходов к представлению семантической информации, можно заключить, что использование семантических сетей стоит совместить с иными способами представления информации. Предполагается, что использование онтологий упростит формирование семантических сетей, позволит их формализовать и повысит количество переиспользуемых техник анализа.

# Эпоха

## Провести анализ текущих решений в области мультиязыкового анализа на прикладном уровне (IDE и инструментальные средства)

На данный момент, в индустрии разработки ПО мало качественных и достаточных инструментов для мультиязыковой разработки, чего очень не хватает пользователям языков программирования [2]. Однако, существует ряд решений, направленных на обеспечение инструментальной поддержки разработки. Многие решения из этой области разработаны для специфичных предметных областей, в первую очередь для веб-разработки. В рамках данной работы были рассмотрены несколько инструментов, поддерживающие (либо имеющие возможность) мультиязыковой разработки для анализа межъязыковых связей.

### JetBrains Rider

JetBrains Rider это кроссплатформенная IDE имеющая поддержку различных языков, в первую очередь .Net семейства (C#, Visual Basic, F# и прочие), а также ряда DSL которые используются в соответствующих фреймворках (Blazor, XAML и др.) [3]. Она использует ReSharper как анализатор и инструмент для обеспечения продуктивности разработчика. Вкупе с IntelliJ IDEA, которая служит базисом для IDE, Rider обеспечивает большое количество удобной для разработчика функциональности. Краткий перечень такой функциональности включает:

1. Обнаружение зависимостей между символами на мультиязыковом уровне;
2. Переименование символов по их зависимостям;
3. Контекстная информация о синтаксическом элементе;
4. Обнаружение несоответствий при межъязыковом взаимодействии;
5. Поиск определения или объявления символа в межъязыковом контексте;
6. Поиск всех использований определенного символа в межъязыковом контексте;
7. Автодополнение, доступное в межъязыковом контексте.

Несмотря на ряд продуктивных возможностей, можно выявить недостатки Rider, касающиеся межъязыкового анализа и рефакторингов. Одним из основных недостатков является узкий спектр поддерживаемых DSL – Rider хорошо взаимодействует с исходным кодом Blazor, но, к примеру файлы конфигурации или файлы проектов/решений им воспринимаются плохо. Ни одна (за исключением пункта 7) из возможностей перечисленных выше не применима к файлам JSON, YAML или SLN.

### SonarQube

SonarQube это автоматический инструмент для проверки кода, предназначенный для увеличения качества кода, раннего устранения ошибок и повышения скорости внедрения. Основным инструментом, привлекающим внимание в рамках данного проекта, является SonarLint – универсальный статический анализатор, поддерживающий множество популярных языков программирования. SonarQube имеет возможность анализировать мультиязыковые проекты, хотя об анализе межъязыковых связей информации нет [4].

Возможности инструмента схожи с JetBrains ReSharper и также предоставляют результаты анализа в реальном времени, для возможности интеграции инструмента в IDE.

Одним из явных недостатков инструмента можно выявить высокую связность с инфраструктурой Sonar – его возможности ограничены, если не использовать полноценный SonarQube сервер и большое количество дополнительных инструментов, что может затруднить интеграцию SonarLint в небольшой проект.

### Mulang

Mulang позиционируется как универсальный, мультиязыковой и мультипарадигменный статический анализатор, ориентированный на обнаружение ошибок и формирование предикатов (ожиданий) о коде [5].

Языки, поддерживаемые Mulang включают:

1. C;
2. Haskell;
3. Java;
4. JavaScript (ES6);
5. Python (2 и 3);
6. Prolog.

Также, имеется поддержка Ruby и PHP через специфические языковые инструменты.

Так как проект является открытым, есть возможность добавить поддержку определенных языков при необходимости.

Mulang является инструментом, позволяющим работать с кодом, как с базой знаний и, соответственно, не поддерживает полноценной интеграции с IDE. Однако, исходя из набора «ожиданий», сформированного разработчиками и из возможности определять свои «ожидания» можно заключить, что проект концептуально применим в IDE при разработке соответствующего плагина.

Небольшой набор «ожиданий» и ошибок, которые способен обнаруживать Mulang включает:

1. «Делает ли элемент вызов определенной функции?»;
2. «Присутствует ли переменная в выражении?»;
3. «Представлено ли вычисление как рекурсия?»;
4. «Используется ли оператор «если»?»;
5. Дупликация кода;
6. Ошибка в имени идентификатора;
7. Недостижимый код;
8. Длинный список параметров.

## Провести анализ моделей представления семантической информации программ в контексте различных ЯП и технологических стеков

Несмотря на наличие инструментов для анализа мультиязыкового кода, большая часть из них является проприетарными решениями с закрытым исходным кодом. В связи с этим, большая часть инструментов работает как «черный ящик» и изучить модель представления семантической информации внутри таких инструментов не представляется возможным.

Однако, существуют анализаторы с открытым исходным кодом, представляющие интерес в данном вопросе.

