

Clustering OSS License Statements Toward Automatic Generation of License Rules

Yunosuke Higashi

Wakayama Univ.

Japan

Manabe Yuki

Kumamoto Univ.

Japan

Masao Ohira

Wakayama Univ.

Japan

OSSライセンスとは

- OSSを利用する上での制限事項を定義したもの
 - 有名な条項 : GPLのソフトウェアを再利用して開発したソフトウェアのライセンスはGPLでなければならない
 - OSSライセンスの定義
 - 広義 : OSSのソースファイルに記述されているライセンス
 - 狭義 : OSIに承認されている69種類のライセンス

OSIに承認されているOSSライセンス



MPL



EPL



BSD



MIT

背景

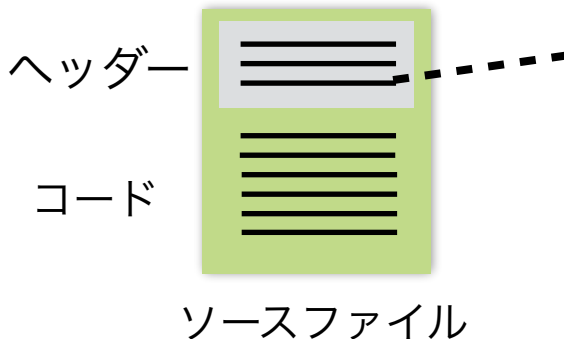
- 近年，開発コスト削減のため，OSSを再利用した開発が積極的に行われている
- ただし，OSSを再利用するには，ソースファイルのライセンスを遵守しなければならない
- 手作業によるライセンス特定は，労力がかかる
- 手作業によるライセンス特定を支援するため，自動でソースファイルのライセンスを特定するツールが複数提案されている

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



ライセンス記述



```
#####  
# This program is free software; you can redistribute it and/or modify  
# it under the terms of the GNU General Public License as published by  
# the Free Software Foundation; either version 3 of the License, or  
# (at your option) any later version.  
#  
# This program is distributed in the hope that it will be useful,  
# but WITHOUT ANY WARRANTY; without even the implied warranty of  
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the  
# GNU General Public License for more details.  
#  
# #####
```

ライセンス特定ツール[1][2]

- ライセンスルール（正規表現）を用いて，自動的にソースファイルのライセンスを特定する



[1] Ninka : Daniel M German, Yuki Manabe, and Katsuro Inoue, "A Sentence-Matching Method for Automatic License Identification of Source Code Files" In Proceedings of ASE, pages 437-446, September 2010

[2] Fossology :R. Gobeille, "The fossology project," in Proc. of MSR'08, 2008, pp. 47–50.

ライセンス特定ツールの弱点

- 新しいライセンスや、表記ゆれは、年々増加している
- 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必要である

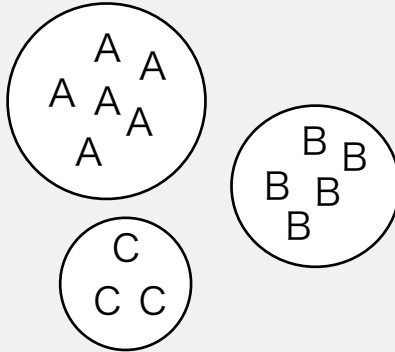


手作業によるルール作成

1. 未知ライセンスファイル集合を ライセンス毎に分類



Unknown
license files



2. 各グループにおいて 表記ゆれを調査

ライセンスグループAの表記ゆれ

This program is~ it under the term of the A license
This software is~ it under the term of the A license
This library is ~ it under the term of the A license

3. ライセンスルール作成

This program is ~
This software is ~
This library is ~

新ライセンスルール



This (program|software|library) is ~

問題点

大量の未知ライセンスファイルが検出された場合、大変な労力がかかる

未知ライセンスが検出される割合



ライセンス名	ファイル数	割合
UNKNOWN	397	未知1.3%
LGPLv3	28,240	既知94.7%
spdxBSD3	244	
LGPL2.1	31	
LGPLv2+	25	
SimpleLic	16	
GPLv2+	8	
LGPLv3,SeeFile	8	
X11	7	
SeeFile	6	
NONE	892	なし4%
合計	29,874	



