

ライセンスルール自動生成手法の構築

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

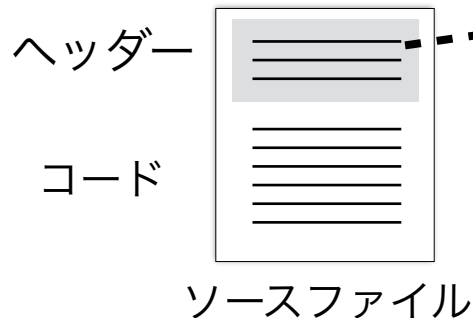


BSD License

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

OSSに含まれるソースファイル数

- OSSには、多くのソースファイルが含まれている



Linux

ソースファイル数 : 22,310

ライセンス種類数 : 53種類以上



FreeBSD®

ソースファイル数 : 20,759

ライセンス種類数 : 112種類以上

- 再利用するOSSに含まれている、全てのソースファイルのライセンスを手作業で確認するのは、労力がかかる

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章ごとに特定するため、その文章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

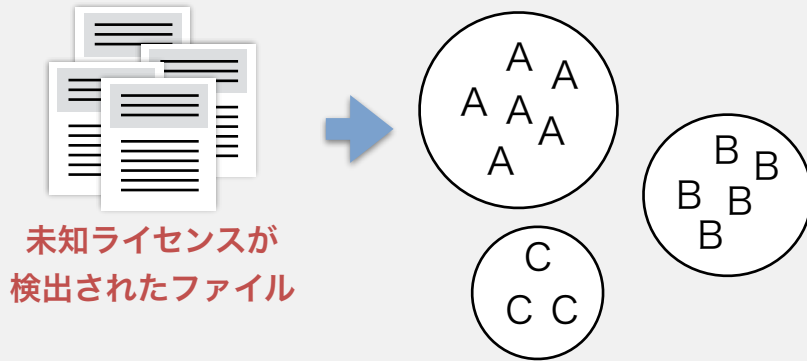
- 未知ライセンスの原因となる新しいライセンスや、未知の表記ゆれは、年々増加してきている
- ライセンス特定ツールは、**コンスタント**に新しいルールを追加し続けていくことが重要

ライセンス特定ツール



手作業でのライセンスルール作成手順

STEP1.未知ライセンス記述を ライセンス名によって分類



STEP2.各クラスタについて 頻出記述と表記ゆれを調査

頻出記述

This[...] is free software.
It under the term of the GPL.

[...]部の表記ゆれ

1	program
2	software
3	library

STEP3.ライセンスルール作成

1	This program is free ~
2	This software is free~
3	This library is free~

ライセンスルール

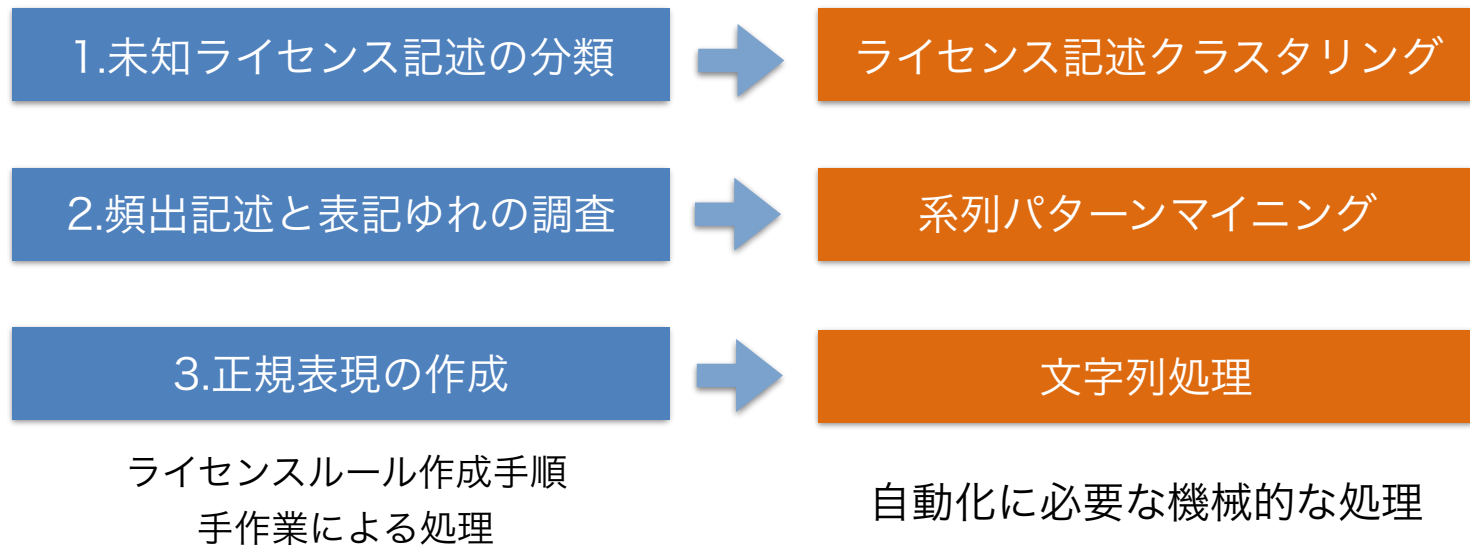
This (program|software|library) is free~

