Существующие решения поиска дубликатов кода

PMD

Основные принципы:

- Michael Wise's Greedy String Tiling algorithm
  Создается временная таблица слов, используемых в коде (tokens). Для каждого слова записываются позиции, на которых оно встречается. Затем производится обработка таблицы, в результате которой рядом стоящие слова преобразуются в одну последовательность. Таким образом, обнаруживаются дубликаты.
  Более подробно об этом алгоритме можно прочитать тут:
  http://onjava.com/pub/a/onjava/2003/03/12/pmd_cpd.html
- Преобразование Барроуза-Уиллера
  Записываются все циклические сдвиги исходной последовательности в алфавитном порядке. Затем кодируются последовательность из последних символов циклических сдвигов и номер исходной последовательности. По закодированной последовательности можно однозначно восстановить исходную. Дубликаты ищутся на основе сравнения закодированных последовательностей.
  Более подробно алгоритм описан тут: http://marknelson.us/1996/09/01/bwt/
- Алгоритм Рабина-Карпа
  Для поиска дубликатов используется основная идея алгоритма Рабина-Карпа, где вместо строк сравниваются их хеши.
  Подробнее тут: http://xlinux.nist.gov/dads//HTML/karpRabin.html

IntelliJ IDEA

Находит структурные дубликаты.

Simian

http://www.harukizaemon.com/simian/index.html

Sonar

Основан на PMD.

Подробнее на http://www.sonarsource.org/manage-duplicated-code-with-sonar/

Другие средства поиска дубликатов перечислены на http://ru.wikipedia.org/wiki/%D0%94%D1%83%D0%B1%D0%BB%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%BA%D0%BE%D0%B4%D0%B0.