氏名（本籍）	東　裕之輔（和歌山県）
学位の種類	博　士（工学）
学位授与番号	甲 第１１２号
学位授与日付	令和５年３月２４日
専攻	システム工学専攻
学位論文題目	コンテナ開発におけるOSSの法的リスク特定自動化に関する研究
学位論文審査委員	（主査）教　授　風間　一洋 （副査）教　授　和田　俊和 准教授　大平　雅雄

論文内容の要旨

研究概要：モダンなソフトウェア開発では，開発コスト削減や最新技術の取り込みを主たる目的として，オープンソースソフトウェア（OSS）を再利用した開発が盛んに行われている．OSSはソースファイル，もしくはパッケージ単位にライセンスが設定されており，それらを遵守することで企業が開発するソフトウェアプロダクトにOSSを統合することができる．ただし，OSSのライセンスは全て均質的に設定されているわけではない．

OSSの中に別のOSSが含まれていることがあり，ファイル単位やパッケージ単位では例外的にライセンスが異なる場合がある．したがって，OSS再利用する場合は，OSSライセンスの特定や整合性の検証が必要となる．

さらに，近年，コンテナ仮想化技術の急速な普及によりOSSの再利用が加速化している．コンテナ仮想化は，OSカーネル部分はホストコンピュータと共有し，アプリケーションと動作環境のみをイメージ化することで，従来の仮想化技術より軽量でかつ複数の環境を跨いだシステム構築を可能にする．コンテナ仮想環境はデファクトスタンダードとなっているDockerにより提供される．

DockerはLinux Container（LXC）技術をベースとした技術を採用しており，Linuxディストリビューション上でのみ動作する．そのため，ホストマシンとカーネルを共有する各コンテナには多数のOSSパッケージが含まれる．また，コンテナ技術の台頭によりリリース方式も変化してきている．

従来，開発されたアプリケーションはソフトウェア単体でリリースされることが主流となっていたが，アプリケーションと動作環境を１つに統合にしたコンテナイメージファイル（Dockerイメージ）としてリリースされるようになってきている．

それらは，Docker HubなどのDockerイメージ共有サイトでリリースされるのが一般化しつつある．Docker Hubでは現在 9百万（2022年9月26日時点)以上のイメージが共有されている．

コンテナ開発では，OSSライセンスの精査はアプリケーションに加えDockerイメージ内のパッケージに対しても行う必要がある．OSSライセンスを正しく遵守しなかった場合，訴訟問題に発展し，社会的信用を失い，多額の損害賠償を負う可能性がある．コンテナ仮想化技術の普及以前にも訴訟問題は多数発生しているため，コンテナ開発の普及により法的リスクは増加するといえる．そのため，コンテナ開発の中で発生する法的リスクを正しく把握することが重要となるが，Docker イメージに含まれるコンポーネント数は数多いため，容易ではない．これらの法的リスクの把握を容易化するためには，アプリケーション，Docker イメージのそれぞれビルドする際に発生する２つの課題を解決する必要がある．その概要を図１に示す．

まず，ライセンスをソースファイル単位で特定するライセンス特定ツールがあるものの，ライセンスルールの保守に労力がかかり，手作業での特定が必要な未知ライセンス（Unknown）が数多く出力されるという課題がある．また，パッケージでは，１つのコンテナイメージに統合されるため，ライセンス間で条項に矛盾が発生していないか（互換性）を検証する必要がある．ただし，パッケージ情報が出揃わなければ検証できないと

いう課題がある．リリースの直前でパッケージを変更することになった場合，リリース時期を遅らせるしかない．

本研究では，コンテナ開発の法的リスクの特定を継続的に行うため，これらの課題を解決するための自動化技術を提案する．まず，ソースファイルのライセンス特定ツールの保守を自動化するため，未知ライセンス特定のためのライセンスルール（正規表現）の自動生成手法を提案する．また，Docker イメージ内のパッケージのライセンス互換性検証結果について機械学習による予測を行い開発早期に行えるように自動化する．

未知ライセンス特定のためのライセンスルール（正規表現）自動生成手法は，主に 6 つの処理群から構成される．(1) 未知ライセンスと判定されたソースファイルから，ライセンス記述を抽出する．(2) ライセンス記述が酷似しているライセンスを分類する．(3) ライセンスルールを抽出するのに適したクラスタを作成するため，未知ライセンスのライセンス記述の階層クラスタリングを行う，(4) 各クラスタに属するライセンス記述を編集距離でフィルタリングする．(5) 各クラスタに対し系列パターンマイニングアルゴリズムを適用し記述パターンを抽出する．(6) 記述パターンを正規表現に変換するしライセンスルールとして出力する．

提案手法を評価するために，FreeBSD-10.3.0，Linux-4.4.6，Debian-7.8.0 から検出した 1,821， 3,561， 2,838 件の未知ライセンス状態を用いたケーススタディを行った．その結果、提案手法は最小限のライセンスルールでより多くのライセンスを識別できること，提案手法で作成したライセンスルールを Ninka に追加することで，ライセンスルールのパフォーマンスが 2%〜10%向上することを示した．今後は，Ninka で使用されているライセンス文のキーワードの見直しや、表記ゆれの一般化などを行う予定である．

Dockerイメージのライセンス互換性検証自動化は，主に３つの処理群から構成される．

(1) Docker イメージをそのパッケージ情報によりベクトル化する

(2) 各開発進捗度に合わせて，パッケージ利用情報をマスキングする

(3) 各開発進捗度において多層パーセプトロンによる予測モデルを構築する

提案手法を評価するために，Github から抽出した Dockerfile をビルドして生成した 598 件の Docker イメージをもとに評価実験を行った．その結果、開発進捗度が 10%の時点でも適合率 94%，再現率 96%，F 値 95%の精度でライセンス検証結果を予測できることが分かった．

図 1：本研究で取り組む課題

論文審査の結果の要旨

Docker などのコンテナ型仮想化技術はアプリケーションと動作環境を統合することで配布・インストール・動作管理コストを大幅に削減できたが，同時に法的検討対象のライセンス数が大幅に増加し，人手でのライセンス違反発見が困難になり，ソフトウェア訴訟のリスクが増大しつつある．本論文では，コンテナ環境における OSS ソフトウェアを用いた開発におけるライセンス違反による訴訟リスクの自動検出を目指し，増大する新規ライセンス対応に注目したライセンス特定ツールのルールの自動作成手法と，開発段階における早期検出を目指した機械学習によるライセンス違反検出手法を提案した．予備審査時に指摘された論文構成と文章の修正や内容の明確化がほぼ満たされており，新規性と有効性が充分に示されていることから，博士論文に値すると判定した．

最終試験の結果の要旨

令和 5 年 2 月 13 日にシステム工学部北 1 号館 5 階 A508 会議室にて公聴会を実施した．参加者は審査委員会委員 3 名と他 5 名であった．公聴会は午前 9 時 30 分より開始し，60 分の発表の後に，40 分の質疑応答を行なった．質疑応答では，凝集型と分散型の階層クラスタリングの違い，Bug of Words の作成方法の詳細，多層パーセプトロンを用いた理由，増加するライセンスルール数の低減，ライセンス判定時の判定漏れ，ライセンス非互換性の種別，法的リスク回避という観点からのライセンス判定ツールの判定方針に関する質疑があり，どの質問に対しても適切な回答が得られた．上記の結果を総合的に判断し，学位申請者は博士の学位を得るに足る学識・能力を有していると判断したため，最終試験は合格と判定する．