

A sentence-matching method for automatic license identification of source code files

Daniel M. German

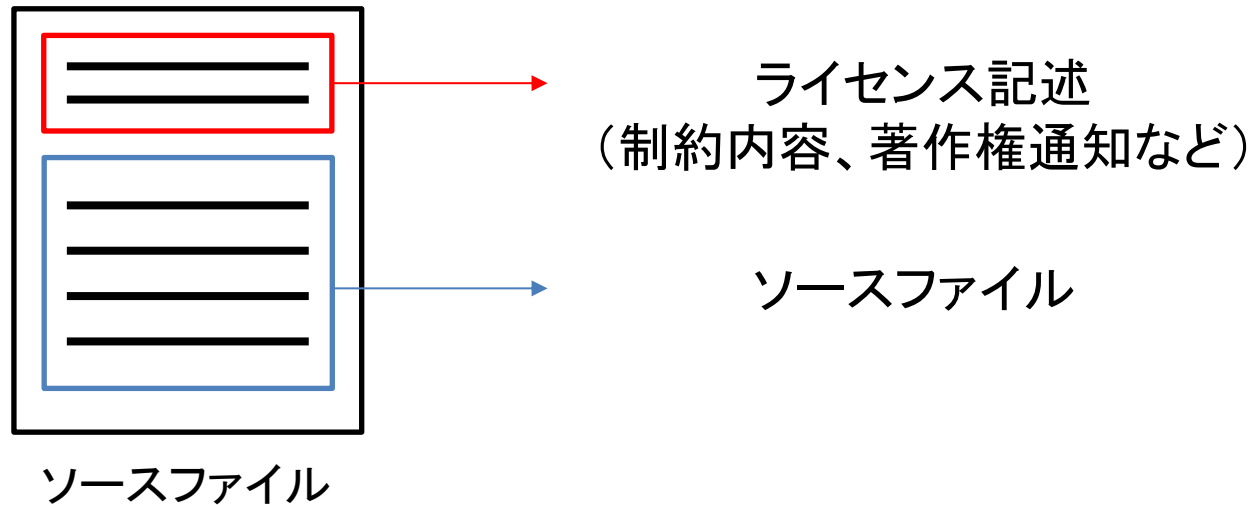
Yuki Manabe

Katsuro Inoue

背景1

- FOSSは再利用できるソースコードとして注目されている
- FOSSは正しく、再利用できるようにライセンス記述として条件をつけている
- ソフトウェアの大多数は、ライセンス記述(ライセンス)が一つだけの下で利用されている

ライセンス記述



ソースファイルの上部にコメントとして記述されており、
ライセンス名、制約内容などが記述してある

ライセンス記述



ライセンス記述の中には、ファイルを参照して
くださいと記述されている場合がある

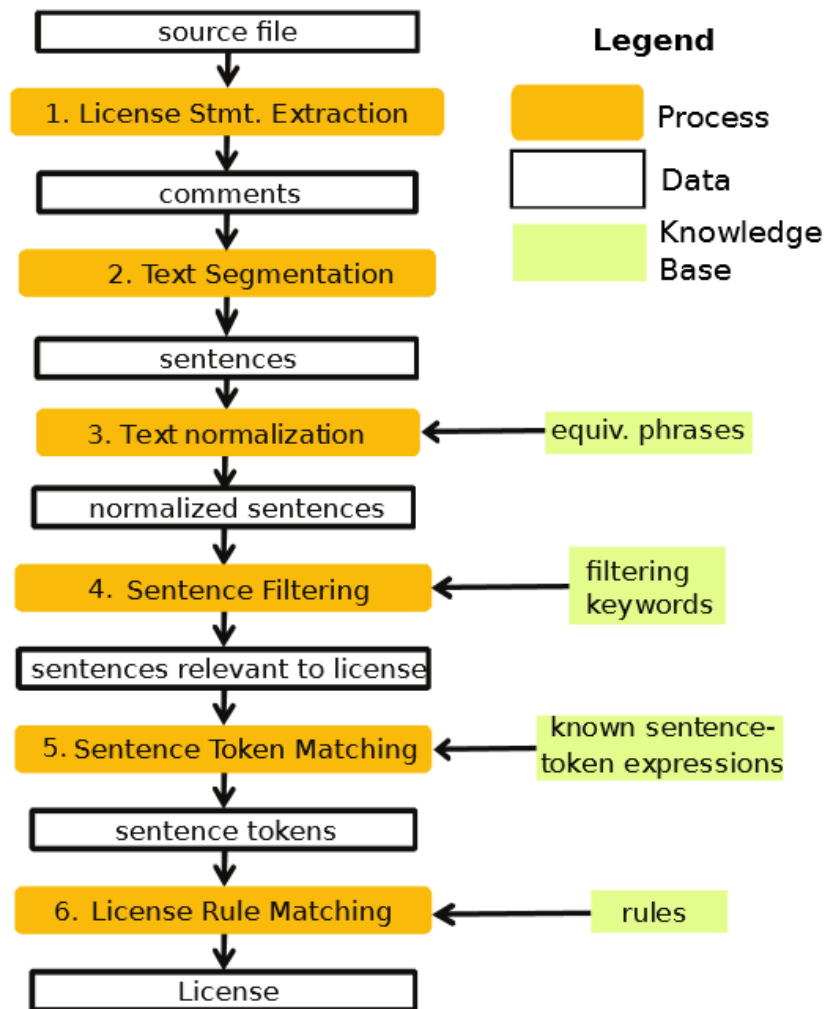
背景2

- これまでの経験的な研究において、会社の37%がOSSの構成要素のソースコードを修正すると報告している
- FOSSを再利用するどのような組織でも、そのような法的意味の変化を理解しなければならないと強調している

目的

- FOSSのコンポーネントから、どのようなライセンスか分かる、ライセンス判定ツールを作成する

ライセンス判定ツール(Ninka)流れ



- 1.ライセンス抽出
- 2.テキスト分割
- 3.テキスト標準化
- 4.意味的フィルター
- 5.意味的
マッチング
- 6.ライセンスルール
マッチング

Figure 1: Diagram depicting the main steps of the algorithm of license identification.

