

未知ライセンス特定のための ライセンスルール自動抽出手法に関する研究

60160087 東 裕之輔

システム知能クラスタ

オープンソースソフトウェア工学研究室

OSSの再利用

- 近年，開発コスト削減のため，**オープンソースソフトウェア(OSS)**を再利用した開発が行なわれている



- ただし，OSSの再利用には，**OSSライセンスを遵守することが義務付けられている**

OSSライセンス

- OSSを利用する上で、著作者が利用者に許可していること、義務付けていることを記述した文書
- OSSライセンスには、多くの種類が存在する
 - OSSに関するコンサルティングを行なっているBlackDuck社調査では、2300種類以上存在する
 - 幅広く利用されているライセンス



GPLv3



LGPLv3



MPLv2.0



MIT License
by ©ExcaliburZero

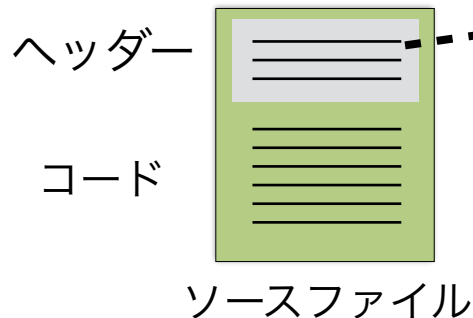


BSD License

ソースファイルのライセンス記述

- 一般的に、OSSライセンスはソースファイルのヘッダー部分に記述されている

BSD2のライセンス記述



```
#####  
# Redistribution and use in source and binary forms, with or without  
# modification, are permitted provided that the following conditions are met:  
#  
# 1. Redistributions of source code must retain the above copyright notice,  
#   this list of conditions and the following disclaimer.  
# 2. Redistributions in binary form must reproduce the above copyright  
#   notice, this list of conditions and the following disclaimer  
#   in the documentation and/or other materials provided  
#   with the distribution.  
#  
#####
```

- OSSライセンスはソースファイル単位で決定されている**
 - OSSライセンスを遵守するには、全てのソースファイルのライセンス記述を確認しなければならない

ライセンス特定ツール : Ninka[1]

- **ライセンスルール**を用いて, ソースファイルのライセンスを自動的に特定する
 - ライセンスルール : ライセンス記述とライセンス名の対応関係



[1] Ninka : Daniel M German, Yuki Manabe, and Katsuro Inoue, "A Sentence-Matching Method for Automatic License Identification of Source Code Files" In Proc. of ASE'10, pp. 437–446

Ninkaのライセンスルール

- ライセンス記述を文ごとに特定するため、その文数に応じたライセンスルールが必要

一般的なBSD2のライセンス記述

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

BSD2のライセンスルール

ルール名 : BSDpre

正規表現 : Redistribution(s)?
and use in source and binary[...]

ルール名 : BSDSourceCond.

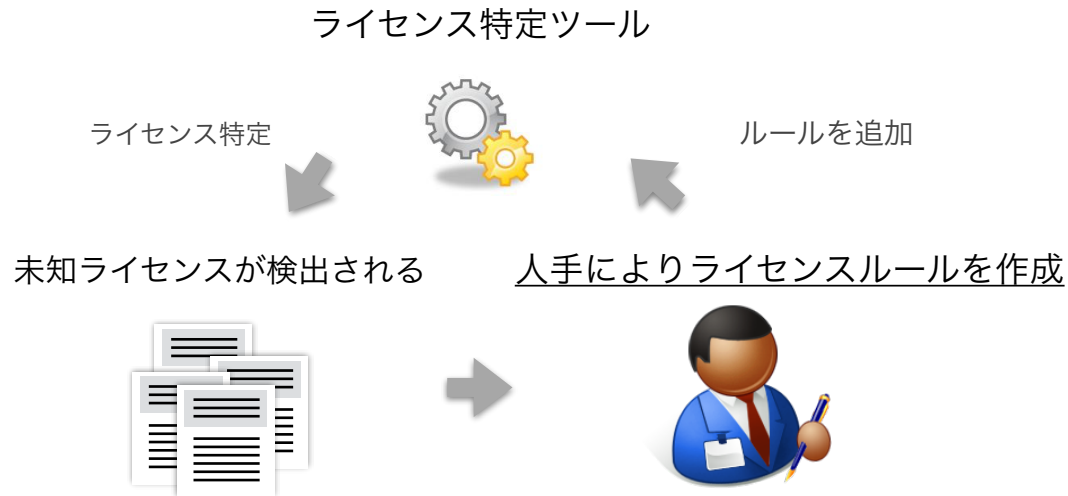
正規表現 : Redistribution(s)? of
source code[...]

