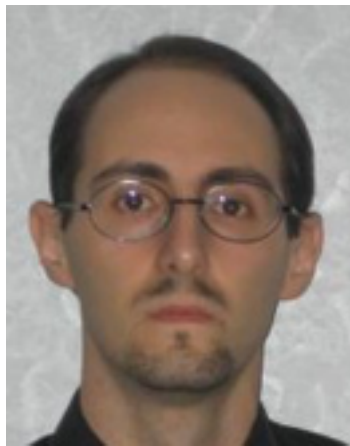


Identifying Licensing of Jar Archives using a Code-Search Approach



Massimiliano Di Penta
University of Sannio, Italy



Daniel M. German
University of Victoria, Canada



Giuliano Antonio
SOCCER Lab. Canada

背景

- ＊ ソフトウェアを再利用することは、ソフトウェア開発において、常識になりつつある
- ＊ Javaにおいては、Jar技術により、リリースを容易にしているため、特に再利用が行なわれている

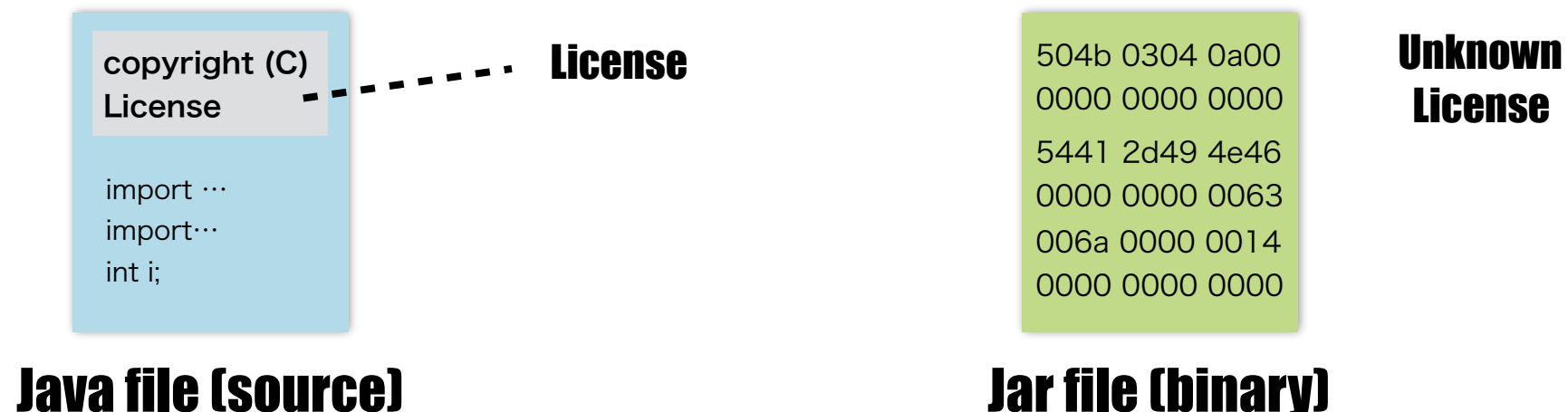


Jar(Java Archives)

- ＊ 複数のJavaファイルをコンパイルし、zipで圧縮したもの
- ＊ メリット：圧縮したまま使用可、ファイルサイズ減少、実行速度
- ＊ ただし、ソフトウェアを再利用するには、ソフトウェアのライセンスを遵守しなければならない

技術的課題

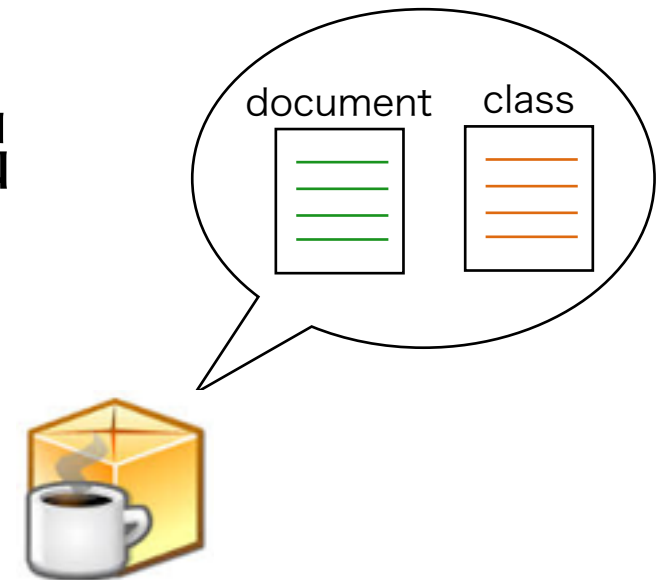
- ❖ 再利用するソフトウェアが、Jarファイルなどのバイナリ形式の場合、ライセンス記述が存在しない
- ❖ 再利用のために必要なライセンス情報が分からない
 - 著作権者情報
 - どのライセンス下でリリースされているか



