Реализация и применение строковых алгоритмов к задаче поиска повторов в документации программного обеспечения

Автор: Мишин Никита Матвеевич, группа 16.Б11-мм
Научный руководитель: к.ф-м.н., доцент С. В. Григорьев
Научный консультант: к.ф-м.н. Д. А. Березун

Рецензент: д. ф. н., Тискин А. В.

Санкт-Петербургский государственный университет
Кафедра системного программирования

13 июня 2020г.

Поиск повторов в документации ПО

Поиск повторяющихся фрагментов текста с целью:

- Избавление от нежелательной избыточности
- Выявление ошибок и несогласованности
- Переиспользование повторяющихся фрагментов
- ...

Задачи поиска повторов

- Поиск шаблона в заданном тексте:

  Луцив Д. В. *Поиск неточных повторов в документации программного обеспечения* (диссертация, 2018)

- Поиск групп повторов:
  - Неясно, насколько часто встречаются древовидные группы повторы
  - Часто ищут в *JavaDoc* документации

semi-local LCS и SA

- Позволяет решать более широкий класс задач:

  Tiskin A. *Semi-local string comparison: algorithmic techniques and applications (draft-book)*

- Алгоритмы с хорошими теоретическими показателями
- Не применялись к поиску к поиску повторов в документации
- Не были реализованы на практике

Постановка задачи

Цель:

- *Адаптация алгоритмов решения задач полулокальных поиска наибольшей общей подпоследовательности и выравнивая строк к задачам поиска повторов в документации ПО*

Задачи:

1. Исследовать существующие теоретические алгоритмы решения задач полулокального поиска наибольшей общей подпоследовательности и выравнивания строк и реализовать их на практике в виде *библиотеки алгоритмов*
2. Адаптировать алгоритмы решения полулокальных задач поиска *LCS* и *SA* к задаче поиска повторов в *JavaDoc* документации и реализовать соответствующее приложение на их основе
3. Провести экспериментальное исследование реализованных алгоритмов и анализ результатов

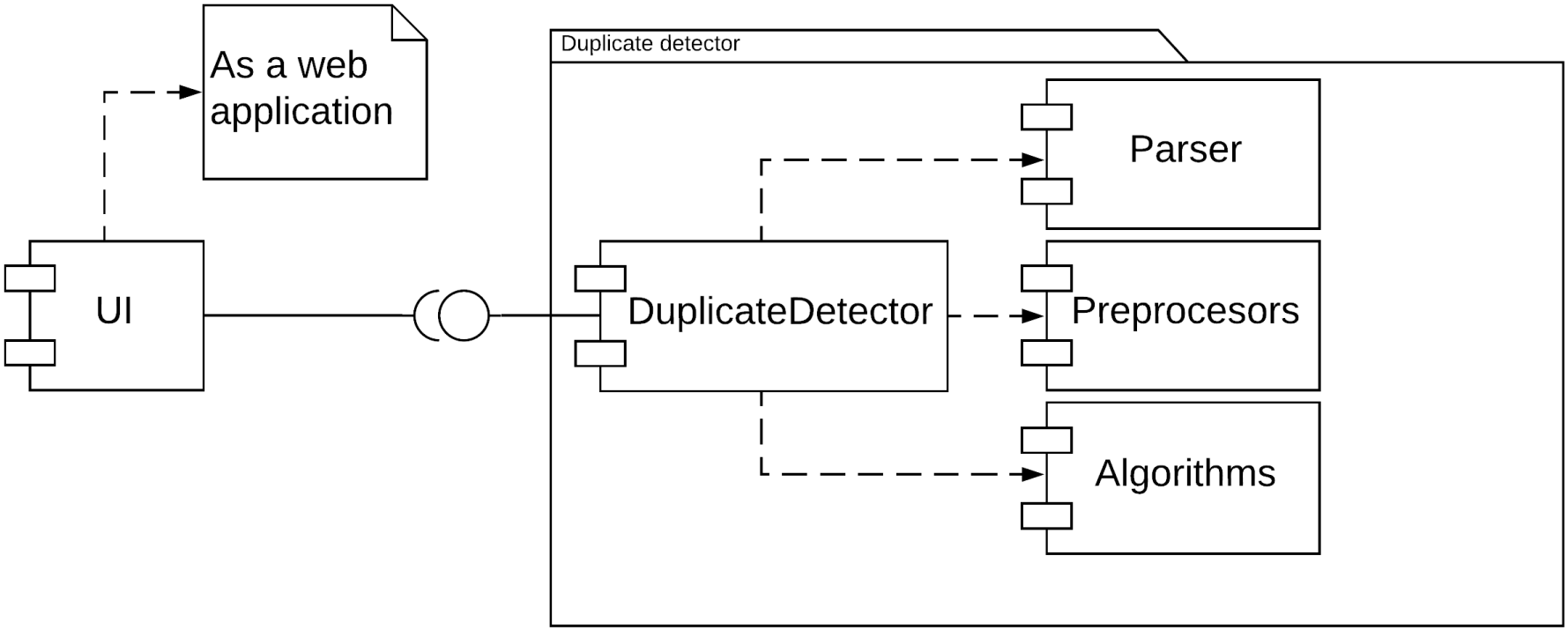
Библиотека алгоритмов *semi-local* задач: Алгоритмы