ルール名 : BSDBinaryCond.

正規表現 : Redistribution(s)? in
binary form[...]

ライセンス特定ツールの課題

- 新しいライセンスや未知の表記ゆれは、年々増加するため、継続的にライセンスルールの作成・追加を行わなければならない

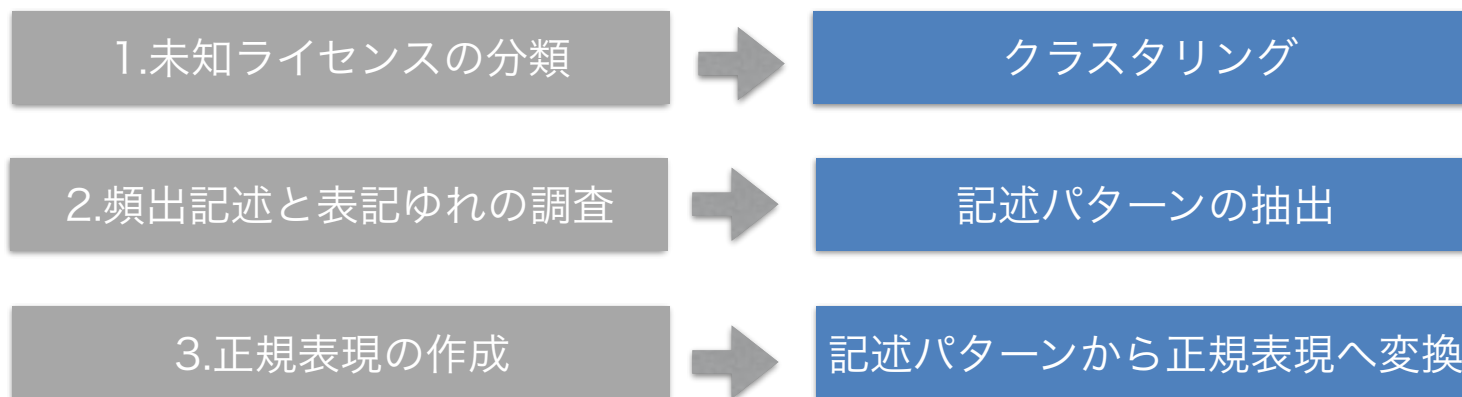


- しかし、Ninkaは評価実験の段階で17.2%のソースファイルに対し未知ライセンスを検出している[1]ため、大量の未知ライセンスが検出される恐れがある
- 大量の未知ライセンスからライセンスルールを人手により作成するのは労力がかかる

[1] Ninka : Daniel M German, Yuki Manabe, and Katsuro Inoue, "A Sentence-Matching Method for Automatic License Identification of Source Code Files" In Proc. of ASE'10, pp. 437–446

