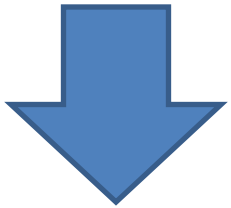
検索エンジンを利用した Covert Channelの検出

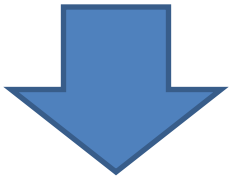
木下研究室
200602824　久保直也

研究の背景・目的

近年、ネットワークの巨大化によりアクセス権限も複雑に絡み合っている。



ネットワーク内では不正な情報経路が発生し、情報流出の危険性が増大してしまっている。



このような情報流出経路の解析法としてCovert Channel解析があるが従来のように把握したコミュニティーのACL（Access Control List）のみを用いたCovert Channelの解析だけでは検出できないアクセス権の矛盾が存在する場合がある。

研究の背景・目的

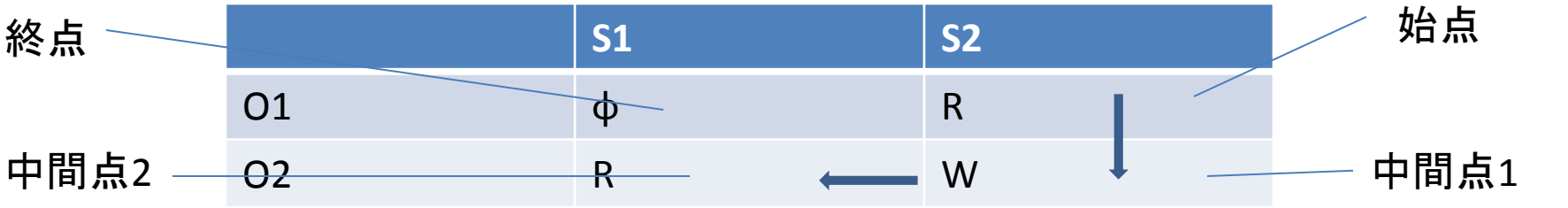
検索エンジンで得られた情報にオントロジーを用いたセマンティックな解析手法を適用することでACLの矛盾や経路を効率よく見つけることを目的とする。

検出できない可能性のあるアクセス権の矛盾を検出するため

オントロジーDBを作成し検索エンジンで収集した情報を形態素解析・構文解析しRDF化し意味まで考慮したマッチングを行うことで外的要因まで考慮したACLの矛盾や経路を見つける方法を提案する。

Covert Channel

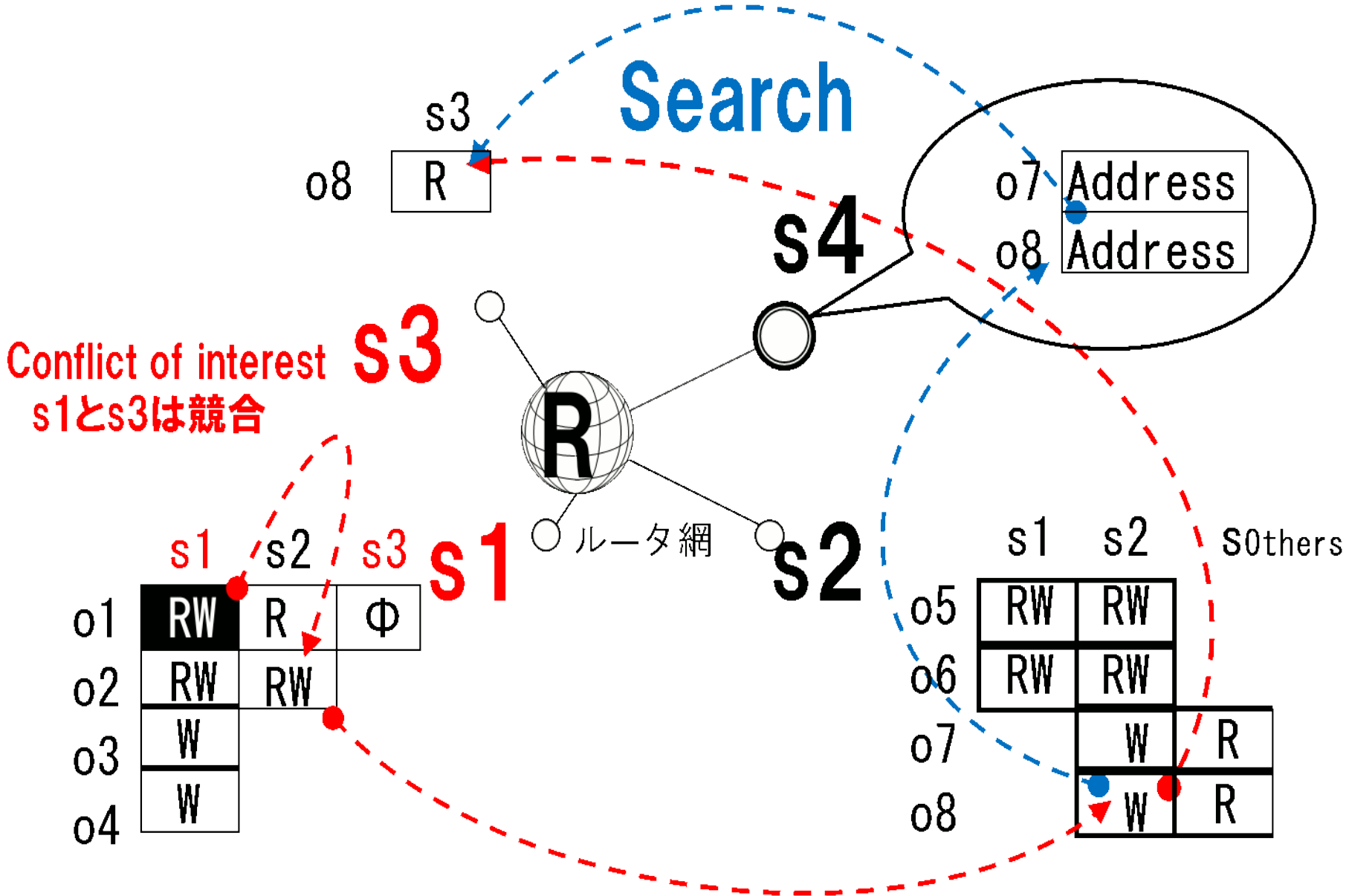
Covert Channelは意図しない情報経路のことで、
アクセス行列において、Subject、Object、permissionを
アクセストリプルと定義した時始点から
終点への流れで、情報流出が発生してしまうことを言う。



