Technische Universität München
Institut für Informatik
Software and Systems Engineering
Prof. Dr. Dr. h.c. Manfred Broy

Masterarbeit

Masterarbeit

Werkzeugarchitektur zur kontinuierlichen Erkennung von Verstößen gegen Open-Source-Lizenzen durch Clone Detection

Thematische Einordnung

Statische Programmanalyse, License Infringement, Clone Detection

Hintergrund

In der Softwareentwicklung wird immer wieder Wiederverwendung durch die Übernahme und Integration von Code aus anderen Codebasen betrieben. Beliebte Quellen hierfür sind beispielsweise frei im Internet zugängliche Codebasen. Häufig stehen derartige Codebasen unter Open-Source-Lizenzen. Dabei passiert es leicht, dass Entwickler Code in ihre Systeme übernehmen und dabei übersehen, dass die Lizenzbedingungen nicht mit der geplanten Lizenzierung des zu entwickelnden Systems kompatibel sind oder dass bestimmte Verpflichtungen (z.B. Erwähnung der Quelle) mit der Übernahme des Codes einhergehen. Besonders problematisch ist oftmals beispielsweise die Übernahme von Code, welcher unter der viralen GPL-Lizenz steht, da dies eine Verpflichtung zur Nutzung der GPL-Lizenz auch für die Zielcodebasis zur Folge hat und somit ggf. zu einer Veröffentlichungspflicht des eigenen Codes führt.
In der Praxis existieren bereits einfache Ansätze zur Identifikation von Open-Source-Codeanteilen in Codebasen. Diese sind jedoch sehr anfällig, wenn der übernommene Code leicht verändert wurde. Auf Basis der im Bereich der Klonerkennung (Clone Detection, Copy&Paste-Erkennung) etablierten Verfahren, wird im Rahmen dieser Arbeit ein robusteres Analyseverfahren konzipiert, welches übernommenen Fremdcode auch im Fall von Modifikationen identifizieren kann.

Konkrete Aufgabenstellung

Im Rahmen dieser Arbeit wird ein Analyseansatz konzipiert und prototypisch umgesetzt, mit dem Ziel, die Übernahme von Fremdcode in

eine Codebasis zu identifizieren, auch wenn dieser Fremdcode leicht modifiziert wurde. Modifikationen können hierbei das Löschen von Kommentaren, Umformatierung, Umbenennung von Bezeichnern etc. umfassen.
Um (modifizierte) Kopien zwischen einer Codebasis und einer Menge von Open-Source-Systemen erkennen zu können, muss der zu detektierende Open-Source-Code in geeigneter Form gesammelt und aufbereitet werden (z.B. in einem Index). Eine derartige Indizierung ist die Voraussetzung dafür, dass dieser über Index-basierte Clone Detection Verfahren als Kopien wiedererkannt werden kann. Insbesondere der Umfang des hierfür zu indizierenden Open-Source-Codes stellt hierbei eine Herausforderung dar, weil auch die Historie der sich über die Zeit ständig weiterentwickelnden Open-Source-Systeme betrachtet werden muss.

Hierfür sind folgende Schritte notwendig:

- Abschätzung des Umfangs der zu indizierenden Open-Source-Codebasen unter viralen Lizenzen
- Konzeption einer geeigneten Indizierungsstrategie
- Definition geeigneter Datenstrukturen zur Ablage aller über die Open-Source-Systeme notwendigen Metadaten
- Konzeption einer geeigneten Architektur, die ein Update der Cloneindizes ermöglicht
- Prototypische Umsetzung einer indexbasierten Clone Detection auf Basis der gewählten Indizierungsstrategie und der konzipierten Architektur im Analysewerkzeug Teamscale

In der Arbeit kann auf Vorarbeiten im Bereich der index-basierten Clone Detection aufgesetzt werden.

Voraussetzungen

- Alle Voraussetzungen der PO für Masterarbeiten
- Kenntnisse der Objekt-Orientierten Programmierung (Java)
- Experimentierfreude und Fähigkeit zum selbstständigen Arbeiten
- Erfahrungen mit Teamscale sind von Vorteil

Aufgabensteller

Prof. Dr. Dr. h.c. Manfred Broy

Betreuer

TUM: Dr. Elmar Jürgens (juergens@in.tum.de)
CQSE: Dr. Benjamin Hummel (hummel@cqse.eu)
CQSE: Dr. Martin Feilkas (feilkas@cqse.eu)

Masterarbeit