提案手法(1/3)

1. Jarに圧縮されているファイルを抽出

- ★ ドキュメント (etc.. “.txt”, “.xml”, “.html”)
- ★ クラスファイル



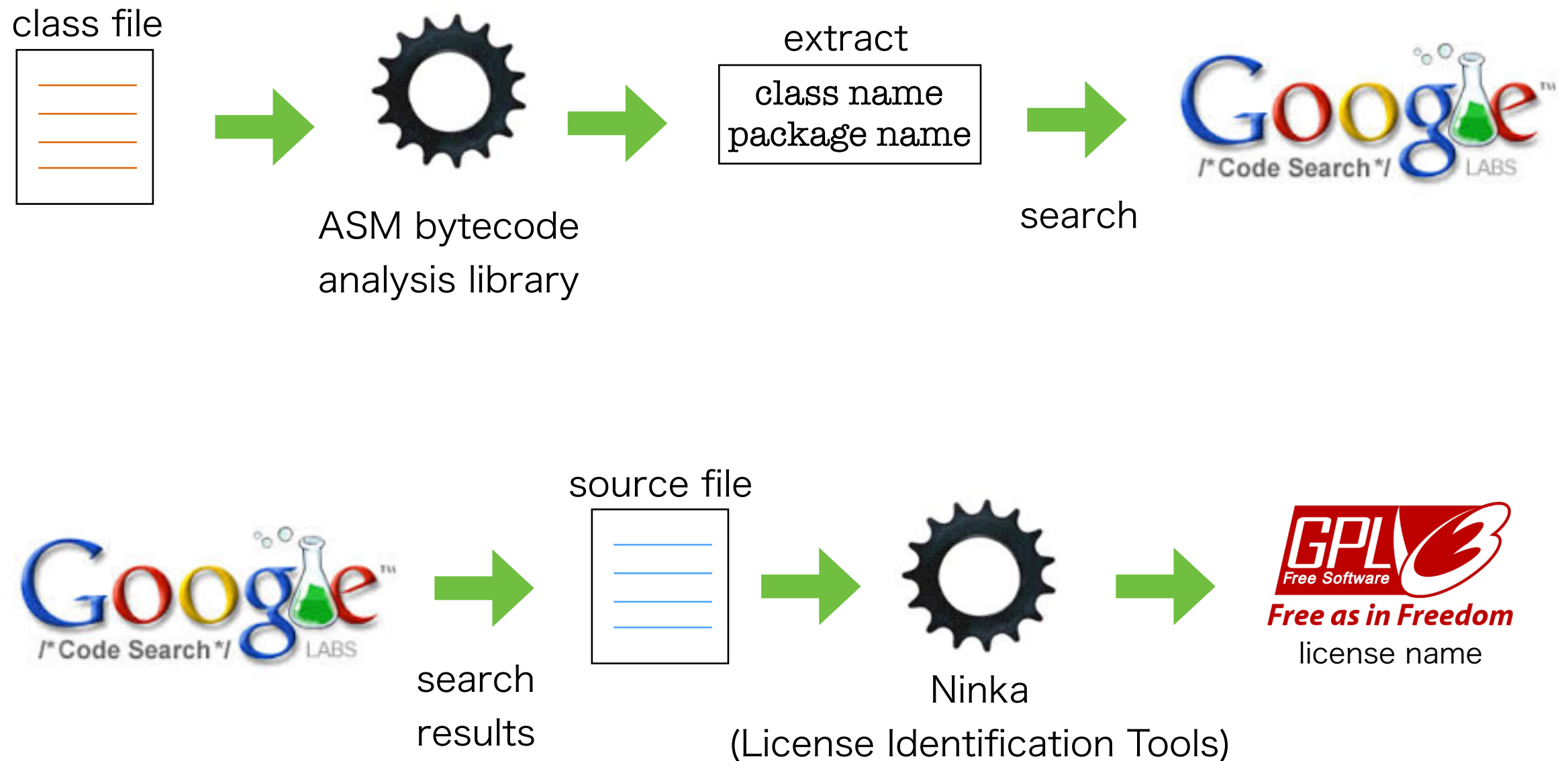
2. ドキュメントのライセンス特定

- ・ ライセンス特定ツール(Fossology)を用いてドキュメントのライセンスを特定



提案手法(2/3)

* 3. クラスファイルのライセンス特定



提案手法(3/3)

✧ 4. クラスファイルのライセンス決定

- マッチしたファイルが1ファイルのみの場合、そのソースコードのライセンスを採用
- 複数のファイルにマッチした場合、多数決により決定



- 検索結果が0件、またライセンス情報がない場合
 - ★ 同パッケージに属する他のクラスファイルで最も多く使われているライセンスを採用

実験対象Jarファイル

Table II
JAR ARCHIVES USED IN THE STUDY.