- Необходимые структуры данных
- Алгоритмы для решения semi-local *LCS* и *SA*:
  1. Распутывание кос
  2. Матричное умножение через тропическую алгебру
  3. Умножение кос
- Алгоритмы, решающие задачи на основе *semi-local*:
  1. Window-substring
  2. Поиск шаблона в тексте
  3. Локальное выравнивание с ограничениями:
     Tiskin A. *Bounded-Length Smith-Waterman Alignment* (WABI 2019)
  4. ...

Алгоритмы поиска дубликатов

- Поиск по шаблону
  1. Улучшенная асимптотическая версия алгоритма из диссертации
  2. Алгоритмы на основе анализа решения *semi-local*
- Поиск групп повторов в *JavaDoc*
  1. Графовые алгоритмы с матрицей расстояний на основе *semi-local* и её производных

Приложение для поиска повторов в *JavaDoc*



Визуализатор

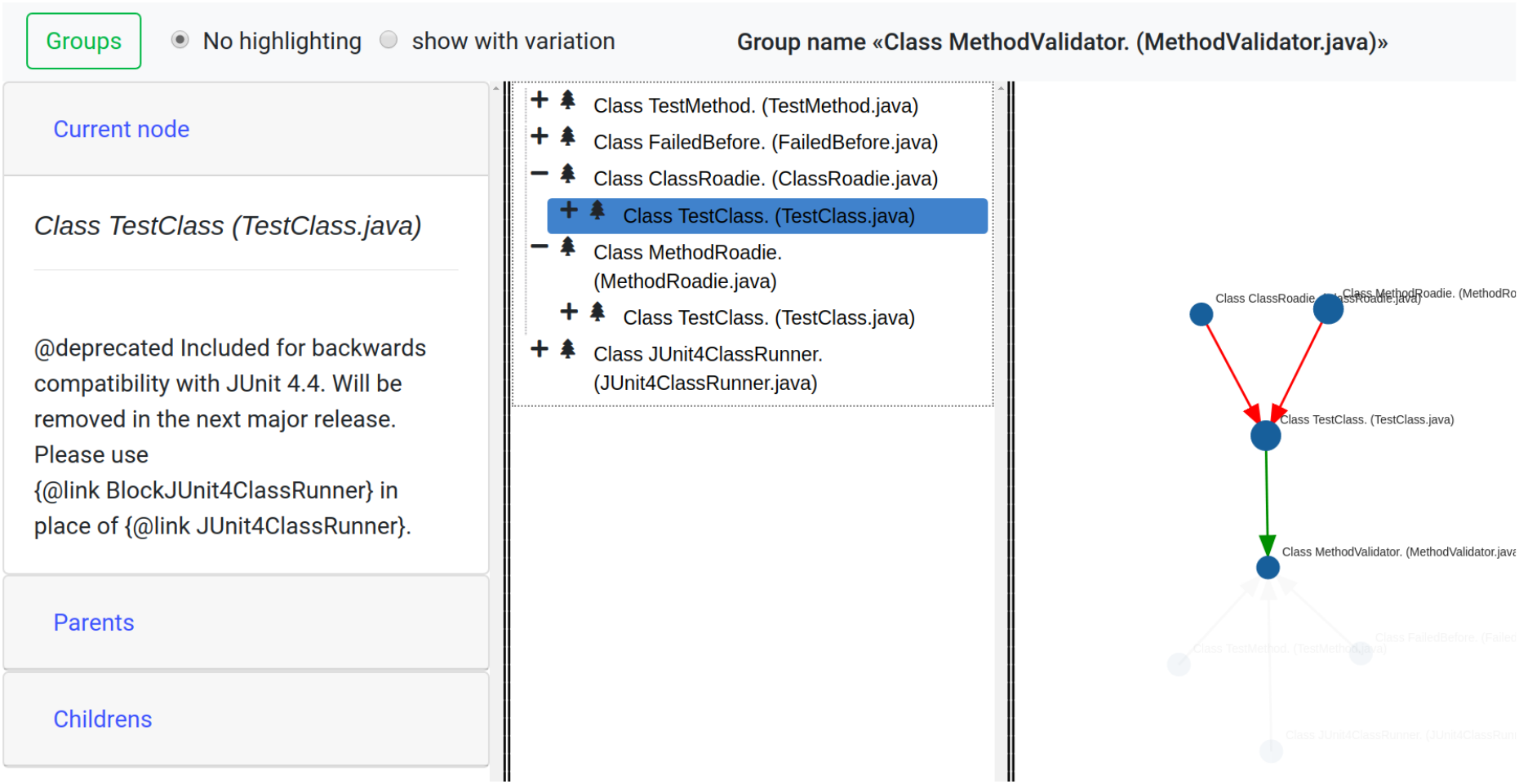
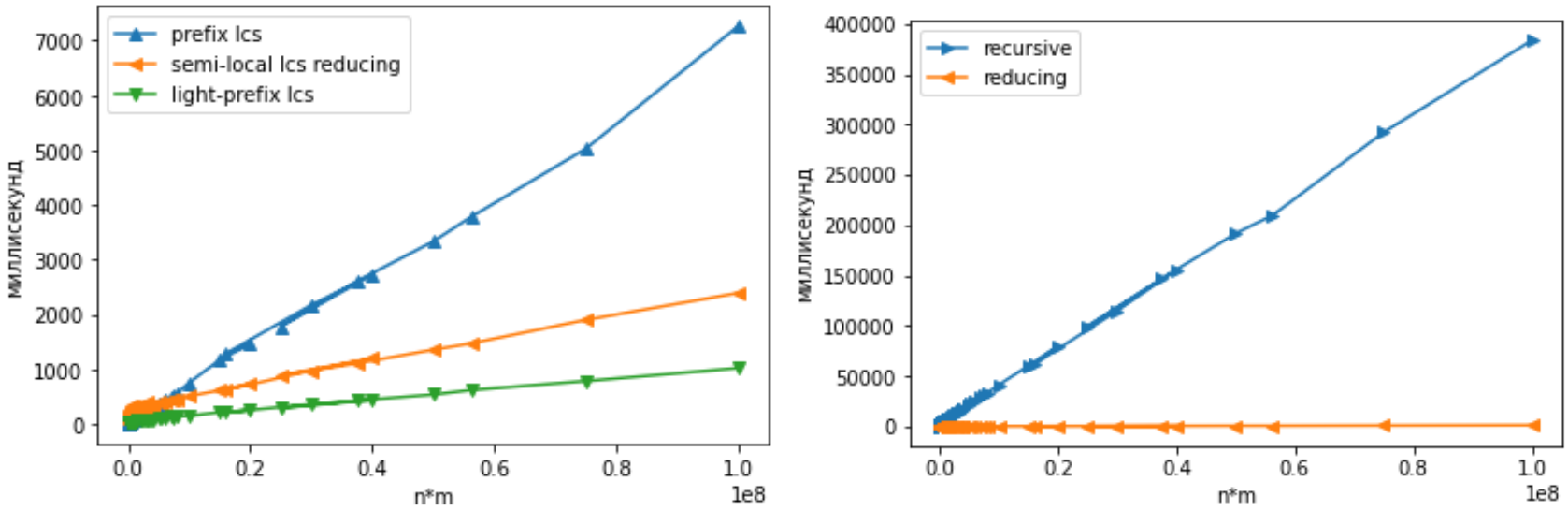


