코드클론 표본 집합체 자동 생성기

이 효 섭

한양대학교 프로그래밍언어연구실

2011. 06. 27

PLASSE RESEARCH ROSAEC

개념

- **코드클론이란?**

  소스 프로그램에서 구문적 생김새가 동일한 코드 조각

- **코드클론 표본 집합체** (reference corpus)
  - ▷ 클론의 미탐(recall) 여부를 확인하기 위해 참조하는 클론들의 모음
  - ▷ 도구의 성능을 판별하는 기준으로 사용

PLASSE RESEARCH

기존 코드클론 표본 집합체의 수동 생성

- **생성방법** [Bellon et al., TSE 2007]
  - 여러 도구에서 찾아낸 클론들의 병합 (union)
  - 여러 도구에서 찾아낸 클론들의 코드 중 겹치는 부분(intersection)
  - 여러 도구에서 공동으로 찾아낸 클론
- **사용한 도구**

기반기술	도구	저자
Token	Dup	Brenda S. Baker
	CCFinder	Toshihiro Kamiya
PDG	Duplix	Jens Krinke
Function metrics	CLAN	Ettore Merlo
Text	Duploc	Matthias Rieger
AST	CloneDR	Ira D. Baxter

PLASSE RESEARCH

기존 코드클론 탐지 기법

- 기존 코드클론 도구들은 토큰들의 나열인 토큰열 비교

기반기술	도구
Token	Dup
	CCFinder
PDG	Duplix
Function metrics	CLAN
Text	Duploc
AST	CloneDR

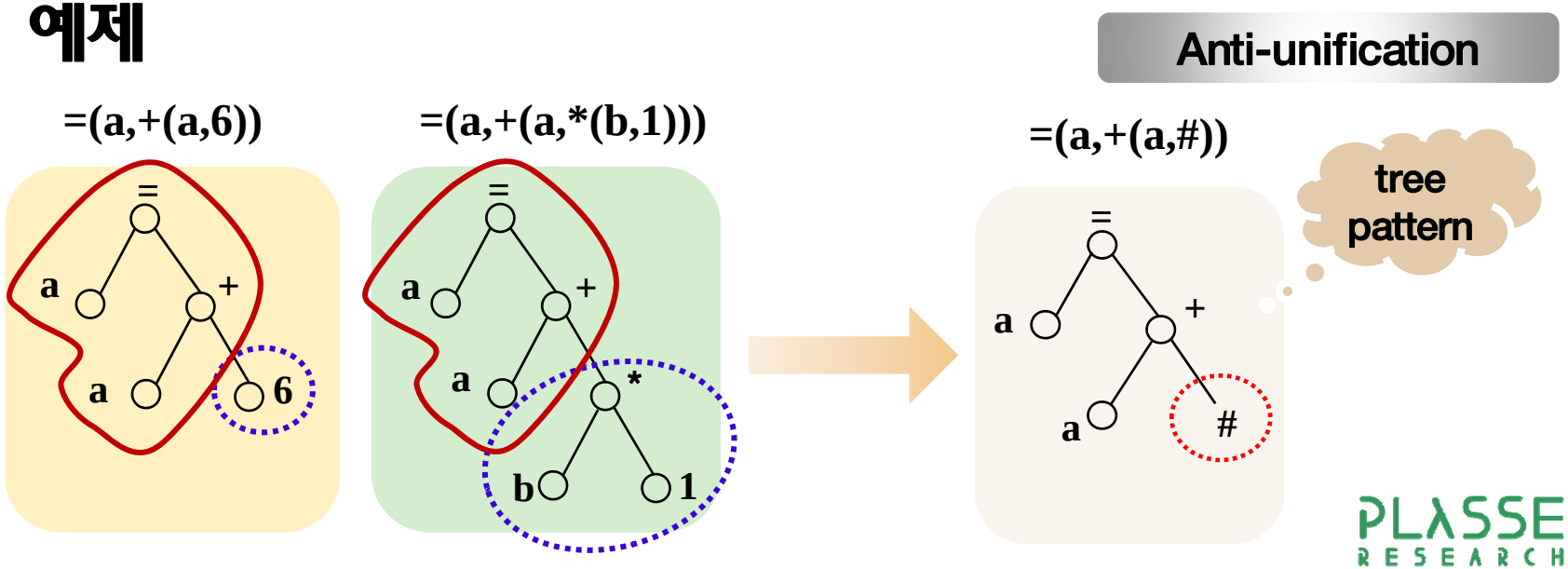
- 토큰열 기반 방식
  - ▷ 1차원 비교로, 프로그램 구문구조를 고려하지 않아 틀린 클론을 찾기 쉽다. (false positive)
- 의존 그래프 기반 방식, 계량치 기반 방식
  - ▷ 모양이 다른 코드의 그래프나 계량치가 동일할 수 있다. (false positive)
- 문자열 기반 방식
  - ▷ 단순한 프로그램 포맷변경에 민감하다. (false negative)

정확한 코드 클론을 찾기 위해서는
코드의 구문구조 정보가 포함되어 있는 AST를
비교하는 방식이 적절하다.

