ライセンス知識に基づくライセンス特定ツールの設計

真鍋 雄貴[†1] Daniel M. German [†2] 井上 克郎[†1]

開発者がオープンソースソフトウェアに含まれるソフトウェア部品（部品）を利用する場合，それらのソフトウェアライセンス（ライセンス）を理解し，尊守する必要がある．しかし，開発者がライセンスを正確に読解することは容易ではない．そこで，部品のライセンスが既知のどのライセンスと合致しているか調べること（特定）で，ライセンス理解にかかるコストを削減したい．本稿では，ライセンス特定ツールに求められる要件を整理し，それに基づき，どのようにライセンス特定ツールを設計するかを述べる．

Design of License knowledge-based License Identification Tool

Yuki Manabe ,[†1] Daniel M. German [†2] and Katuro Inoue[†1]

Developers which use software components included in open source software must understand and follow each license of them. However, it is difficult for developers to understand them accurately. To reduce this effort, we aim to retrieve the license matching the license of the component from known licenses. In this paper, we describes requirements for license identification tool and designs this tool on the basis of these requirements.

1. はじめに

開発コスト削減のため，オープンソースソフトウェア (OSS) に含まれる関数やクラスなどのソフトウェア部品（部品）が新規開発に利用されることが多い．

OSS の部品を利用するためには，多くの場合，ソースファイルのコメントとして記述されている，利用に関する許諾と許諾を受けるために負う義務を示したソフトウェアライセンス（ライセンス）を読み，理解し，それを尊守しなければならない．

一般に，ライセンスは自然言語で記述された長文であり，また，前提知識が必要だったり，他のファイルへの参照を含んだりするため，開発者がライセンスを正確に読解することは容易ではない．一方，多くの OSS 部品のライセンスは，定型の文で構成された既知のライセンスが用いられている場合が多い．この場合，ライセンスの文章全体を詳細に読解せずに，既知のどのライセンスと合致しているかが分かれば（これをライセンスの特定と呼ぶ），どのような許諾条件や義務があるか比較的容易に認識でき，その許諾条件や義務をそのまま適用することで，効率的な部品の利用を促進することができる．

しかし，ライセンスの特定も容易ではない．ソースコード中のライセンスの記述には，綴り間違いや単語の同義語への変更などの表記揺れ，ソースファイルにより変化する著者名や組織名が含まれるため，簡単な照合だけで特定することは困難である．そのため，細かい差異を認識して，ライセンスの特定を効率良く行えるシステムが望まれる．

本稿では，ライセンス特定ツールに求められる要件を述べ，それに基づくライセンス特定ツールの設計について述べる．

提案するツールを用いることで，再配布可能なソースファイルを容易に特定することができるようになり，プログラム解析技術をその適用結果だけでなく，使用したソースファイルとともに公開することが容易になる．そのため，一般の利用者が評価，検証可能になり，プログラム解析技術を利用してもらいやすくなるのではないかと考えている．

2. ライセンス特定ツールに求められる要件

1 節で述べたライセンス特定の問題と，OSS におけるライセンスの特徴に基づき，ライセンス特定ライセンス特定ツールに求められる要件を以下のように定めた．

要件１：現状の調査に基づいていて，多くのライセ

†1 大阪大学
Osaka University
†2 ビクトリア大学
University of Victoria, Canada

ンスを特定できる　実際に，どのようにライセンスの記述に表記揺れが含まれるか，類似したライセンスの記述があるかどうかについては明らかになっていない．そのため，実際に開発されている OSS を多数調査し，その多くのライセンスの特定ができる必要がある．さらに，調査により明らかにできた表記揺れや，類似したライセンスの記述を認識してライセンスを特定するため，細かい表現の差異を正しく認識できる機構を備えていることが望ましい．