Linux

ライセンス名	ファイル数	割合
UNKNOWN	3,561	未知17%
GPLv2	6,847	既知74%
GPLv2+	6,131	
GPL,SeeFile	923	
MIT	891	
GPL	668	
GPL2orOpenIB	314	
LGPLv3,SeeFile	201	
MITX11BSDvar	143	
SeeFile	118	
Other (42種類)	613	なし9%
NONE	1,900	
合計	22,310	

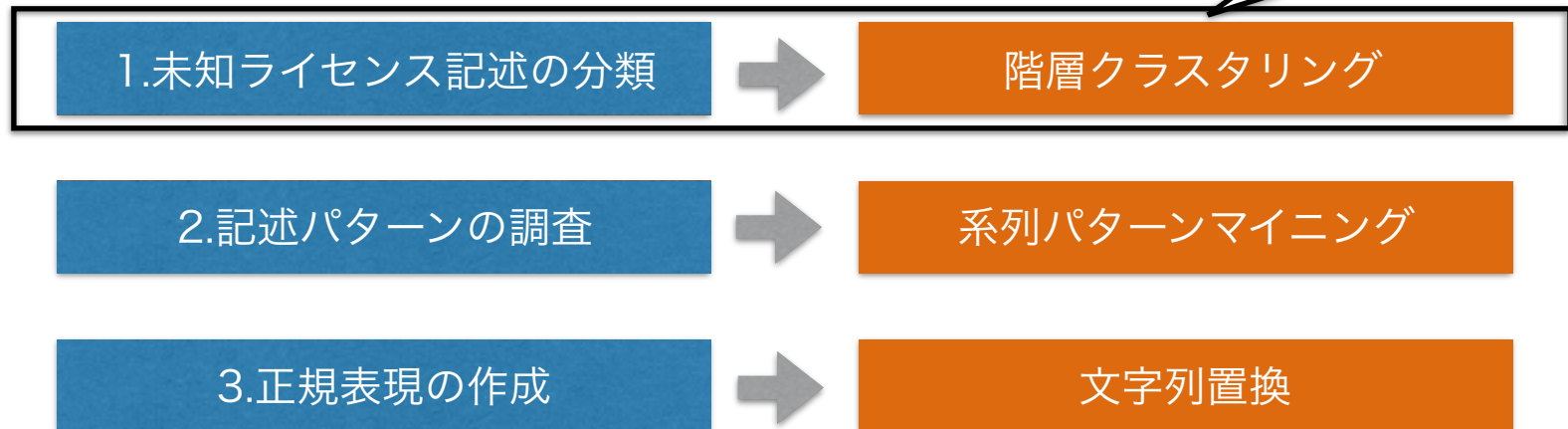
最終的なゴール

✧ ゴール：ライセンスルール候補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3までの各工程を、機械的処理に置き換える

