

軽量な類似度計算によるプロジェクト間の ソースファイル集合の再利用検出

伊藤 薫

大阪大学

石尾 隆

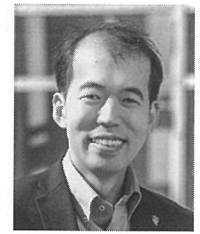
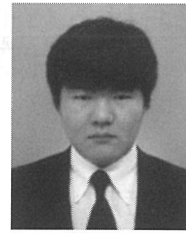
奈良先端科学技術大学院大学

神田 哲也

大阪大学

フェロー 井上 克郎

大阪大学



このたびは、私どもの論文[1]に対して電子情報通信学会論文賞という栄誉ある賞を頂き、大変光栄に存じます。本研究を進めるにあたり、御指導、御支援頂いた皆様、貴重な御意見を賜りました査読者の皆様、そして本研究を評価して下さいました選定委員や学会関係者の皆様に、この場を借りて御礼申し上げます。

ソフトウェア開発ではオープンソースソフトウェアとして公開されているライブラリを利用することが多く、セキュリティ対策や予期せぬバグを防ぐためには、利用しているライブラリを常に最新の状態に保ち、潜在的なリスクを把握する保守作業が重要です。ライブラリをバイナリ形式で利用したり、改変せずに利用する場合はどこから再利用したのかなど、再利用に際しての情報が失われたとしても再利用元の特定は容易ですが、再利用時に自身のソフトウェア向けに改変している場合は、単純な手法では再利用元の情報を特定することは困難です。

そこで本論文では、ソフトウェア中で改変して再利用しているライブラリについて、改変後のソースファイル集合とライブラリの各バージョンのソースファイル集合をそれぞれ比較することで高精度に再利用元のライブラリのバージョンを特定する手法を提案しています。既存手法

では計算速度が制約となりソースファイル単位での分析にとどまっていたが、本手法では高速な類似度計算手法である b -bit MinHash 法を導入することで処理の高速化を図り、ファイル群同士の比較により高精度に分析することが可能になりました。本手法を幾つかのソフトウェアとライブラリの組に適応したところ、愚直に類似度を計算する場合と比べて、平均で計算時間を 75.9%削減、使用メモリサイズを理論値で 99.5%削減しました。

ソフトウェア開発においてはオープンソースソフトウェアやライブラリはますます重要となっており、保守作業の効率化はソフトウェアの安全性を保つために非常に重要です。パッケージマネージャなどの仕組みが存在しない、若しくは普及していないプログラミング言語においては、本研究で提案した手法は非常に有用であると考えます。本研究がソフトウェア工学研究ひいてはソフトウェア開発に資するところがありましたら幸いです。

参考文献

- [1] 伊藤 薫, 石尾 隆, 神田 哲也, 井上 克郎, “軽量な類似度計算によるプロジェクト間のソースファイル集合の再利用検出,” 信学論 (D), vol.J103-D, no.7, pp.542–554, July 2020.