要件２：新しいライセンスの記述への適合が容易　Open Source Initiative[⋆1] に承認されている OSS の定義を満たすライセンスは 76 種あるが，それ以外のライセンスも多数あり，必要に応じてそれらを特定できるようにしたい．また，ひとつのライセンスが複数の記述を持つ場合がある．これらに対応するため，新しいライセンスに容易に適合できるようにするための機構を備えていることが望ましい．

要件３：高速に処理できる　OSS の部品を利用するためには，事前に各部品のライセンスを特定しておくことが便利である，しかし OSS が大きいと，部品の数も増え，全てのライセンスの特定に要する時間も大きくなる．従って，個々のライセンスの特定には，高速な処理が望まれる．

ライセンス特定ツールの既存研究として，Fossology[1] と，ASLA[2] がある．

Fossology は bSAM アルゴリズムを用いて既存ライセンスの記述とコメントを比較し，近似の記述を持つライセンスを特定結果として出力する．しかし，このアルゴリズムの計算量が大きいため，要件３を満たさない．ASLA は既存ライセンスの記述に対応する正規表現のパターンを用いることで，差異を吸収して特定する．しかし，ふつう各記述の正規表現は複雑で容易に作成できないため，要件２を満たさない．また，これらの研究論文には背景の調査が示されていないため，要件１を満たしているかは不明である．

3. ツールの設計

図 1 に提案するツールの構成を示す．本ツールの設計に当たって FreeBSD，OpenBSD，Linux，Eclipse JDT，Mozilla といった大規模 OSS を調査した．調査ではコメントを文に分割し，どのような文が存在するかを調べた．その結果，多数の表記揺れを検出した．これらを吸収するため，本ツールでは，まず，既存のライセンスで用いられる各々の文（ライセンス文）を

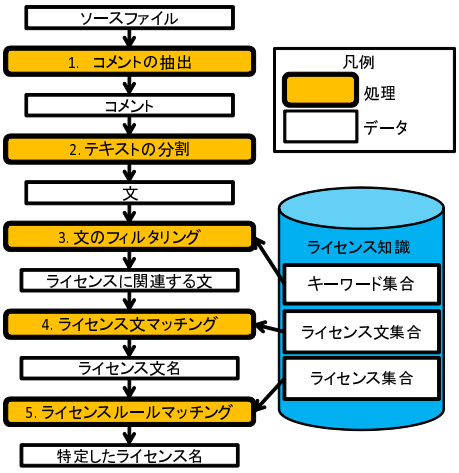

図 1　ライセンス特定ツールの構成

あらかじめ表記揺れを含む正規表現で表し，調べる対象の各文との照合を行う (4)．次に，得られたライセンス文の系列に対し，類似する系列を持つライセンスを探す (5)．このように照合を２段階に分割することにより，表記揺れの吸収，ライセンスの追加での変更を簡単に行えるようになる．これらはいずれも正規表現の照合であり，高速に実行可能である．また，ライセンスと関係のない文とライセンス文の照合を避けるため，特定のキーワードを持たない文を除去することでさらに高速化を図る (3)．

4. まとめと今後の課題

本稿では，ライセンス特定ツールの要件と，それに基づくライセンス特定ツールの設計について述べた．今後の課題として，ツールの実装と評価，そして，より多くの調査に基づくライセンス文やキーワードの追加，洗練がある．

謝辞　本研究は，文部科学省科学研究費補助金若手研究（B）（課題番号:20700024）の助成を得た．また，本研究は，文部科学省グローバルＣＯＥプログラム（研究拠点形成費）の補助によるものである．

参　考　文　献

1) Gobeille, R.: The FOSSology project, *MSR '08: Proceedings of the 2008 International Conference on Mining Software Repositories*, New York, NY, USA, ACM, pp.47–50 (2008).
2) Tuunanen, T., Koskinen, J. and Kärkkäinen, T.: Automated software license analysis, *Automated Software Engineering*, Vol.16, No.3-4, pp.455–490 (2009).

⋆1 http://www.opensource.org/