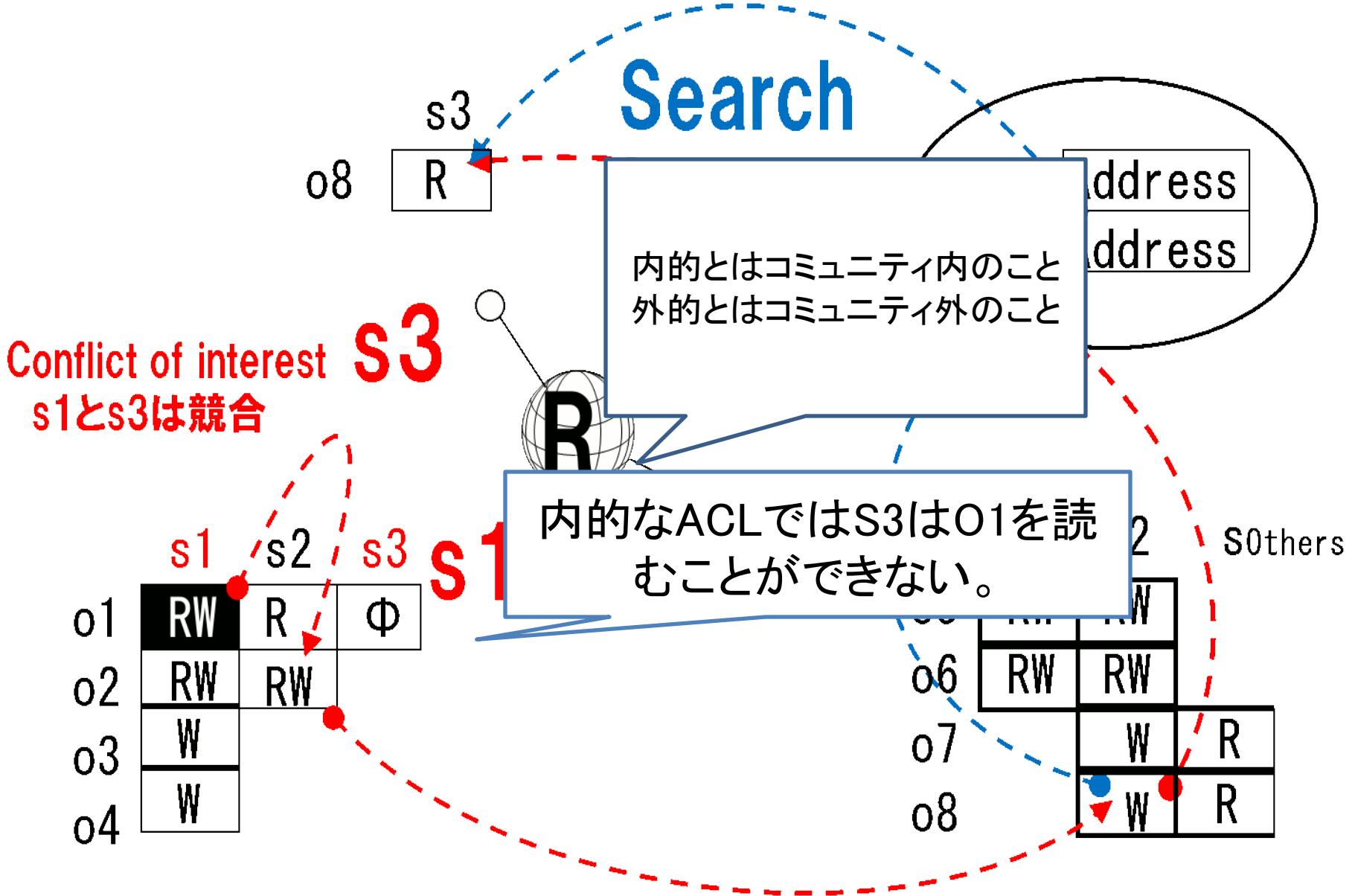
アクセス行列

S:Subject
O:Object
R:READ