ライセンス抽出

- コメントからライセンス文を抽出する

テキスト分割

- コメントを単語毎に分割する

テキスト標準化

- 読み込んだ単語を統一させる
- 例 you mayという記述とyou canは意味として「してもよい」、「することができる」となっているので、同じとして扱う
- このような同じ意味の単語を12種類用意した

キーワードによるフィルター

- ライセンス記述の文と関連しているキーワード
- キーワードが一つも含まれていない文は判定の対象外とする
- キーワードの例は
 - Permission, license, conditions, disclaimerなど
 - キーワードは82種類用意した

意味的トークンマッチング

- 既知のライセンス文を参照するために、トークンを用いる
- トークンは正規表現を用いて一般化できる
- 427種類のトークンを用いる
- 例
 - SeeFile,^For the licensing terms see the file ([^,;]+)\$
 - SeeFile,^See ([^,;]+) for license\$

ライセンスルールマッチング

- ライセンスを特定するためのトークンの集合（ライセンスルール）と一致しているか確認
- 今回は112種類のライセンスを特定するために、126のライセンスルールを用いる
- 例 BSD2のライセンスルール
 - BSDpre,BSDcondSource,BSDcondBinary,BSDasls,BSDWarrの意味的トークンが存在する
（BSDは著作権表示と免責表示のみなので様々な記述形式が存在する）

BSD2-clause “Simplified” License

- Copyright (c) <year> <owner> All rights reserved. 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.
- THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

BSD 2-clause FreeBSD License

- The FreeBSD Copyright Copyright 1992-2012 The FreeBSD Project. All rights reserved.
- 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.
- THIS SOFTWARE IS PROVIDED BY THE FREEBSD PROJECT ``AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE FREEBSD PROJECT OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
- The views and conclusions contained in the software and documentation are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the FreeBSD Project.

実験

- 他のライセンス特定ツールとの性能を適合率、再現率、F値を用いて比較する
- 比較対象ツール
 - FOSSology version 1.0.0[11]
 - ohcount version 3.0.0rc
 - OSLC 3.0

実験対象と方法

- Debian 5.0.2 (ソースファイル数 80万)
- Debian 5.0.2から250のソースコードを選択し、ライセンス特定する

計算式

$$Recall = \frac{|C|}{|C| + |U|}$$

$$Precision = \frac{|C|}{|C| + |I|}$$

$$F\text{-measure} = \frac{2 * |C|}{2 * |C| + |I| + |U|}$$

- Cは正確に判定できたファイル
- Iが正確に判定できなかったファイル
- Uは判定をしなかったファイル
- (今回のシステムは判定が難しい場合、判定を行わない)

結果

Table 3: Results of the evaluation of the tool. Best result in each category is depicted in bold.

	Ninka	FOSSo.	ohcount	OSLC
$ C $	200	137	83	57
$ I $	7	112	167	193
$ U $	43	1	0	0
<i>Recall</i>	82.3%	99.2%	100.0%	100.0%
<i>Precision</i>	96.6%	55.0%	33.2%	29.5%
F-measure	0.889	0.708	0.498	0.371
Execution Time	22s	923 s	27s	372s

- Ninkaは適合率が最も高く、実行時間が最も速い

Research question

1. FOSSで使われるライセンスは何か？
2. ライセンス記述にはどのような問題が存在するか？

追加調査

- RQ1
 - アプリケーションにNinkaを使用して結果をまとめる
- RQ2
 - Ninkaを使用して、その結果から、分析し、開発者に質問して判断していく

追加調査対象

- Debian5.0.2の11101のアプリケーション
(794622ファイル)

結果

License	Files	Perc.	Aps.
NONE	210147	31.5%	8241
GPLv2+	147535	22.1%	5486
LesserGPLv2.1+	42692	6.4%	767
CDDLv1orGPLv2	37623	5.6%	2
SeeFile	31685	4.7%	1252
Apachev2	25023	3.7%	151
LibraryGPLv2+	23705	3.6%	1150
GPLv2	18930	2.8%	582
GPLv2 or LGPLv2.1 or MPLv1.1	18183	2.7%	30

Table 4: Most common licenses per number of files where they occur in Debian 5.0.2. The last column is the number of applications in which they are found.

RQ1:FOSSで使われるライセンス

- ライセンス記述の存在しないファイルが最も頻繁になっている
- GPLとLGPLは全てのバージョンで顕著にリストに記載されている
- しかし一部に極端に多い場合が存在している
 - 例(2つのアプリケーションだけで多くのファイルが存在している時)
- ライセンス記述がないファイルはサイズが小さい傾向がある

RQ2 ライセンス記述にはどのような問題が存在するか？

- ライセンス記述が存在しない
- 間違ったコピーをする
- ライセンス記述に異なったものが混ざっている
- 間違ったライセンス名になっている
- 著作権記述者しか変更することができない

まとめ

- 新しいライセンス判定ツールを作成した
- 他のライセンス判定ツールよりも適合率が高い結果がでた
- 追加で分析を行い、ライセンス記述が存在しないファイルが多く存在し、問題があることが理解できた