問題点

大量の未知ライセンスが
検出された場合、大変労力がかかる

本研究の目的とアプローチ

- ✧ 目的：ライセンスルール自動生成手法の構築
- ✧ アプローチ
 - ライセンスルール作成手順における各ステップを，機械的処理に置き換える



ライセンスルール自動生成における課題

- 単一のライセンスから成るクラスタのみからライセンスルールを生成する

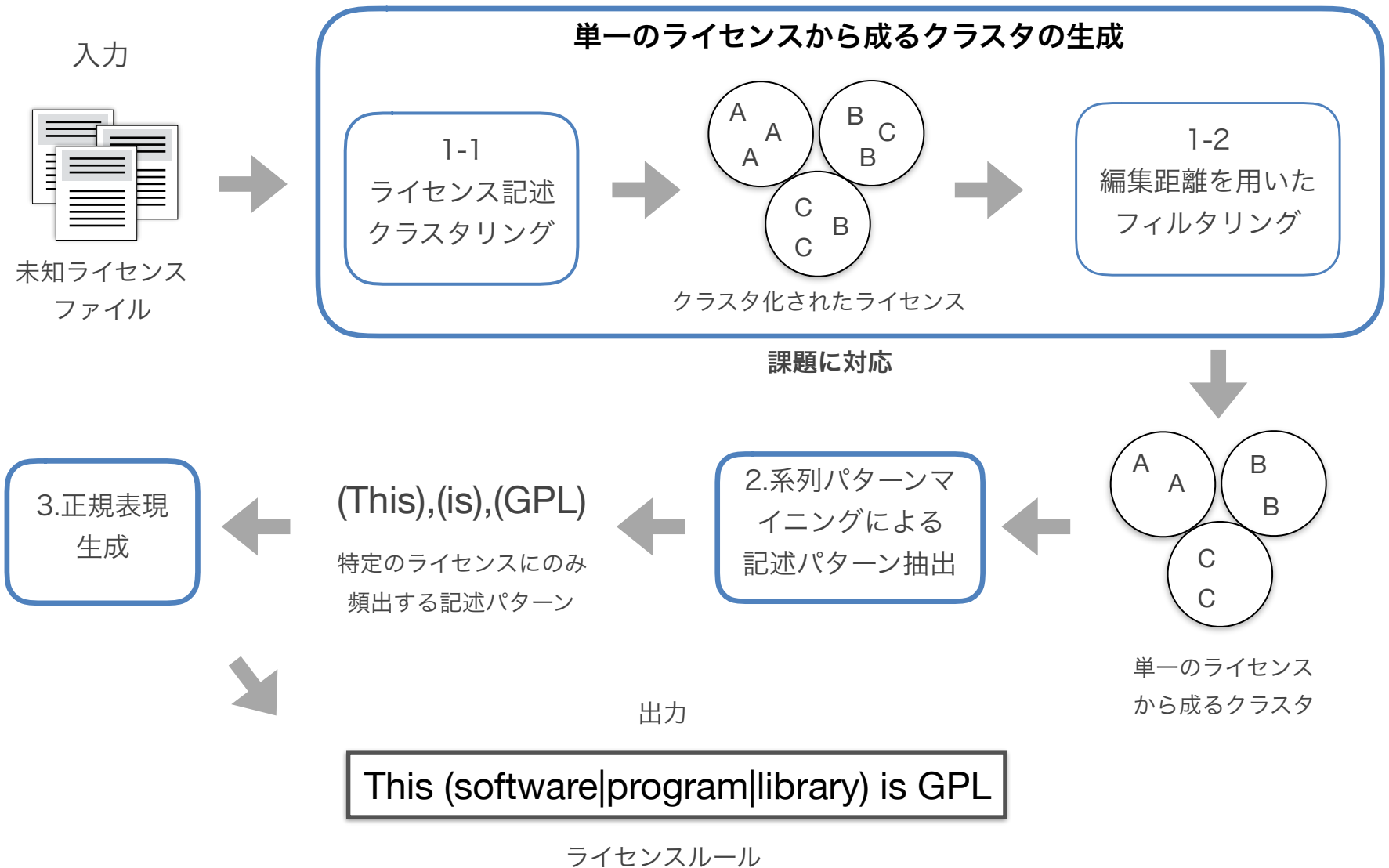
- 複数のライセンスが混入しているクラスタからは適切なライセンスルールを生成することができない