まず、手順1の自動化を行う



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

技術的課題

- ステップ1自動化における技術的課題

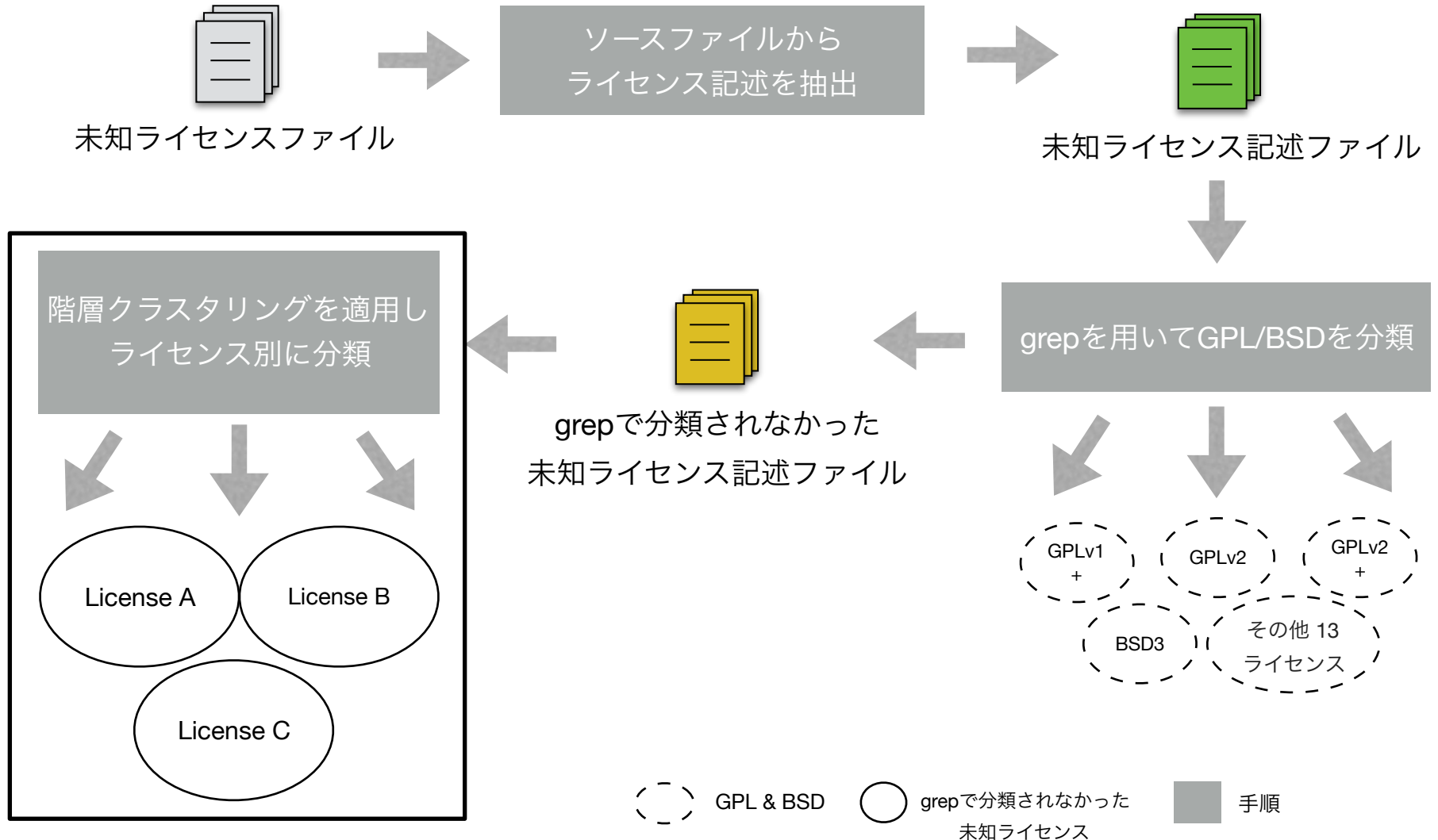
A. GPL/BSDの分類

- 予備調査で多くのGPL/BSDファイルが存在することが分かった
- GPL/BSDは、長い歴史を持つライセンスであるため、多くのOSSで使用されている
- 正確に分類すべきである