### Multilingual static analysis tool

В контексте MLSA используются различные модели представления семантической информации, а именно:

1. Граф вызовов функций;
2. CFG;
3. Граф присваиваний;
4. Граф зависимостей на уровне файлов и директорий.

Исходя из того, что большая часть представлений является внутренними структурами данных исходного языка, возможность расширения такого анализатора как MLSA является трудоемкой задачей. И хотя для каждого формата есть своё сериализованное унифицированное представление, зависимость проекта от специфической техники парсинга (Island grammar parsing) снижает его расширяемость.

Также, инструмент обязательно включает в себя парсер исходного языка, что затрудняет его интеграцию с существующими анализаторами.

### Mulang

Mulang в свою очередь представляет особый интерес благодаря наличию унифицированного представления называемого авторами проекта «Abstract semantic tree».

Такое дерево состоит из узлов, попадающих в одну из 5 категорий:

1. Выражения;
2. Шаблоны;
3. Типы;
4. Равенства;
5. Генераторы.

Равенства можно описать как отображение из списка шаблонов (параметров функции или параметров match выражения) в тело (выражение либо набор частичных выражений). Они семантически представляют собой математическое отображение из входных параметров/шаблонов в одно или несколько выражений и, следовательно, подходят в первую очередь для функциональных языков. Генераторы семантически представляют собой выражения, имеющие ленивый (по надобности) тип вычислений.

На остальных трех категориях стоит остановится поподробнее.

#### Выражения

Выражение является основной описания любого типа вычисления. Mulang AST моделирует через выражения иные конструкции, описывающие вычисления – в первую очередь операторы и объявления. Такой подход является гибким, хотя и размывает границу между чистыми вычислениями и вычислениями с сайд-эффектами.

В Mulang представлен весь стандартный, несколько упрощенный набор семантических конструкций, описывающих выражения, встречающийся в многих популярных императивных и функциональных языках. Из-за высокого уровня абстракции можно выявить некоторые недоработки онтологии.

Она выражает общие языковые конструкции, но делает это неоднородно, к примеру цикл языка C описан как отдельная конструкция ForLoop, хотя семантически она может являться общим случаем итерации, так как через неё выражаема конструкция For (являющаяся генератором). Также, онтология не покрывает семантику некоторых языков, хотя причин для этого не имеет. К примеру, конструкция Class имеет лишь одно поле для обозначения базового класса, однако к C++ это в общем случае не подходит.

#### Шаблоны

Шаблоны в терминах Mulang AST обозначают параметры или, собственно, шаблоны, предназначенные для обозначения хода вычислений. Они не несут значений сами по себе и являются абстракцией над многими управляющими и структурирующими конструкциями языков. К примеру, они могут быть использованы для обозначения параметров функций, параметров в выражениях match и switch, а также для обозначения литеральных конструкций (числа, строки, списки, кортежи и т. д.).

При этом многие шаблонные конструкции удаляют изначальную семантическую информацию о моделируемых семантических концепциях и служат, скорее, как Ad-hoc решение, направленное на выражение минимальной информации о конструкции.

#### Типы

Типы в Mulang AST представляют собой довольно слабый, но тем не менее, достаточный набор семантик для описания операций над типами. Типы обозначают операции, которые могу проводиться над заданными типами как над множествами или функциями над множествами. К примеру, есть возможность задать параметризированный тип или задать ограничения на тип. Также есть возможность явного обозначения типа у определенного идентификатора.

Стоит заметить, что все конструкции оперируют над строками – таким образом, большая часть специфичной семантической информации остается необработанной и, следовательно, неиспользованной.

Если подытожить, то основные характеристики данной онтологии применимо к данной работе включают:

1. Онтология достаточно полная для многих языков (особенно высокоуровневых);
2. Некоторые конструкции выражений могут быть доработаны или расширены, в зависимости от задачи;
3. Конструкции шаблонов и равенств позволяют поверхностно описать необходимую семантику управляющих и структурных конструкций языка;
4. Базовое описание типов является малопригодным в общем случае, т. к. чаще проще обозначать операции над типами как стандартные выражения;
5. Онтология из-за специфики решаемой инструментом задачи не задает явным образом контекст времени связывания сущностей и идентификаторов, что очень важно для полного описания семантики языков.

## Исследование влияния прикладных областей языков на возможность обобщения метода мультиязыкового анализа

Согласно исследованию [6], проведенному для выяснения важных аспектов мультиязыковой разработки, было выяснено что большая часть разработчиков, использующих несколько языков в проекте, занимается веб-разработкой и клиент-серверными решениями. Также, было выяснено, что наиболее частым сочетанием языков была пара GPL/DSL, что объясняется большим количеством DSL существующих в индустрии веб-разработки.