Рис.: Пример визуализации для группы повторов

Апробация: Применимость на практике алгоритмов решения задач *semi-local*

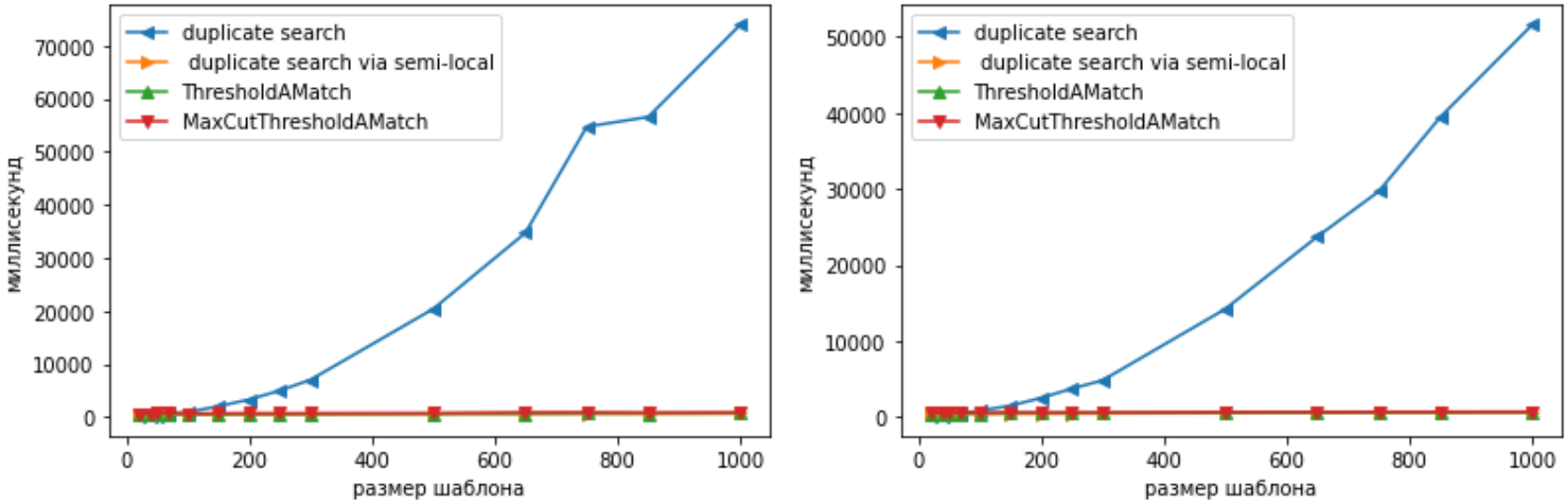


(a) semi-local lcs vs prefix lcs

(b) recursive vs reducing approach

Рис.: Сравнение скорости реализаций, решающих полулокальные задачи

Апробация: Применимость *semi-local* к задаче поиска по шаблону



(a) Сценарий с малым размером алфавита

(b) Сценарий с большим размером алфавита

Рис.: Сравнение скорости различных алгоритмов решения задачи *поиска по шаблону*

Апробация: Применимость *semi-local* к задаче поиска групп повторов

Название проекта	Кол-во комментариев	Кол-во повторов	Кол-во групп	Время исполнения (сек)
slf4j	188	157	25	8
apache commons io	1284	1180	92	569
apache commons collection	610	495	50	408
gson	498	356	81	96
junit	680	539	87	163
mockito	2979	2812	164	2012
guava	4340	3662	418	8505