B. 未知ライセンスファイルの分類

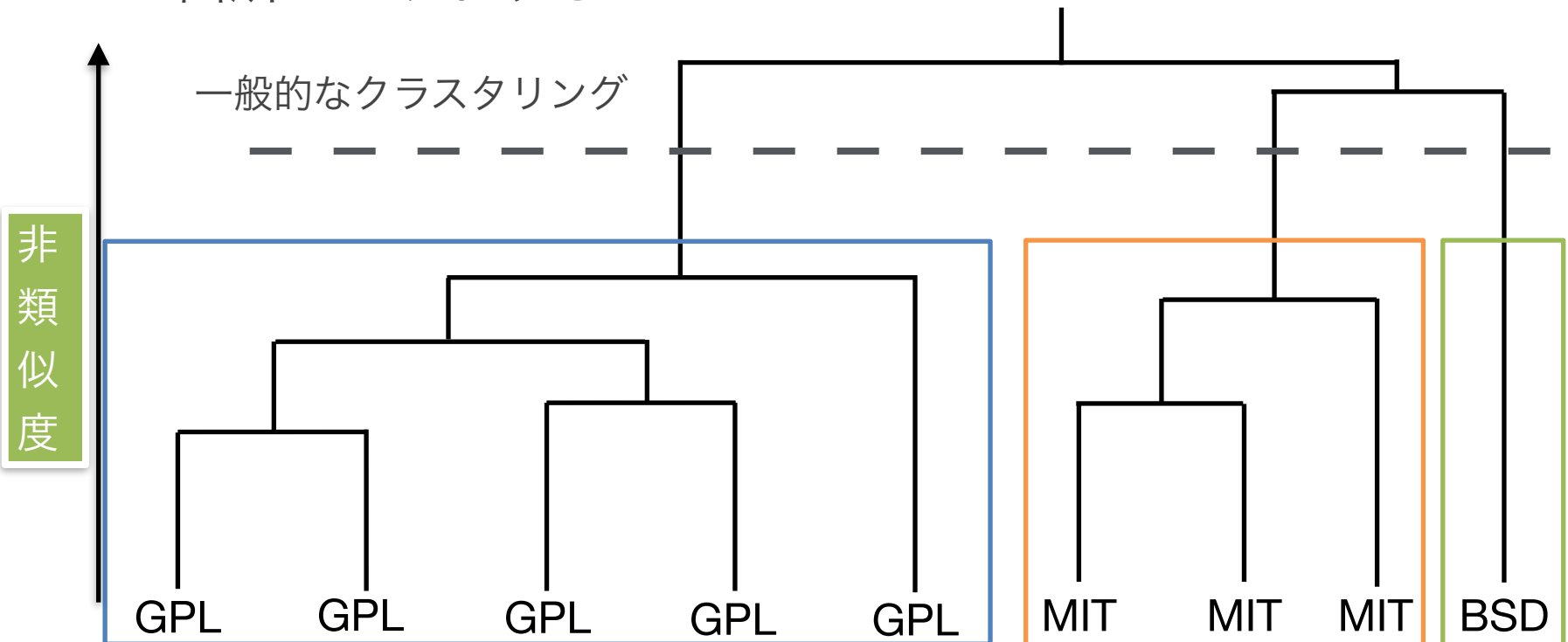
- 1種類のライセンスから成るグループを作成する必要がある
- 生成されたグループ数は、再利用するソフトウェアに存在するライセンスの種類数と一致するのが理想