Application	Description	Jars	Classes
APACHE			
commons-attributes-2.1	Enables C#/.Net-style attributes in Java	3	36
commons-beanutils-1.8.2	Wrapper for Java introspection API	5	137
commons-betwixt-0.6	Maps Java beans into XML	1	135
commons-chain-1.2	Implementation of Chain of Responsibility	3	61
commons-cli-1.0	Parsing of command-line options	3	22
commons-codec-1.1	Encoders and decoders (<i>e.g.</i> , Base64, Hex, Phonetic and URLs)	1	14
commons-logging-1.1.1	Library to log various information	6	83
commons-pool-1.5.4	DB Connection Pooling	3	52
collections-3.2.1	Collection of data structures	4	497
apache-log4j-1.2.15	Manages execution logging	1	259
ECLIPSE BUNDLE			
com.jcraft.jsch_0.1.28	Java Secure CHannel	1	95
com.sun.syndication_0.9.0	Feed management	1	120
javax.servlet.jsp_1.2.0	Servlet/JSP API	1	25
javax.xml.ws_2.0.0	Java Web services library	1	41
org.mozilla.javascript_1.6.2	JavaScript engine (Rhino)	1	199
JLAYER			
JLayer 0.4	Java MP3 library	1	142
JLayer 1.0	—	1	142
TOTAL:		37	2060

Research Question

- ✧ RQ1 : バイナリファイルからソースコードのライセンスをどれくらい特定することができるか
- ✧ RQ2 : クラスファイルのライセンス特定結果は、ドキュメントのライセンスと一致しているか

Google Code Searchでヒットしたクラスファイル数

Table IV
NUMBER AND PERCENTAGES OF FILES FOUND ON GOOGLE CODE SEARCH.