- ライセンス記述には、バージョンのみが異なるなどライセンス記述が酷似していて区別することが難しいものが存在する

- GPLv2, GPLv3など

提案手法の概要

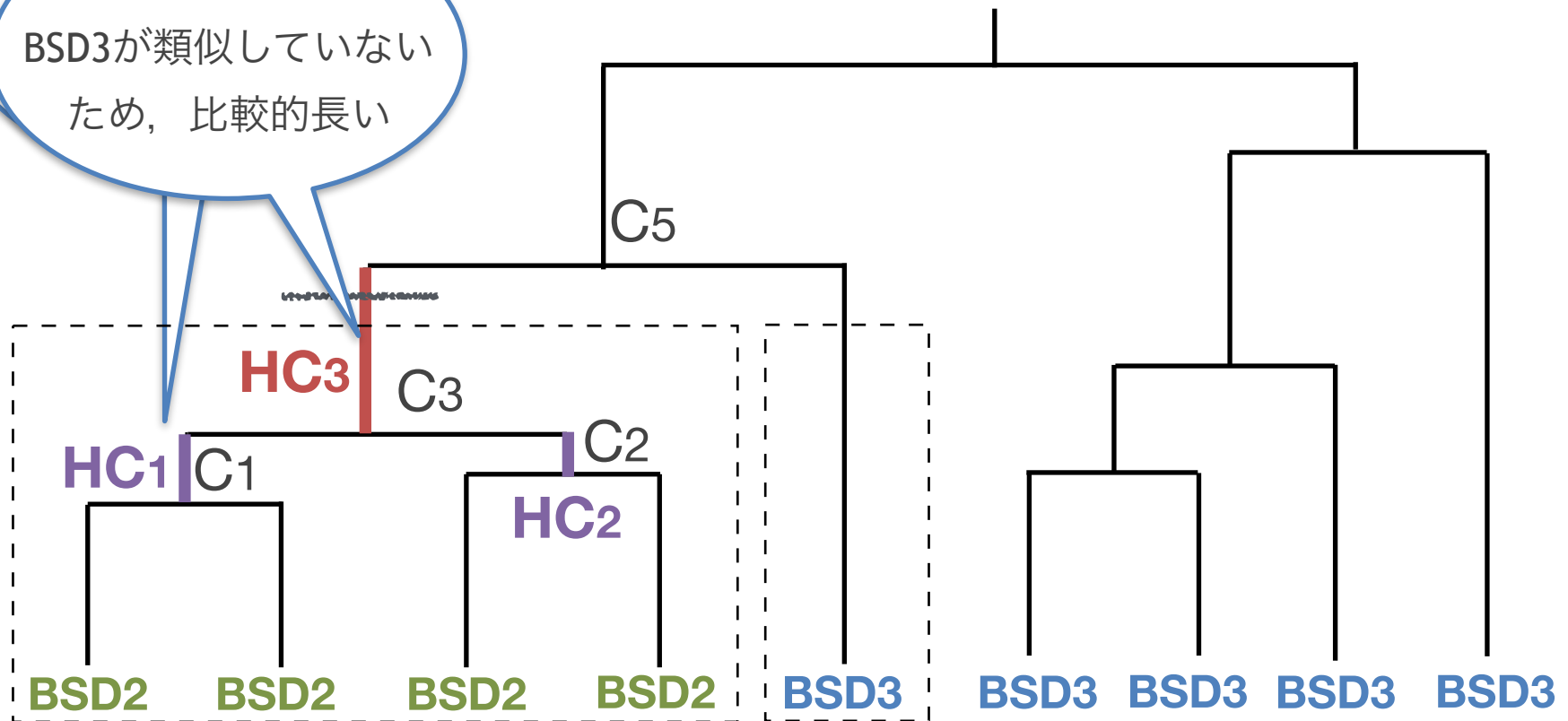


1-1 : ライセンス記述階層クラスタリング

- 次式が成り立つとき, 枝刈りを行う

- C5を作成しない : $((HC3 > HC1) \text{ かつ } (HC3 > HC2))$

HC3は, C3と
BSD3が類似していない
ため, 比較的長い



1-2：編集距離を用いたフィルタリング