提案手法の概要



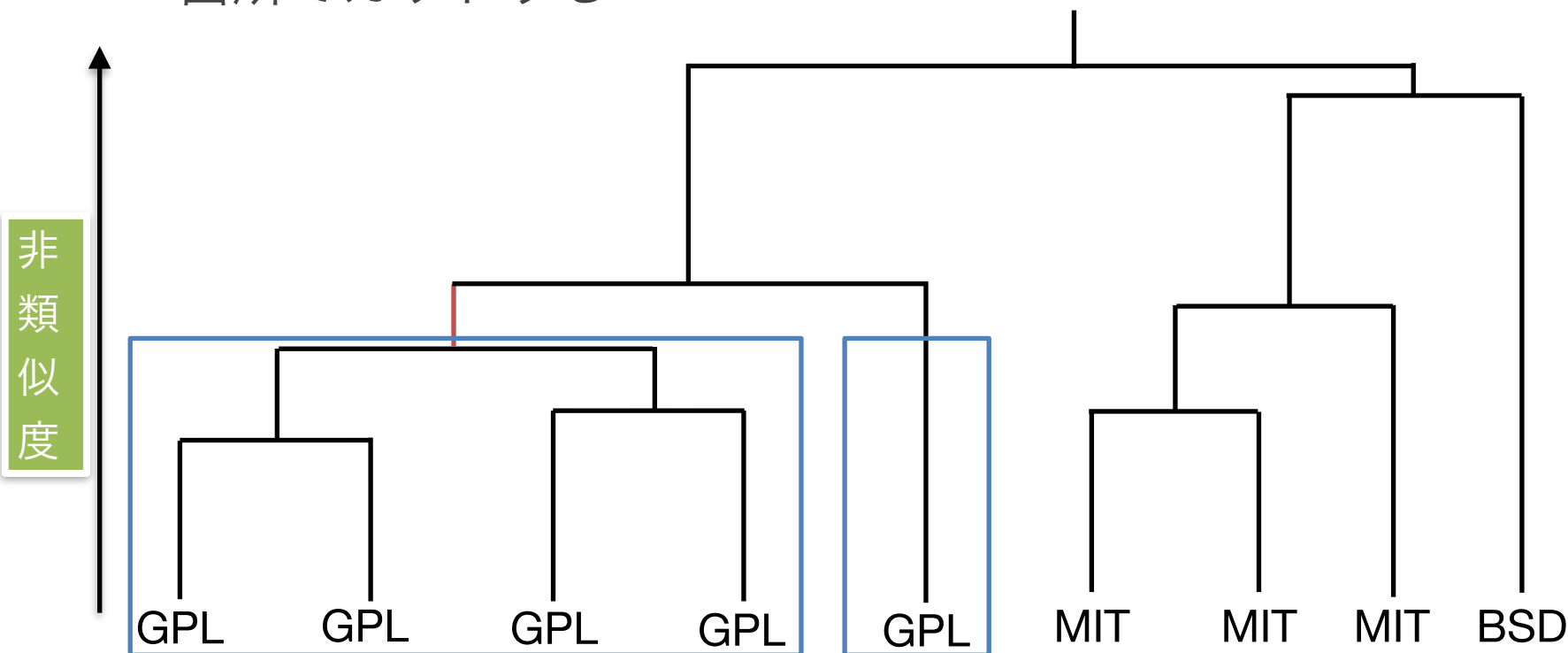
階層クラスタリングによるライセンスグループ生成

1. 未知ライセンス記述ファイルをBag of Wordsに変換
2. ウォード法による樹形図（デンドログラム）を生成
3. データ間の類似度に条件を設けてデンドログラムを複数箇所でカットする



