УДК 004.421

Разработка diff модуль для 1C:Enterprise Development Tools

А.Р. Насибуллин[1], Г.С. Суаридзе[2]
[1]Московский физико-технический институт (государственный университет)
[2]Фирма «1С»

Аннотация
В работе рассмотрена задача семантического сравнения документов и предложено решение diff модуля для 1C:Enterprise Development Tools.

Программистам в своей работе довольно-таки часто приходится сравнивать файлы между собой, искать изменения и анализировать их. Для облегчения данного процесса необходим инструмент, который бы мог корректно сравнивать файлы и на выходе выдавать правильный, с точки зрения изменений между файлами, результат сравнения.

Введем общие определения, используемые в данной работе.

1. Текстовый файл – это набор строк текста, например, исходный код программы, написанный на каком-либо языке программирования.
2. Сравнение файлов – процесс поиска разницы между двумя файлами.
3. Diff модуль – механизм, реализующий сравнение файлов.

Существуют два класса алгоритмов сравнения файлов. К первому из них относятся текстовые алгоритмы сравнения, достаточно распространенный класс алгоритмов, ко второму – семантические алгоритмы сравнения. Рассмотрим каждый из подходов подробнее.

Текстовые алгоритмы сравнивают файлы, не учитывая их внутреннюю структуру, внутренние зависимости, минимальные единицы языка программирования (переменные, ключевые слова) и т.д. Алгоритмы данного класса оперируют файлами как текстовыми документами. К основным недостаткам текстовых алгоритмов можно отнести упущение семантики документов и неумение находить moved-блоки.

Семантические алгоритмы сравнивают файлы с учетом их внутренней структуры. Строится AST (Abstract syntax tree) либо его модификация, а далее семантические алгоритмы работают именно с AST. Сравниваются уже не файлы, а сравниваются их AST. Главным недостатком семантических алгоритмов является неумение находить изменения форматирования (замена табов на пробелы и наоборот и др.).

Целью работы являлось улучшение механизмов сравнения, а именно, учет семантики документов, нахождение moved-блоков и изменений форматирования.

Опишем математическую модель для класса семантических алгоритмов сравнения. На вход алгоритм сравнения получает два AST (Abstract syntax tree), построенных по файлам, которые будем сравнивать.

У каждой вершины есть тип и значение. Например, одна вершина имеет тип «строка» и значение «age», другая вершина имеет тип «число» и значение «18».

В процессе работы семантический алгоритм сравнения сопоставляет вершины, руководствуясь следующими правилами:

1. возможно сопоставление лишь тех вершин, которые имеют одинаковый тип;
2. сопоставление вершин имеет минимальное значение по метрике Tree edit distance. Метрика Tree edit distance описывает стоимость преобразования одного дерева в другое. Доступные операции – удаление/вставка/замена вершины, при этом каждая операция может иметь свою стоимость.

На выходе алгоритм сравнения выдает набор пар вершин из разных AST, которые сопоставлены друг с другом.

Рассмотрены несколько семантических алгоритмов и выбран алгоритм Gumtree, требующий квадратичных (по количеству вершин в AST) затрат времени и памяти.

Алгоритм Gumtree состоит из трех этапов:

1. жадный top-down алгоритм: находит изоморфные поддеревья;
2. bottom-up алгоритм: сопоставляет вершины, у которых много парных вершин-потомков;
3. bottom-up алгоритм: сопоставляет вершины в тех поддеревьях, у которых сопоставлены их корни.

Алгоритм Gumtree является лишь основой для нашего алгоритма, наш алгоритм состоит из следующих этапов:

1. построение AST для каждого из документов;

2. сопоставление вершин AST (алгоритм Gumtree);
3. анализ полученных сопоставлений вершин AST;
4. поиск изменений форматирования:
    a. сопоставление строк на основе результатов из п.2;
    b. парное сравнение строк;
5. поиск moved-блоков.

Стоит отметить, что реализованный алгоритм является универсальным, достаточно падать ему на вход AST сравниваемых документов и получить результат сравнения. Также наш алгоритм позволяет определять свою стратегию поиска moved-блоков.

Реализованный diff модуль на основе семантического алгоритма Gumtree написан на языке программирования Java и будет встроен в финальную релизную версию 1C:Enterprise Development Tools. 1C:Enterprise Development Tools – это среда для разработки бизнес-приложений на платформе “1С: Предприятие”.

Литература

1. *Falleri J.-R., Morandat F., Blanc X., Martinez M., Monperrus M.* Fine-grained and Accurate Source Code Differencing September 12th 2014
2. *Pawlik M., Augsten N.* RTED: A Robust Algorithm for the Tree Edit Distance August 27th -31st 2012