Рис.: Результат работы приложения

Апробация: Применимость *semi-local* к задаче поиска групп повторов

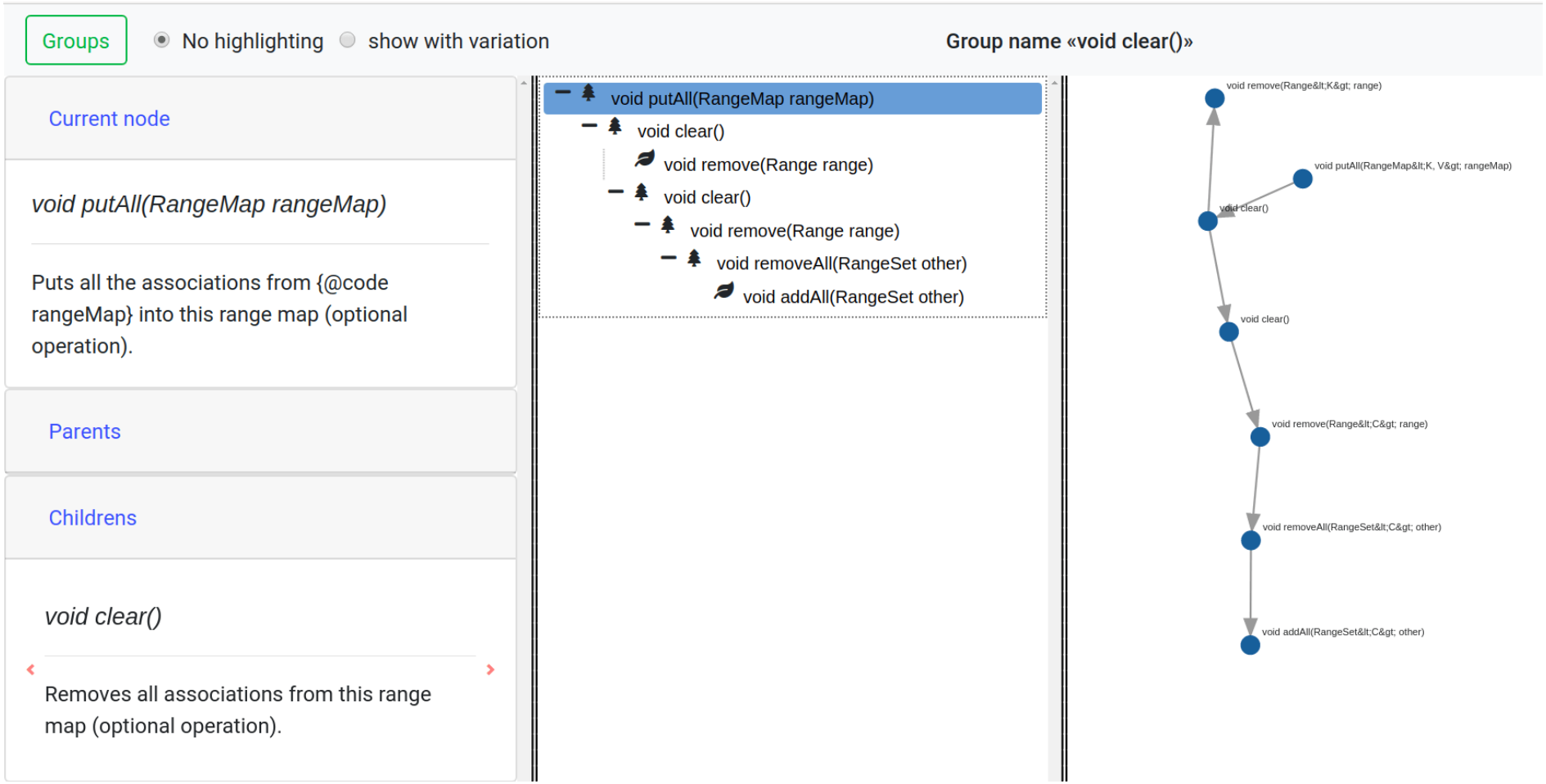


Рис.: Пример группы из проекта *guava*

Результаты

- Исследованы существующие теоретические алгоритмы решения задачи полулокального поиска наибольшей общей подпоследовательности и выравнивания строк и реализованы в виде *библиотеки алгоритмов* на языке *Kotlin*
- Создано приложение на языке *Kotlin* для поиска повторов в *JavaDoc* документации на основе адаптации алгоритмов решения полулокальных задач
- Проведено экспериментальное исследование реализованных алгоритмов и анализ результатов