Resumen

La detección automática de reutilización en códigos fuente consiste en determinar si un (fragmento de) código ha sido creado considerando otro como fuente. El plagio y las ramificaciones en proyectos software son dos ejemplos de tipos de reutilización en códigos fuente. Con la llegada de la Web y los medios electrónicos ha crecido enormemente la facilidad de acceso a códigos fuente para ser leídos, copiados o modificados. Esto supone una gran tentación para programadores que, con propósitos de reducir costes (temporales o económicos), deciden utilizar códigos fuente previamente depurados y probados. Este fenómeno ha causado que expertos en lenguajes de programación estudien el problema.

La gran cantidad de recursos disponibles en la Web hace imposible un análisis manual de códigos fuente sospechosos de haber sido reutilizados. Por ello, existe una necesidad urgente de desarrollar herramientas automáticas capaces de detectar con precisión los casos de reutilización. Basándose en técnicas del procesamiento del lenguaje natural y recuperación de información, las herramientas de detección automáticas de reutilización son capaces de realizar multitud de comparaciones de códigos fuente de forma eficiente.

En esta tesis proponemos un conjunto de modelos que pueden aplicarse indistintamente a nivel monolingüe o translingüe. Es decir, se pueden comparar dos códigos que están escritos en el mismo, o en distinto, lenguaje de programación. Por lo tanto, nos permite realizar comparaciones entre casi cualquier par de lenguajes de programación a diferencia de las propuestas del estado de la cuestión. Inicialmente, hemos estudiado las modificaciones más comunes realizadas por los programadores para evitar ser detectados. Para tratar estas modificaciones y mejorar la detección, hemos propuesto una serie de preprocesos. Se han evaluado y analizado los modelos tanto en un escenario académico real como en un escenario de detección a gran escala. Finalmente, nuestras mejores propuestas se han comparado con otras propuestas del estado de la cuestión dentro de un mismo marco de evaluación.

Estas pruebas de nuestros modelos se han realizado mediante millones de comparaciones tanto a nivel monolingüe como translingüe empleando diversas técnicas que fueron efectivas al aplicarlas sobre textos escritos en lenguaje natural. La mayor parte de los recursos desarrollados en el marco de esta tesis están a libre disposición de la comunidad científica. Utilizando parte de estos recursos, hemos configurado dos escenarios (monolingües y translingües) de evaluación que son un referente para que actuales y futuros trabajos de investigación puedan ajustar y comparar sus propuestas.