- 各クラスタのライセンス記述を，編集距離を用いてフィルタリングし，単一のライセンスからなるクラスタを作成する

- 編集距離：文字列Aから文字列Bにするまでに必要な削除，置換，挿入の回数

- 例 “kitten”と”sitting” 編集距離=3

文字の出現順序も
距離に考慮することができる

- 本手法で用いる編集距離によって定義される類似度

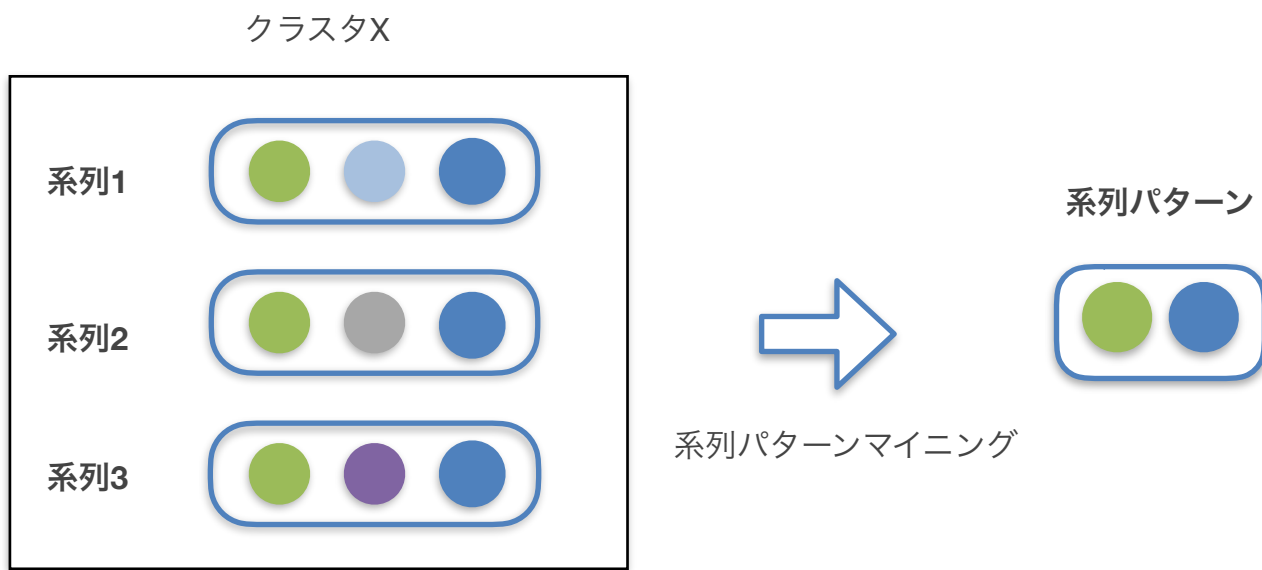
$$NLD = 1 - \frac{LD(A, B)}{\maxlength(A, B)}$$

