Detección de reutilización de código fuente monolingüe y translingüe

Enrique Flores Sáez - eflores@dsic.upv.es
Programa de Doctorado en Informática
Directores: Lidia Moreno y Paolo Rosso
Departamento de Sistemas Informáticos y Computación

Introducción

Con el crecimiento de Internet la reutilización de código fuente ha aumentado al tener una mayor facilidad de acceso a la información. Además, una de las premisas de la programación consiste en que si "una algoritmo ya está implementado, no lo programes de nuevo, reutilízalo", ya sea utilizando librerías, código con licencias abiertas, etc.

Algunas áreas de interés para este tipo de sistemas son:

Ámbito académico: Posibles copias entre trabajos de alumnos para una misma tarea.

Empresas de software: Pérdidas millonarias en litigios por violación de licencias o patentes. Ej. Oracle vs Google

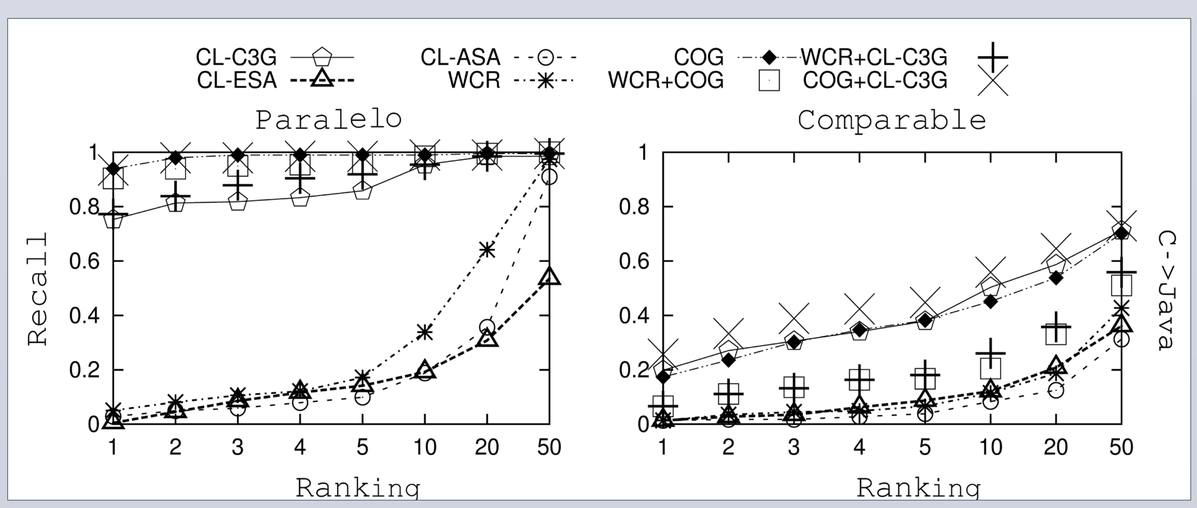
Repositorios de código fuente: Localización de un algoritmo implementado en diferentes lenguajes de programación.

Objetivos

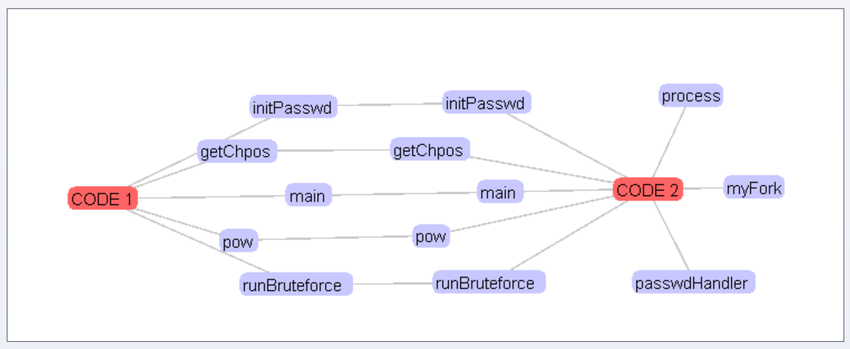
- Proponer sistemas que permitan detectar casos de reutilización de código fuente, a nivel monolingüe y translingüe.
- Aplicar sobre códigos fuente técnicas que han sido eficaces en la detección de reutilización de textos en lenguaje natural, considerando como un lenguaje más los lenguajes de programación.
- Identificar y abordar las principales técnicas utilizadas en modificación de códigos para evitar ser detectados.
- Facilitar la detección de uso ilegitimo de código fuente.
- Facilitar la búsqueda de códigos fuente similares para reducir costes temporales y minimizar errores en los proyectos software.
- Implantar en asignaturas, competiciones de programación, etc. herramientas de detección de reutilización de código fuente como ayuda al profesor en el control de trabajos y exámenes.