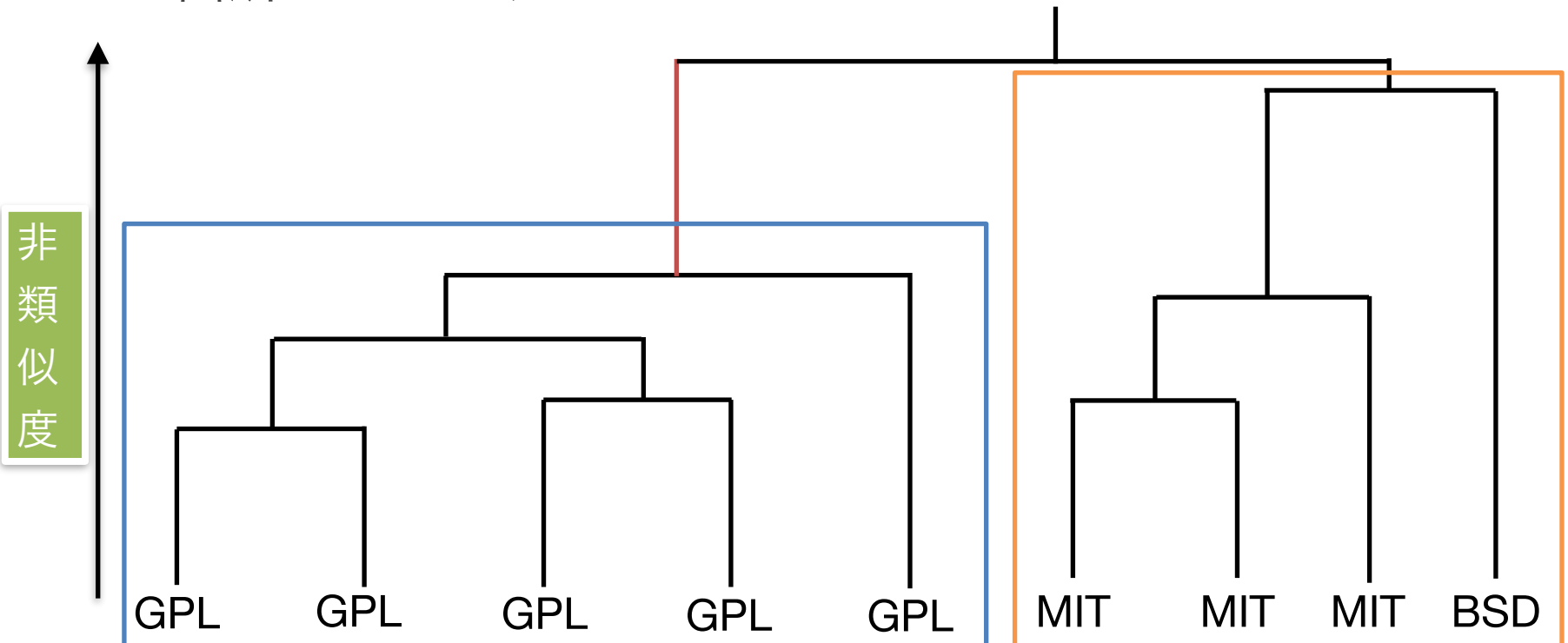
階層クラスタリングによるライセンスグループ生成

1. 未知ライセンス記述ファイルをBag of Wordsに変換
2. ウォード法による樹形図（デンドログラム）を生成
3. データ間の類似度に条件を設定しデンドログラムを複数箇所でカットする



階層クラスタリングによるライセンスグループ生成

1. 未知ライセンス記述ファイルをBag of Wordsに変換
2. ウォード法による樹形図（デンドログラム）を生成
3. データ間の類似度に条件を設定しデンドログラムを複数箇所でカットする



Case Study

- Purpose
 - ライセンスルール作成に適したクラスタをどれだけ作成できるかを調査する
- Target files
 - Debian v7.8.0からサンプリングした2,475の未知ライセンスファイル
 - 2,475ファイルのライセンスを目視で特定

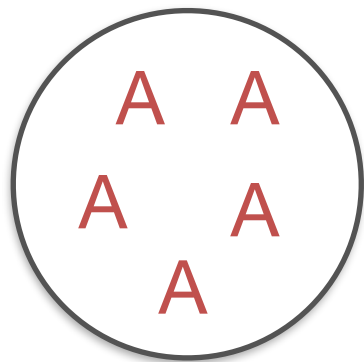
プロジェクト	未知ライセンスファイル	ライセンス数
Debian v7.8.0	2,475	118

- 一般的な階層クラスタリングを行う
- 118に分かれるように分類する

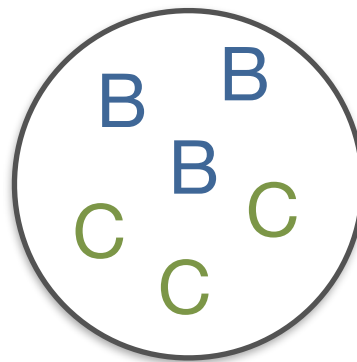


SLC (Single License Cluster)