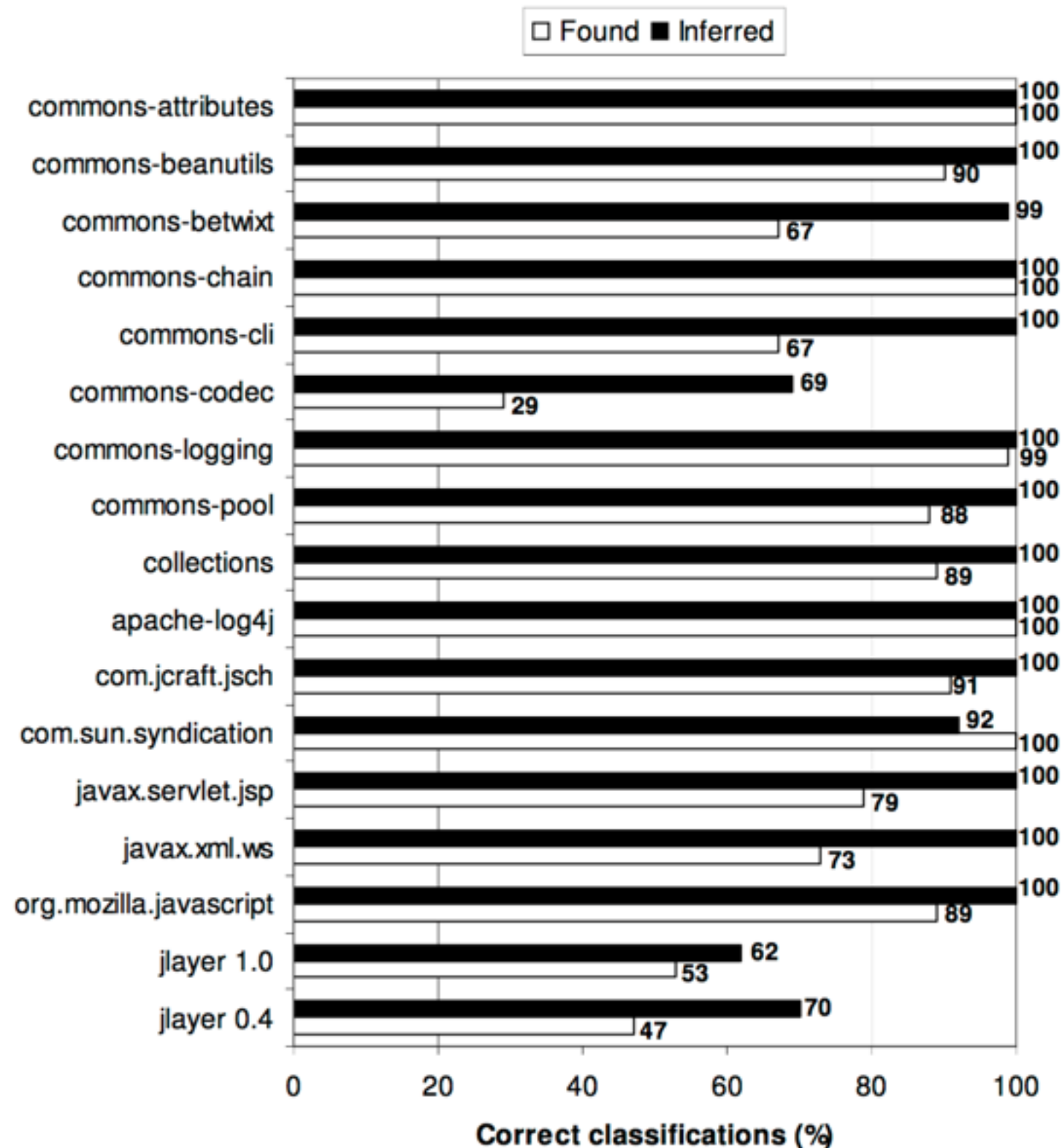
Application	Overall found (%)	Found Licenses (%)						
		AGPL	GPL	LGPL	BSD	Apache	MPL	MIT
commons-attributes	5 (14%)	—	—	—	—	5 (100%)	—	—
commons-beanutils	62 (45%)	—	—	—	6 (10%)	56 (90%)	—	—
commons-betwixt	9 (7%)	—	—	—	3 (33%)	6 (67%)	—	—
commons-chain	18 (30%)	—	—	—	—	18 (100%)	—	—
commons-cli	6 (27%)	—	—	—	2 (33%)	4 (67%)	—	—
commons-codec	8 (57%)	—	—	—	6 (75%)	2 (25%)	—	—
commons-logging	71 (62%)	—	—	—	1 (1%)	70 (99%)	—	—
commons-pool	26 (50%)	—	—	—	3 (12%)	23 (88%)	—	—
collections	81 (16%)	—	—	—	9 (11%)	72 (89%)	—	—
apache-log4j	102 (39%)	—	—	—	—	102 (100%)	—	—
com.jcraft.jsch	22 (23%)	—	—	2 (9%)	20 (91%)	—	—	—
com.sun.syndication	28 (23%)	—	—	—	—	28 (100%)	—	—
javax.servlet.jsp	14 (56%)	—	1 (7%)	2 (14%)	—	11 (79%)	—	—
javax.xml.ws	22 (54%)	—	16 (73%)	2 (9%)	—	4 (18%)	—	—
org.mozilla.javascript	51 (26%)	1 (2%)	—	—	—	1 (2%)	47 (92%)	2 (4%)
JLayer 0.4	86 (61%)	—	40 (47%)	46 (53%)	—	—	—	—
JLayer 1.0	86 (61%)	—	40 (47%)	46 (53%)	—	—	—	—

多数決により決定されたクラスファイル数

Table V
NUMBER AND PERCENTAGES OF LICENSES INFERRED USING GOOGLE CODE SEARCH.

Application	Overall inferred (%)	Inferred Licenses (%)						
		AGPL	GPL	LGPL	BSD	Apache	MPL	MIT
commons-attributes	36 (100%)	—	—	—	—	36 (100%)	—	—
commons-beanutils	268 (98%)	—	—	—	6 (2%)	262 (98%)	—	—
commons-betwixt	135 (100%)	—	—	—	5 (4%)	130 (96%)	—	—
commons-chain	61 (100%)	—	—	—	—	61 (100%)	—	—
commons-cli	22 (100%)	—	—	—	2 (9%)	20 (91%)	—	—
commons-codec	14 (100%)	—	—	—	6 (43%)	8 (57%)	—	—
commons-logging	115 (100%)	—	—	—	1 (1%)	114 (99%)	—	—
commons-pool	52 (100%)	—	—	—	3 (6%)	49 (94%)	—	—
collections	495 (100%)	—	—	—	10 (2%)	485 (98%)	—	—
apache-log4j	236 (91%)	—	—	—	—	236 (100%)	—	—
com.jcraft.jsch	95 (100%)	—	—	2 (2%)	93 (98%)	—	—	—
com.sun.syndication	120 (100%)	—	—	—	—	120 (100%)	—	—
javax.servlet.jsp	25 (100%)	—	1 (4%)	2 (8%)	—	22 (88%)	—	—
javax.xml.ws	41 (100%)	—	35 (85%)	2 (5%)	—	4 (10%)	—	—
org.mozilla.javascript	198 (99%)	1 (0.5%)	—	—	—	1 (0.5%)	194 (98%)	2 (1%)
JLayer 0.4	142 (100%)	—	84 (59%)	58 (41%)	—	—	—	—
JLayer 1.0	142 (100%)	—	84 (59%)	58 (41%)	—	—	—	—