Detección de reutilización translingüe en corpus paralelo y comparable

En la detección de reutilización translingüe se pueden reconocer dos tipos de escenarios distintos a explorar. El escenario paralelo se considera como caso de reutilización que un código fuente sea una "traducción" literal de otro código fuente. Esto se puede conseguir mediante traductores de código fuente o manualmente. El escenario comparable se consideran como casos de reutilización aquellos códigos que resuelven el mismo problema pero de manera distinta. En el marco de esta investigación se han aplicado (combinaciones de) varios modelos: modelos que no requerían recursos externos (COG, CL-C3G, WCR) para realizar la detección han mostrado un comportamiento notablemente mejor que los que requieren recursos (CL-ESA, CL-ASA).



DeSoCoRe: Sistema de detección de funciones similares



http://memex2.dsic.upv.es/DeSoCoRe/

DeSoCoRe es una herramienta web que permite detectar reutilización de código fuente entre lenguajes de programación realizando comparaciones a nivel de función. El usuario puede elegir el grado de similitud que debe existir entre dos funciones para ser consideradas reutilizadas. Si dos funciones tienen una similitud mayor de la establecida por el usuario, la aplicación muestra la relación gráficamente y permite visualizar el contenido que ha sido considerado como reutilizado. Queda en manos de un revisor humano la decisión de considerar como reutilizado o no un código fuente o parte de éste.

Detección de reutilización en Google Code Jam Contest

Google celebra cada año la competición Google Code Jam, una competición internacional de programación para desafiar a programadores profesionales y estudiantes de todo el mundo en la resolución de problemas algorítmicos complejos. La fase inicial consta de 6 tareas a resolver en 25h en cualquier ordenador. Se ha aplicado el sistema sobre más de 58 millones de pares de códigos fuente para obtener el ranking en base a su similitud. Se han analizado manualmente los 20 pares de códigos fuente más similares por cada uno de los 3 lenguajes y por cada tarea encontrándose 216 pares reutilizados de un total de 360.

Lenguaje de programación	Tarea1	Tarea2	Tarea3	Tarea4	Tarea5	Tarea6	Total
C++	20	18	18	18	8	0	**82**
Java	18	18	20	20	1	1	**78**
Python	17	15	20	4	0	0	**56**
Total por tarea	55	51	58	42	9	1	**216**

Competición de detección de reutilización de código fuente

Detection of SOurce COde Re-use es la primera competición internacional sobre detección de reutilización de código fuente. Se celebrará en diciembre en la India en el marco del **FIRE 2014** Forum for Information Retrieval Evaluation.

La tarea consiste en detectar reutilización entre un conjunto de códigos fuente escritos en un mismo lenguaje. Aunque la reutilización se encuentre a nivel monolingüe, los códigos fuente pueden estar escritos en los lenguajes de programación C y Java. En próximas ediciones la tarea se extenderá a nivel translingüe, es decir, también reutilización entre lenguajes.

Más info en: http://users.dsic.upv.es/grupos/nle/soco/

AGRADECIMIENTOS

• DIANA-APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01)