目的とアプローチ

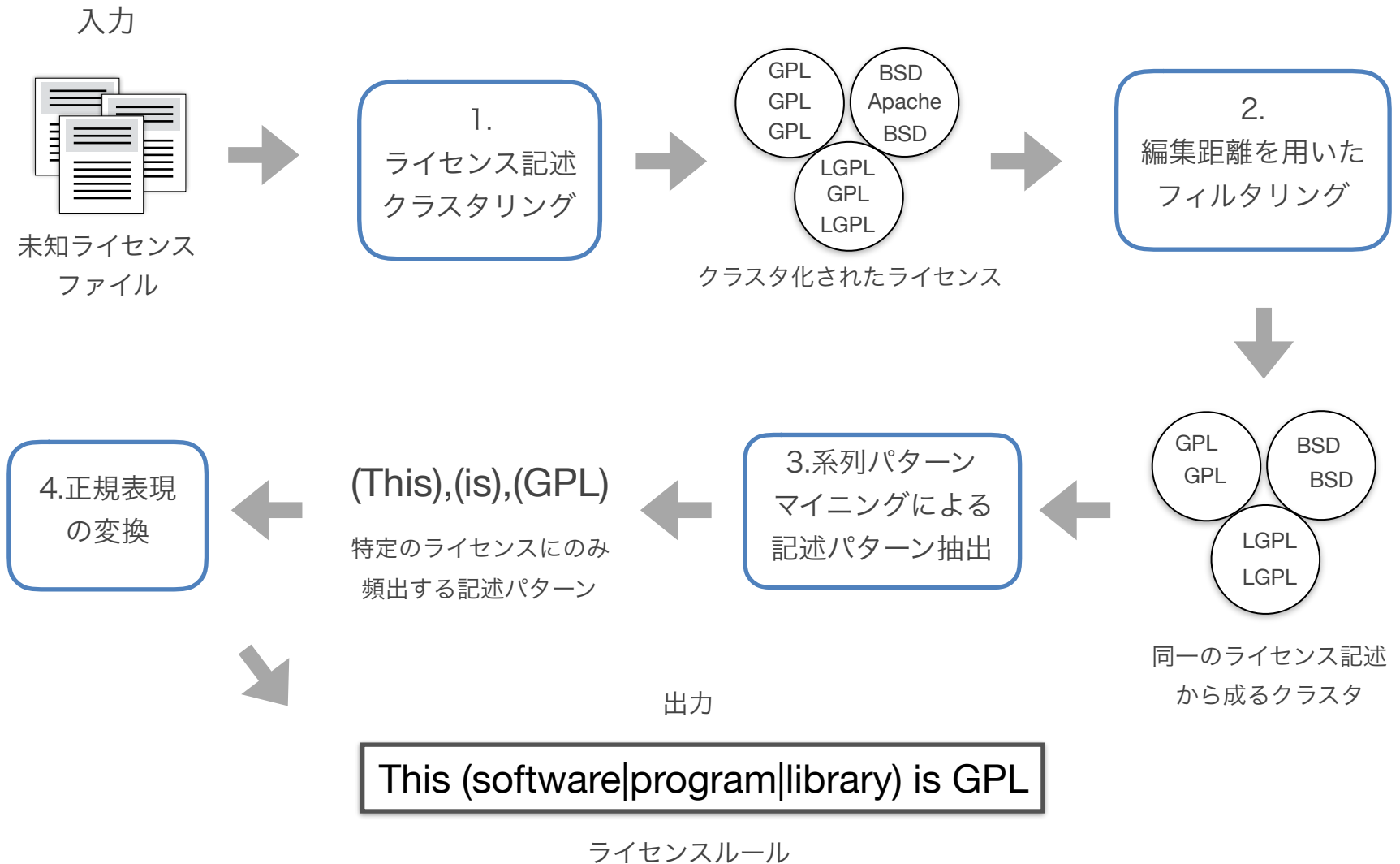
- ✧ 目的 : 未知ライセンスを特定するためのライセンスルールを自動抽出する手法の構築
- ✧ アプローチ
 - 手作業でのライセンスルール作成を機械的処理に置き換える



ライセンスルール作成手順
手作業による処理

自動化に必要な機械的な処理

ライセンスルール自動抽出手法



ケーススタディ

- 目的：提案手法の有用性の評価
- 対象ソフトウェア：FreeBSD, Linux, Debian
- Research Question
 - RQ1 : 全体の何%の検出された未知ライセンスのライセンス記述からライセンスルールを抽出できるか
 - RQ2 : 抽出したライセンスルールのうち1種類のライセンス記述のみに対応したライセンスルールは何%あるか
 - RQ3 : 複数の異なるライセンス記述を集約したライセンスルールを抽出することができたか
 - RQ4 : 抽出したライセンスルールによって何%の未知ライセンスが検出されたファイルを削減できるか