- 1種類のライセンスから成るクラスター
- 本研究は、SLCを多く作成できることが望ましい
 - SLCからは、容易にライセンスルールを作成できる



SLC

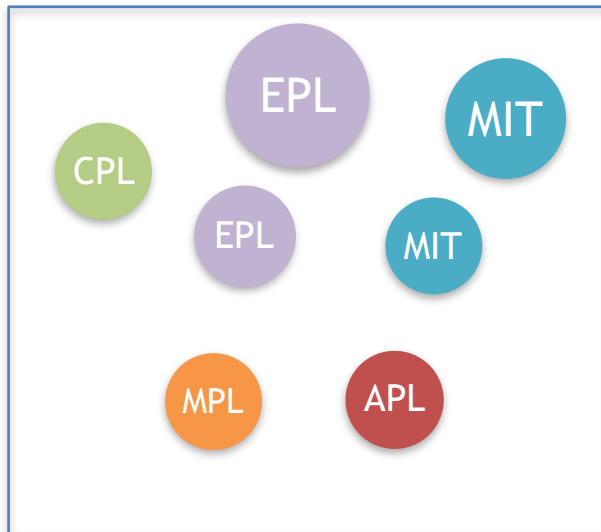


Non-SLC

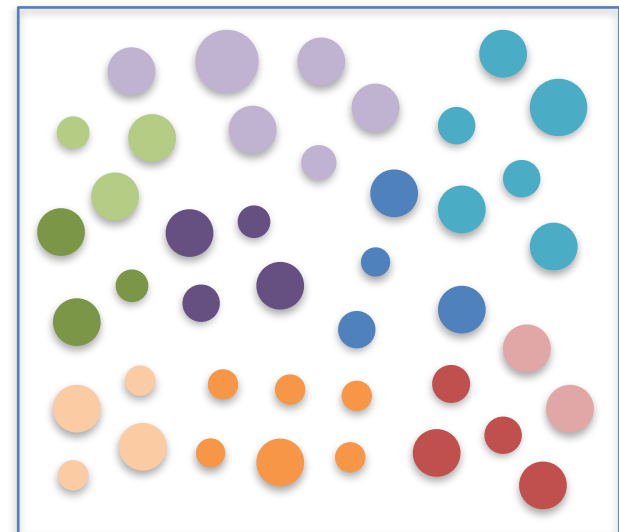
実験結果

	#C	#SLC	SLCのライセンス種類数
cond. α	247	153	35
baseline	81	31	17

baseline



condition α



Circle : SLC **Color** : License **Diameter** : #files in SLC

まとめ & 今後の課題

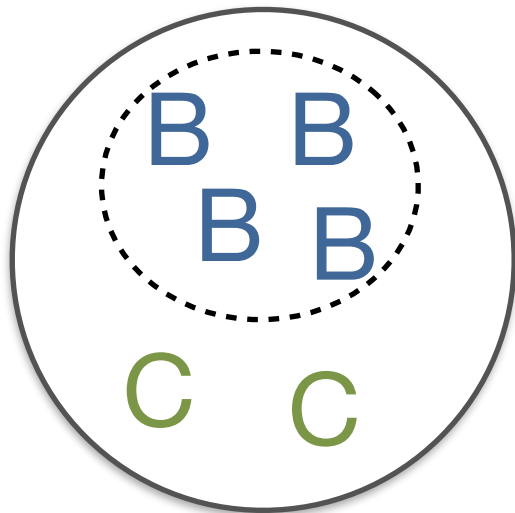
- まとめ
 - 提案手法は、一般的な階層的クラスタリングを行うより、多くのSLCを作成することができた
- 今後の課題
 - 同じライセンスのSLCをマージする
 - Non-SLCからSLCを作成する

End

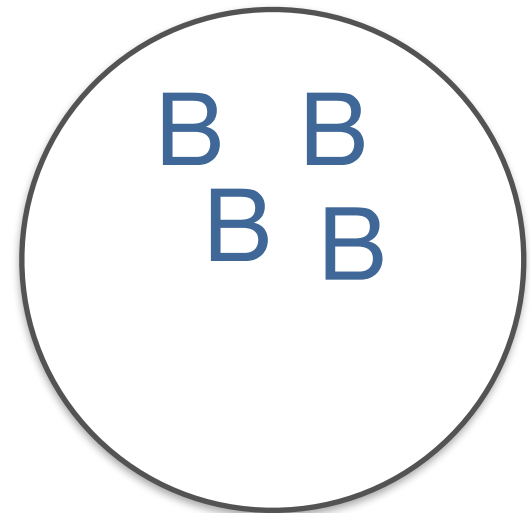
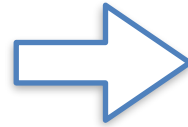