Согласно этому исследованию, наиболее популярные возможности, предоставляемые инструментами разработчика, включали в себя (от частых к редким по инструментальной поддержке):

1. Подсветка синтаксиса;
2. Переименование схожих семантически идентификаторов;
3. Навигация по коду;
4. Анализ ошибок и нарушений консистентности.

Исходя из исследования [7], основные прикладные области, имеющие явную гетерогенную структуру в отношении языков программирования, включают:

1. Энтерпрайз разработку (J2EE, .NET Platform);
2. Веб-разработку (современные JavaScript фреймворки, Electron);
3. Встроенные системы.

Также, стоит учесть, что и в иных областях нередки ситуации использования различных языков (в первую очередь это касается файлов конфигурации или сборки).

Основные сложности, которые существуют при анализе мультиязыковых текстов программ, применимо к разным предметным областям, представлены в таблице 5.

|  |  |  |
| --- | --- | --- |
| Таблица 5 – Сложности мультиязыкового анализа | | |
| Предметная область | Проблема | Пояснение |
| Энтерпрайз, Веб-разработка | Большое количество различных языков, созданных для разных целей | Использование разных языков для фронтенд и бекенд разработки, а также для операций с БД является очень частым явлением и вкупе с другими языками для конфигурации и сборки может очень усложнить проект |
| Энтерпрайз | Широко распределенная система зависимостей | Зависимости между языками могут существовать не только на привычном уровне исходного кода, но также и на уровне файлов или даже целых отдельных API (например HTTP) |
| Встроенные системы | Большое разнообразие аппаратных средств | Из-за высокого разнообразия аппаратных средств, такие универсальные языки как C или ассемблер не имеют достаточной выразительной силы, в связи с чем возникает необходимость в дополнительных средствах (например, в кодогенерации) |
| Встроенные системы | Специфическая семантика в зависимости от аппаратных средств | Многие компиляторы для встроенных систем (C, C++, Python) не всегда следуют стандартам в угоду машинно-специфичным возможностям, из-за чего семантика исходного языка может быть нарушена, что затрудняет статический анализ |

Таким образом, обобщение мультиязыкового анализатора выглядит как достаточно сложная задача. Однако, благодаря использованию фиксированной формальной онтологии и соответствующих семантик представляется возможным реализация поверхностного анализа, отвечающего требованиям целевой платформы.

# Эпоха

Следует описать предлагаемый метод языкового анализа, разделив описание на три секции:

1. Входная информация для анализа;
2. Формальная структура анализатора, как реализации метода анализа;
3. Результаты анализа, его формат и возможные сценарии использования.

## Рассмотреть природу входной метаинформации и способы её обработки и хранения

Метод, описываемый в данной работе, является агрегатором различных специфичных анализаторов, каждый из который проводит свой собственный анализ для конкретного языка программирования.

Таковой анализ проводится над AST, полученных каким-либо образом для избранных программ, написанных на избранных языках программирования. Способ получения AST, специфический анализ и язык программирования, на котором реализован анализ не имеют для метода анализа, рассматриваемого в данной работе, особого значения.

Однако, имеются накладываемые ограничения на природу данных, используемых анализатором и формат данных результата анализа.

Для реализации конкретного анализа конкретного языка предполагается использование определенной онтологии – набора формально определенных семантических конструкций, которые будут использоваться для описания семантики избранного языка. Онтология является унифицированным языком, что позволяет совместить результаты работы разных специфических анализаторов для дальнейшего использования в мультиязыковом контексте. Таким образом, каждый специфический анализатор, помимо использования AST конкретного языка будет также использовать определенную онтологию для кодирования результатов анализа.

В отношении формата выходных данных такого анализатора, его результатом работы должен быть список узлов семантической сети, имеющий структуру, представленную в таблице 6.

// откуда анализ начинается, можно упомянуть Island Grammar

// че делает

// куда идут результаты

###### СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. <https://homes.cs.aau.dk/~normark/prog3-03/html/notes/paradigms_themes-paradigm-overview-section.html>
2. Pfeiffer, RH., Wąsowski, A. (2012). Cross-Language Support Mechanisms Significantly Aid Software Development. In: France, R.B., Kazmeier, J., Breu, R., Atkinson, C. (eds) Model Driven Engineering Languages and Systems. MODELS 2012. Lecture Notes in Computer Science, vol 7590. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33666-9\_12
3. https://www.jetbrains.com/rider/
4. https://www.sonarsource.com/products/sonarlint/
5. <https://mumuki.github.io/mulang/>
6. Mayer, P., Kirsch, M. & Le, M.A. On multi-language software development, cross-language links and accompanying tools: a survey of professional software developers. *J Softw Eng Res Dev* **5**, 1 (2017). <https://doi.org/10.1186/s40411-017-0035-z>
7. Z. Mushtaq and G. Rasool, "Multilingual source code analysis: State of the art and challenges," 2015 International Conference on Open Source Systems & Technologies (ICOSST), Lahore, Pakistan, 2015, pp. 170-175, doi: 10.1109/ICOSST.2015.7396422.