データセットの作成

- 未知ライセンスの取得方法

- FreeBSD-10.3.0, Linux-4.4.6

- ソースファイルのライセンスをライセンス特定ツールNinkaで特定し、検出された未知ライセンスをデータセットとして取得

- Debian-7.8.0

- 提案手法に不利な条件のデータセットを作成するため、Debianのソフトウェアパッケージに含まれる12,725の各ソフトウェアから未知ライセンスを1ファイルずつサンプリング

	未知ライセンス数	未知ライセンスの種類数（目視）
FreeBSD	1,821	69
Linux	3,561	33
Debian	2,838	194

実験方法

- 1.各データセットに提案手法を適用し，ライセンスルールを抽出
 - ただし，計算コストの問題で，Copyrightの文を除外して，系列パターンマイニングアルゴリズムBIDE[2]を適用
 - Copyrightの文：“Copyright (c) [yyyy] *** inc. All rights reserved”
- 2.抽出した各ライセンスルールを条件とし，ライセンスルールを抽出したソフトウェアにおける全ての未知ライセンスに対し，マッチングを行う
- 3.1種類のライセンス記述のみに対応したライセンスルールは何%あるか（RQ2）を確認する
- 4.抽出したライセンスルールによって何%の未知ライセンスが検出されたファイルを削減できるかを確認する（RQ4）

RQ2：1種類のライセンス記述のみに対応した ライセンスルールは何%あるか

抽出したライセンスルール	ライセンス	適合したライセンス種類数
You should have received a copy of the GNU Lesser General Public License[...]	LGPL	1
1 This program is distributed in the hope that it will be useful, but WITH[...]	LGPL	
software(is governed by/ under) the(terms of the)? CeCILLC license(under French law and abidingl as circulated)[...]	CeCILL-C	
2 - This Source Code Form is subject to the terms of the Mozilla Public License v2.0[...]	MPLv2.0	1
If a copy of the MPL was not distributed with this file, You can[...]	MPLv2.0	1

MPLv2.0は、Ninkaが特定することができないライセンス[5]

• RQ2の検証結果

	抽出したライセンスルール数	1種類のライセンスに対応したライセンスルール
FreeBSD	456	268 (58.8%)
Linux	269	213 (79.2%)
Debian	419	300 (71.6%)

-ライセンスルールを組み合わせて用いることで、ライセンスを一意に特定できると考えている

RQ4：抽出したライセンスルールによって何%の未知ライセンスがファイルを削減できるか

- 削減可能な未知ライセンス
 - ライセンス記述の全ての文に、抽出したライセンスルールのいずれかがマッチしているライセンス記述
- RQ4の検証結果

	未知ライセンス数※1	削減可能な未知ライセンス
FreeBSD	1,719	971 (56.4%)
Linux	1,941	1,125 (57.2%)
Debian	2,474	455 (18.4%)

※1：Copyright文のみのライセンス記述を除外した未知ライセンス数

- 1つのソフトウェアから検出された未知ライセンスに対しては有効である

まとめ

- 研究の目的：未知ライセンスを特定するためのライセンスルールを自動で抽出する
- ケーススタディの結果
 - Ninkaが対応していないライセンスルールを抽出できた
 - 抽出したライセンスルールのうち、平均して69.9%が1種類のライセンスのライセンス記述のみに対応した
 - 抽出したライセンスルールを用いることで、1つのソフトウェアから検出された未知ライセンスに対しては、約半数の未知ライセンスを削減可能であることが分かった
- 今後の展望
 - フィルタリングされたライセンス記述に対して、提案手法を繰り返し適用することで、さらにライセンスルールを抽出できないか確認する
 - Ninka以外のライセンス特定ツールにも応用できないか検証する