NLD : 正規化された編集距離
LD : 編集距離
maxlength : 最大字数

- フィルタリング方法：総当たりで類似度を求め，類似度が94%以上あるライセンス記述のペアをクラスタから取り出す

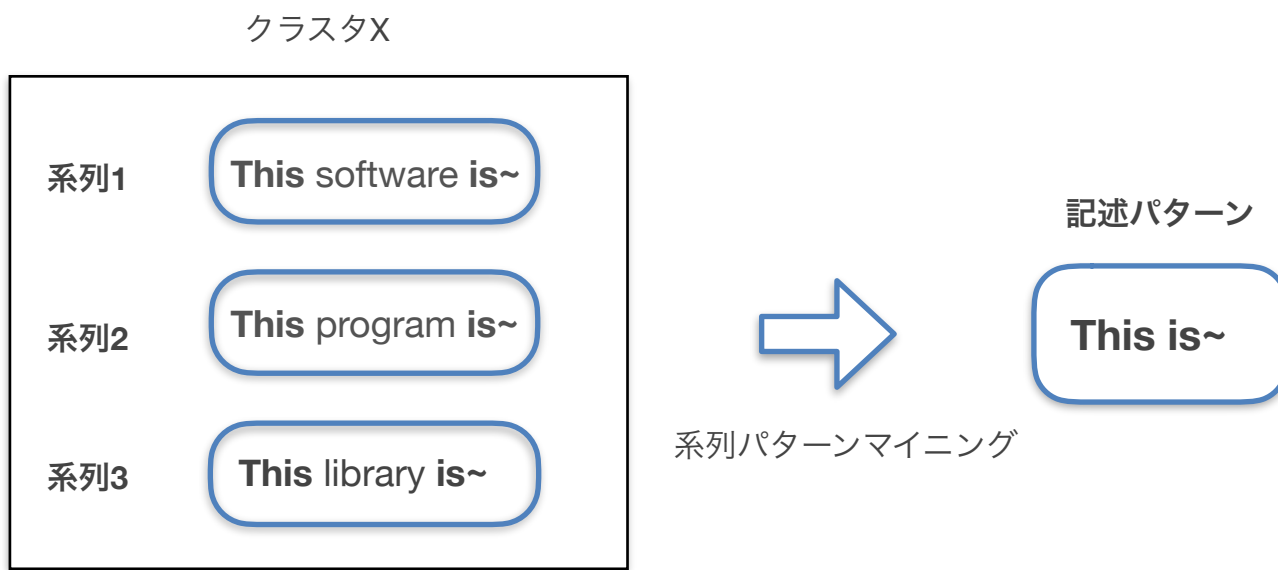
2.記述パターンの抽出

- 系列パターンマイニング：複数の系列に共通する要素の出現順序（系列パターン）をすべて列挙する
- 系列パターンマイニングを用いて、各クラスタに頻出する記述パターンを抽出する



2.記述パターンの抽出

- 系列パターンマイニング：複数の系列に共通する要素の出現順序（系列パターン）をすべて列挙する
- 系列パターンマイニングを用いて、各クラスタに頻出する記述パターンを抽出する



3.正規表現の生成

- 記述パターンが抽出されたライセンスの文章から表記揺れを抽出し，正規表現を生成する

記述パターン	This	.*	is	.*	free	.*	software
ライセンス文1	This	software	is		free		software
ライセンス文2	This	library	is		free		software
ライセンス文3	This	program	is		free		software
ライセンス文4	This		is		free		software

生成されるべきライセンスルール

This (software| library| program)? is free software

ケーススタディ

- 目的：提案手法により生成されたライセンスルールは未知ライセンスをどれくらい特定することができるか
- 対象プロジェクト：Debian, FreeBSD, Linux
- データセット
 - ソースファイルのライセンスをNinkaを用いて特定
 - 検出された未知ライセンスのライセンスを目視で確認

プロジェクト	未知ライセンス ファイル	ライセンス文	ライセンスの種類
Debian 7.8.0	2838	9487	194
FreeBSD 10.3.0	1821	9165	69
Linux 4.4.6	3561	7767	33

生成したライセンスルール

- Debian-7.8.0 : 419個のライセンスルールを生成
 - CeCILLライセンス
 - software(is governed by| under) the(terms of the)? CeCILL\ -C license(under French law| as circulated by CEA, CNRS) and(abiding by| INRIA at) the
- Linux : 270個のライセンスルールを生成
 - GPLライセンス
 - May be copied or modified under the terms of the GNU General Public License
 - This file is \ (c\) under GNU PUBLIC LICENSE
 - Released under the term of the GNU GPL v2
 - Released according to the GNU GPL, version 2 or any later version
- FreeBSD : 456個のライセンスルールを生成

実験結果

- ライセンス文単位での特定結果

プロジェクト	未知 ライセンス文	生成したルールと マッチしたライセンス文
Debian 7.8.0	9487	2155 (22.7%)
FreeBSD 10.3.0	9165	5082 (55.4%)
Linux 4.4.6	7767	3166 (40.8%)

- ファイル単位での特定結果

-全てのライセンス文章にマッチしたファイル

プロジェクト	未知ライセン スファイル	ライセンス記述の全ライセン ス文にマッチしたファイル
Debian 7.8.0	2838	210 (7.3%)
FreeBSD 10.3.0	1821	281 (15.4%)
Linux 4.4.6	3561	411 (11.5%)

まとめ

- 未知ライセンスが検出されたファイルから，ライセンスルールを自動生成する手法を提案
- ケーススタディの結果
- 今後の課題
 - Ninkaが特定できなかったライセンス文をどれくらい特定することができるかを調査
 - 適合率，再現率はどれくらい向上するかを調査

単一のライセンスから成るクラスタの生成

- 既存手法によって作成された未知ライセンス記述の各クラスタに対し，編集距離を用いたフィルタリングを行う
- 実験結果
 - 全体の95%のクラスタを単一のライセンスから成るクラスタにできた

Debian-7.8.0	フィルタリング	クラスタ数	単一のラ から成る	ライセンス記述に 関係ない記述が混入	フィルタリン グされた記述
2,446ファイル	フィルタリング前	246	170(69%)	76(31%)	-
	フィルタリング後	126	120(95%)	6(5%)	1960