PLASSE RESEARCH

우리의 방식

- 코드내의 클론을 정확하고 빠짐없이 찾는 것이 목표
- 방법
  - ▷ 코드의 구문구조 정보가 포함되어 있는 AST를 비교
  - ▷ AST를 샅샅이 뒤져서 비교하되, 이미 방문한 트리는 재탐지하지 않도록 하여 계산비용을 줄인다.
- 예제



구현 및 실험대상

- **구현**
  - ▷ 구현언어 : Objective Caml 3.09, Python 2.5.1
  - ▷ 파서
    - CIL 1.3.6 (C Intermediate Language)
    - Joust 0.8 for Java

- **실험대상**

응용 프로그램	크기(MB)	파일수	줄수
netbeans-javadoc	0.69	97	14,301
eclipse-ant	1.35	149	29,880
eclipse-jdtcore	6.37	687	135,675
j2sdk1.4.0-javax-swing	8.32	533	202,943

PLASSE RESEARCH

상용도구 CloneDR와의 비교

- **정확도**
  - ▷ 오탐 없음 (No false positive)
  - ▷ CloneDR가 찾아낸 클론을 100% 모두 포함

- **찾아낸 클론의 개수**

응용 프로그램	CloneDR	우리도구
netbeans-javadoc	66	177
eclipse-ant	122	247
eclipse-jdtcore	930	2,791
j2sdk1.4.0-javax-swing	529	1,517

PLASSE RESEARCH

상용도구 CloneDR와의 비교

- **정확도**
  - ▷ 오탐 없음 (No false positive)
  - ▷ CloneDR가 찾아낸 클론을 100% 모두 포함

- **클론을 찾는데 걸린 시간**

응용 프로그램	CloneDR	우리도구
netbeans-javadoc	0h 00m 38s	0h 01m 40s
eclipse-ant	0h 01m 22s	0h 03m 56s
eclipse-jdtcore	0h 28m 39s	2h 40m 11s
j2sdk1.4.0-javax-swing	0h 51m 34s	5h 06m 09s

PLASSE RESEARCH

Bellon의 클론 표본 집합체와의 비교

- 기존 표본 집합체를 거의 대부분 포함한다.

응용 프로그램	기존 표본집합체	우리의 표본 집합체 자동 생성기	
		기본값	기본값 변경
netbeans-javadoc	55	44 (80.0%)	54 (98.2%)
eclipse-ant	30	28 (93.3%)	30 (100%)
eclipse-jdtcore	1,345	62 (93.0%)	62 (93.0%)
j2sdk1.4.0-javax-swing	777	33 (86.8%)	38 (100%)

5%

우리 도구에서 입력값은 총 4개로,
클론의 노드 수, 클론에서 일치하지 않는 부분의 개수, 일치하지 않는 부분의 크기와 최소 줄 수이다.

PLASSE RESEARCH

Bellon의 클론 표본 집합체와의 비교

- 기존 표본 집합체를 거의 대부분 포함한다.

응용 프로그램	기존 표본집합체	우리의 표본 집합체 자동 생성기	
		기본값	기본값 변경
netbeans-javadoc	55	44 (80.0%)	54 (98.2%)
eclipse-ant	30	28 (93.3%)	30 (100%)
eclipse-jdtcore	67	62 (93.0%)	62 (93.0%)
j2sdk1.4.0-javax-swing	38	33 (86.8%)	38 (100%)

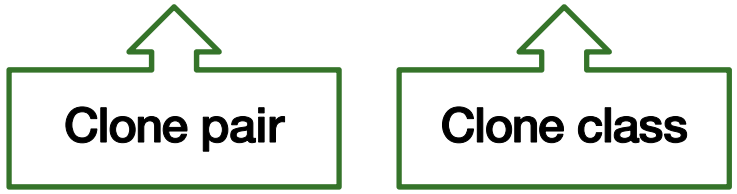
우리 도구에서 입력값은 총 4개로,
클론의 노드 수, 클론에서 일치하지 않는 부분의 개수, 일치하지 않는 부분의 크기와 최소 줄 수이다.

PLASSE RESEARCH

Bellon의 클론 표본 집합체와의 비교

- 기존 표본 집합체보다 더 많은 클론을 찾아낸다.

응용 프로그램	기존 표본집합체	우리의 표본 집합체 자동 생성기	
		기본값	clone pair 개수
netbeans-javadoc	55	113	80 (70.8%)
eclipse-ant	30	146	120 (82.2%)
eclipse-jdtcore	1,345	2,618	1,537 (58.7%)
j2sdk1.4.0-javax-swing	777	2,196	1,650 ㄴ(75.1%)



PLASSE RESEARCH

지금까지 한일과 할 일

- **지금까지 한 일**
  - ▷ 정확하면서 빠짐없이 클론을 찾는 알고리즘 개발 및 구현
  - ▷ 우리의 도구는 기존 코드클론 표본 집합체의 클론들을 거의 대부분 포함하면서 더 많은 클론들을 자동으로 탐지

- **할 일**
  - ▷ 기존 코드클론 표본집합체보다 정확히 얼마나 더 많은 클론을 찾아내는지 조사
  - ▷ 지원 언어 확대
  - ▷ 속도 개선
  - ▷ 다른 응용 분야의 적용
    - 리팩토링, 소프트웨어 형상관리, 복사 후 붙이기 오류 찾기, 복제 감정평가

PLASSE RESEARCH