Non-SLCからSLCを作成

- クラスタの中心となる記述を編集距離を用いて抽出
 - クラスタの中心となる記述 : 類似度が93%以上の関係にある記述を最も多く持つ記述



Non-SLC



SLC

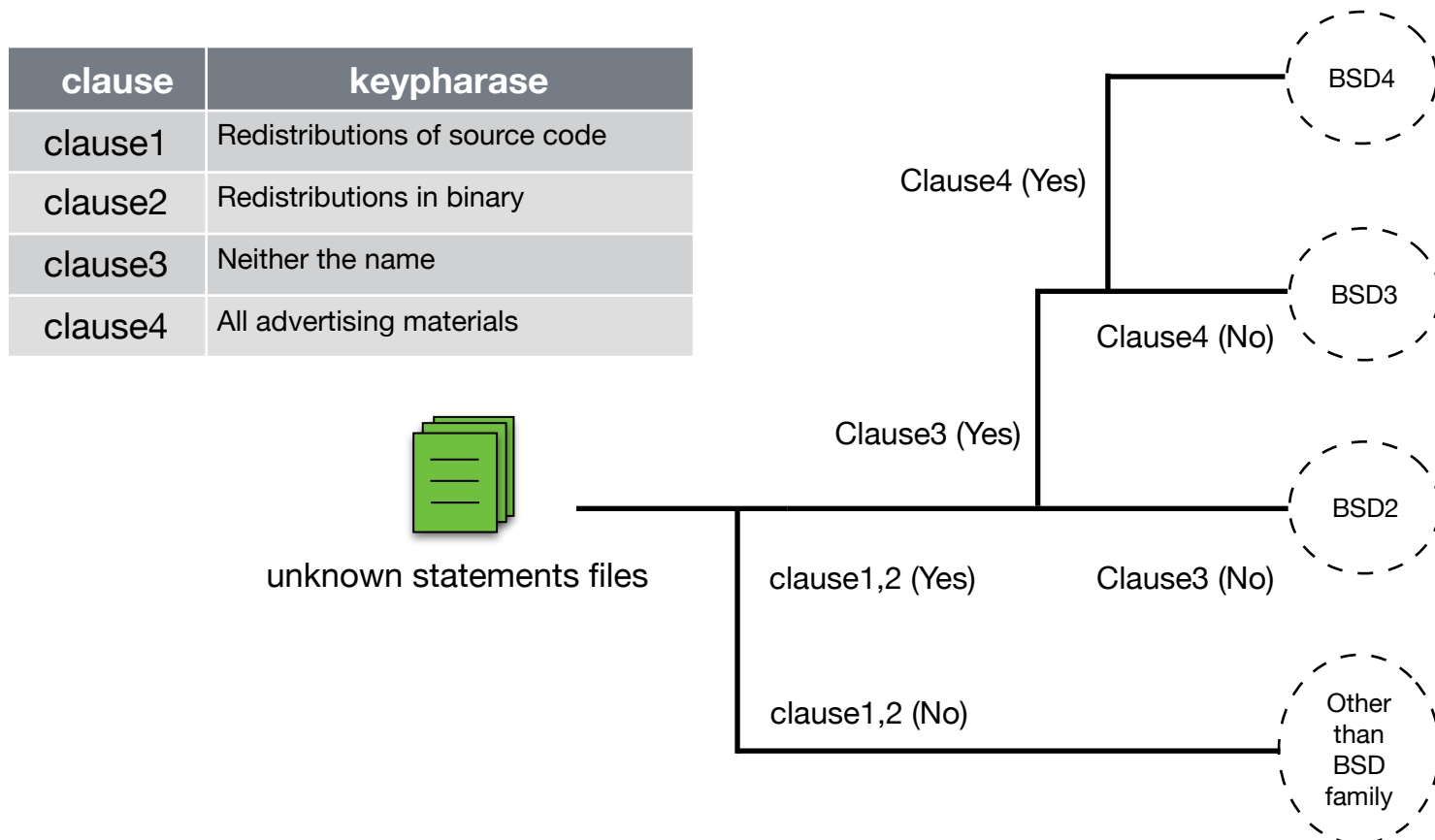
GPL/BSD-Related Lic. Extraction with grep

- We try to correctly discriminate GPL/BSD related licenses with keyphrase
 - GPL Family Licenses

GPL license family	Keyphrase	version name
GPLv1	GNU General Public License	version 1
GPLv2	GNU General Public License	version 2
GPLv3	GNU General Public License	version 3
LibraryGPLv2	GNU Library General Public License	version 2
LGPLv2.1	GNU Lesser General Public License	version 2.1
LGPLv3	GNU Lesser General Public License	version 3
AGPLv3	GNU Affero General Public License	version 3

Extraction of BSD licenses with grep

- Discriminant conditions and flow for BSD license family



Step 1: Extraction of GPL/BSD

- Evaluation
 - We evaluate that grep can correctly find GPL/BSD related licenses from the dataset with precision

- Result

Table : Accuracy of Grep For GPL/BSD Licenses.

Licenses	TP	FP	Precision
GPLv2	170	4	0.98
GPLv3	45	5	0.90
LibraryGPLv2	19	1	0.95
LGPLv2.1	31	2	0.94
LGPLv3	21	3	0.88
AGPLv3	15	0	1.00
BSD2	17	27	0.39
BSD3	16	1	0.94
BSD4	2	0	1.00
all	342	43	0.89