XIII МЕЖДУНАРОДНАЯ МОЛОДЕЖНАЯ НАУЧНО-ПРАКТИЧЕСКАЯ КОНФЕРЕНЦИЯ С ЭЛЕМЕНТАМИ НАУЧНОЙ ШКОЛЫ "ПРИКЛАДНАЯ МАТЕМАТИКА И ФУНДАМЕНТАЛЬНАЯ ИНФОРМАТИКА"

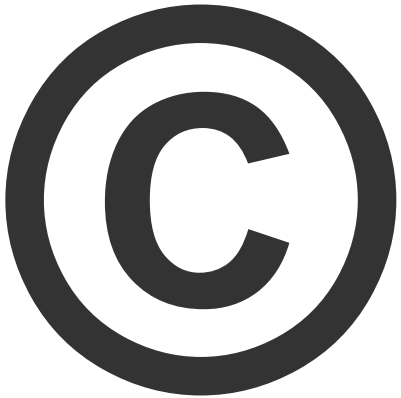
МЕТОДЫ СРАВНЕНИЯ РАССТОЯНИЙ МЕЖДУ АБСТРАКТНЫМИ ДЕРЕВЬЯМИ ДЛЯ ПРЕДОТВРАЩЕНИЯ ПЛАГИАТА КОДА

Подготовил:
Студент гр. МО-191 Шепелев Н.С.

План

- Проблема плагиата кода;
- Существующие алгоритмы сравнения текста их их недостатки;
- Понятие абстрактного синтаксического дерева;
- Алгоритмы на деревьях;
- Сравнение результатов.

Введение



Плагиат кода становится все более серьезной проблемой в сфере компьютерного образования и промышленности.

Соблюдение авторских прав стало важным аспектом разработки, и теперь актуальной задачей является создание методов и инструментов для автоматического выявления плагиата.

Цель исследования

Целью исследования является разработка алгоритмов, способных определять плагиат кода, а также их сравнение с существующими методами.

Расстояние Левенштейна

	“	T	E	S	T
“	0 →	1 →	2 →	3 →	4
S	1				
E	2				
T	3				

Алгоритм шинглов

1) Разбиение текста на последовательности (шинглы);
2) Вычисление хеша каждой последовательности;
3) Создание хеш-таблицы из полученных значений;
4) Сравнение двух хеш-таблиц;
5) Расчет оценки схожести с помощью различных метрик.

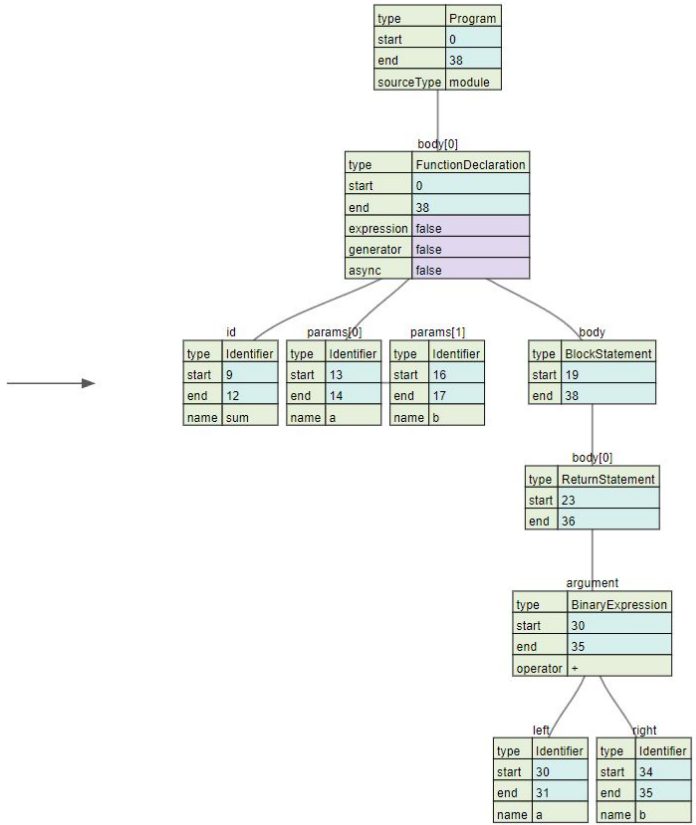
Проблема алгоритмов сравнения текста

```
function sumFunction(¶
firstArgument,¶
secondArgument¶
) {¶
return firstArgument +
secondArgument;¶
}¶
```

```
function sum(a, b) {¶
return a + b¶
}
```

Анализ AST

```
function sum(a, b) {
  return a + b;
}
```



Наложение деревьев

1) Построение AST для пары входящих программ;

2) Очистка дерева от шумовых узлов;

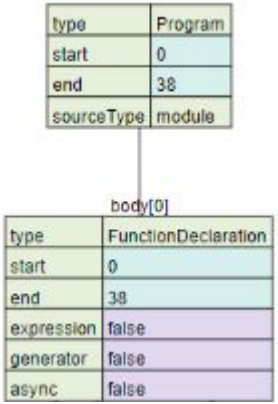
3) Построение множества узлов двух деревьев;

4) Поиск пересечения двух множеств;

5) Получение оценки схожести по метрике Жаккара

$$J = \frac{O}{A + B - O}$$

Редакционное расстояние



$$A((B)) \longrightarrow$$

$$A(B(D)(E))(C(D)(E)) \longrightarrow A((B)(D))((C)(E))$$

Сравнение результатов

Сравнение / Алгоритм	1	2	3	4
Алгоритм шинглов	116%	56%	80%	12%
Алгоритм наложения	93%	76%	37%	8%
Алгоритм расстояний	96%	78%	39%	13%

Заключение

1) Решена задача статического анализа;
2) Разработаны алгоритмы анализа кода;
3) Проведено сравнение с существующими алгоритмами сравнения текста;

Список литературы

1) Roy C., Cordy J. A Survey on Software Clone Detection Research. Ontario, School of Computing, 2007. 115 p.
2) Grune D., Jacobs C. Parsing Techniques. Ithaca, Cornell University, 2008. 662 p.
3) Zezula P., Amato G., Dohnal V. Similarity Search - The Metric Space Approach, 2006. 233 p.
4) Aho A., Lam M., Sethi R. Compilers: Principles, Techniques, and Tools, 2006. 1035 p.
5) Scott M., Programming Language Pragmatics, 2006. 867 p.
6) Оре О. Теория графов. М.: Наука. 2003. 336 с.
7) Харари Ф. Теория графов М.: Мир. 2003. 300 с.
8) Чувилин К. В. Эффективный алгоритм сравнения документов в формате LaTeX // Компьютерные исследования и моделирование. МФТИ 2015 Т.7 №2 С. 329–345.