クラスファイルのライセンス特定精度 (RQ1)



多くのファイルセットにおいて
Inferred > Found

完全一致したファイルのライセンスより、
同じパッケージで最も多く使用されている
ライセンスを用いるほうが精度が高い

クラスファイルのライセンス特定結果は、 ドキュメントのライセンス特定結果と一致しているか (RQ2)

Table VII
SUMMARY OF LICENSES DECLARED IN TEXTUAL FILES, INFERRED, AND EXTRACTED FROM THE SOURCE CODE.

Application	Declared	Inferred	Source
commons-attributes	–	Apache (100%)	Apache v2 (100%)
commons-beanutils	Apache v2.0	(BSD 2%), Apache (98%)	Apache v2 (100%)
commons-betwixt	Apache v2.0	BSD (4%), Apache (96%)	Apache v2 (100%)
commons-chain	Apache v2.0	Apache (100%)	Apache v2 (100%)
commons-cli	Apache v2.0	BSD (9%), Apache (91%)	Apache v2 (100%)
commons-codec	'Apache v1.1'-style	BSD (43%), Apache (57%)	Apache v1.1 (100%)
commons-logging	Apache v2.0	BSD (1%), Apache (99%)	Apache v2 (100%)
commons-pool	Apache v2.0	BSD (6%), Apache (94%)	Apache v2 (100%)
collections	Apache v2.0	BSD (2%), Apache (98%)	Apache v2 (100%)
apache-log4j	–	Apache (100%)	Apache v2 (100%)
com.jcraft.jsch	EPL v1.0, BSD-3	LGPL (2%), BSD (98%)	BSD 3 (100%)
com.sun.syndication	EPL v1.0, Apache 2.0	Apache (100%)	Apache v2 (100%)
javax.servlet.jsp	EPL 1.0, Apache v1.1	GPL (4%), LGPL (8%), Apache (88%)	Apache v2 (100%)
javax.xml.ws	CDDL v1.0	GPL (85%), LGPL (5%), Apache (10%)	GPL v2 ClasspathException (100%)
org.mozilla.javascript	NPLv1.1	AGPL (0.5%), Apache (0.5%), MPL (98%), MIT (1%)	MPL (99.5%) MIT-old (0.5%)
JLayer 0.4	–	GPL (59%), LGPL (41%)	GPL v2 (100%)
JLayer 1.0	–	GPL (59%), LGPL (41%)	LGPL v2 (100%)

- Apacheは、ほとんど一致している
- Eclipseでは、クラスファイルには存在しないライセンスが検出されている
 - ドキュメントはEPLである確率が高い
 - ドキュメントのライセンスを鵜呑みにしてはいけない

まとめ

- ✧ バイナリ形式のファイルのライセンスを特定する手法を提案した
 - 17プロジェクト、37ファイルのjarに提案手法を適用
 - 結果：平均して、95%のクラスファイルのライセンスを正しく特定することができた
- ✧ 今後の課題
 - 同じバージョンのクラスファイルを検索できるようにして、精度を上げる
 - もっと多くのjarファイルで実験して信頼性を上げる

所感

* Lessons Learn Sectionが特殊

- 今回の実験から学ぶべきことがまとめられている
- 1) パッケージのドキュメントのライセンスを信じるべからず
 - * ドキュメントとそのファイルのライセンスは異なる場合がある
- 2) Google Code Search Engineに登録されているライセンス情報を信じるべからず
 - * ライセンス情報が間違っている、また、バージョンが登録されていないことが多い
- 3) ライセンス記述挿入ミスにより、ライセンスが複数検出される場合がある
 - * Eclipseの3rd-party製 jarファイルで、EPLが検出された原因として、全てのファイルに対し、ライセンス記述を挿入していることが考えられる