W:WRITE

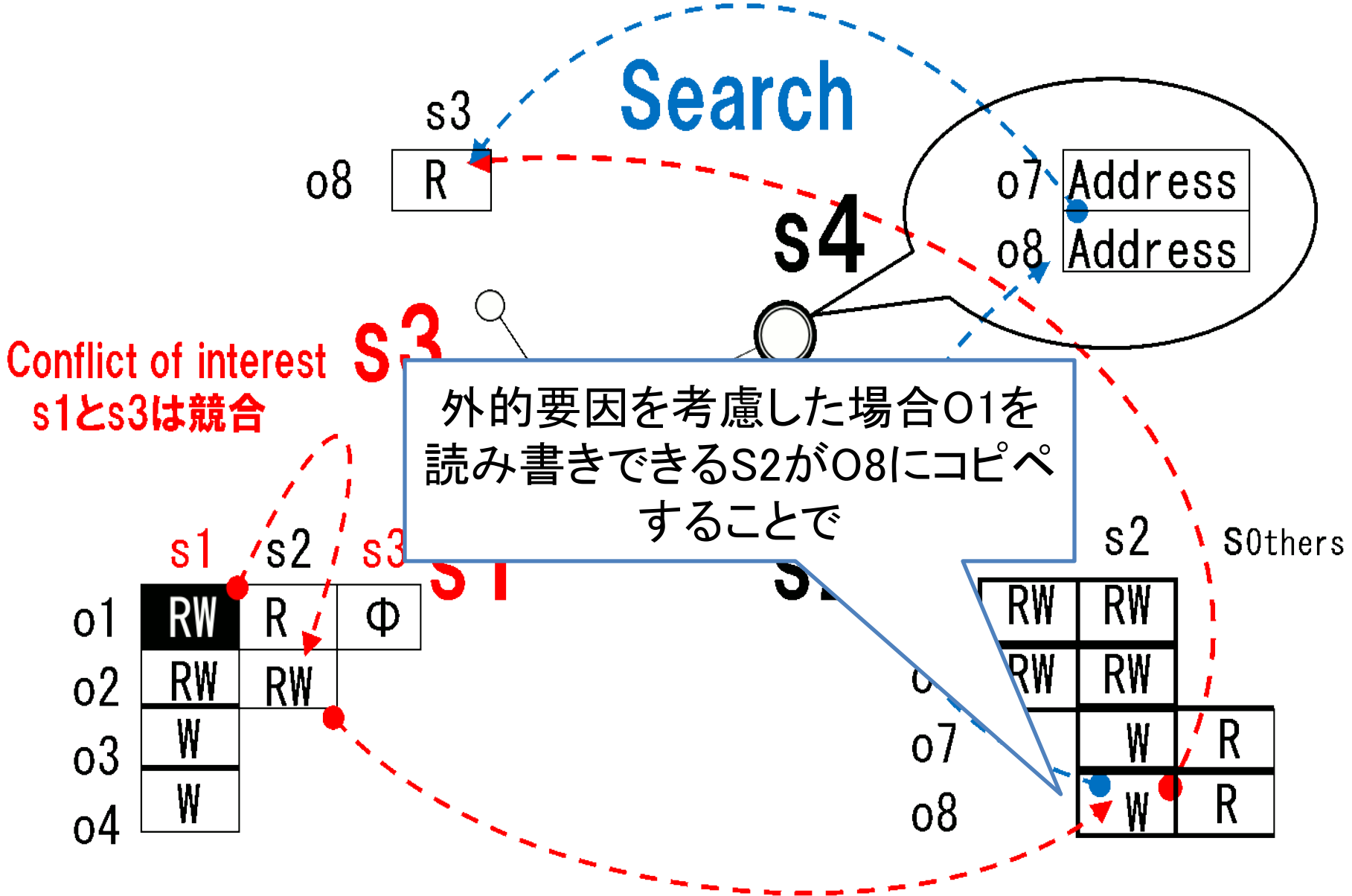
検索エンジンとCovert Channel



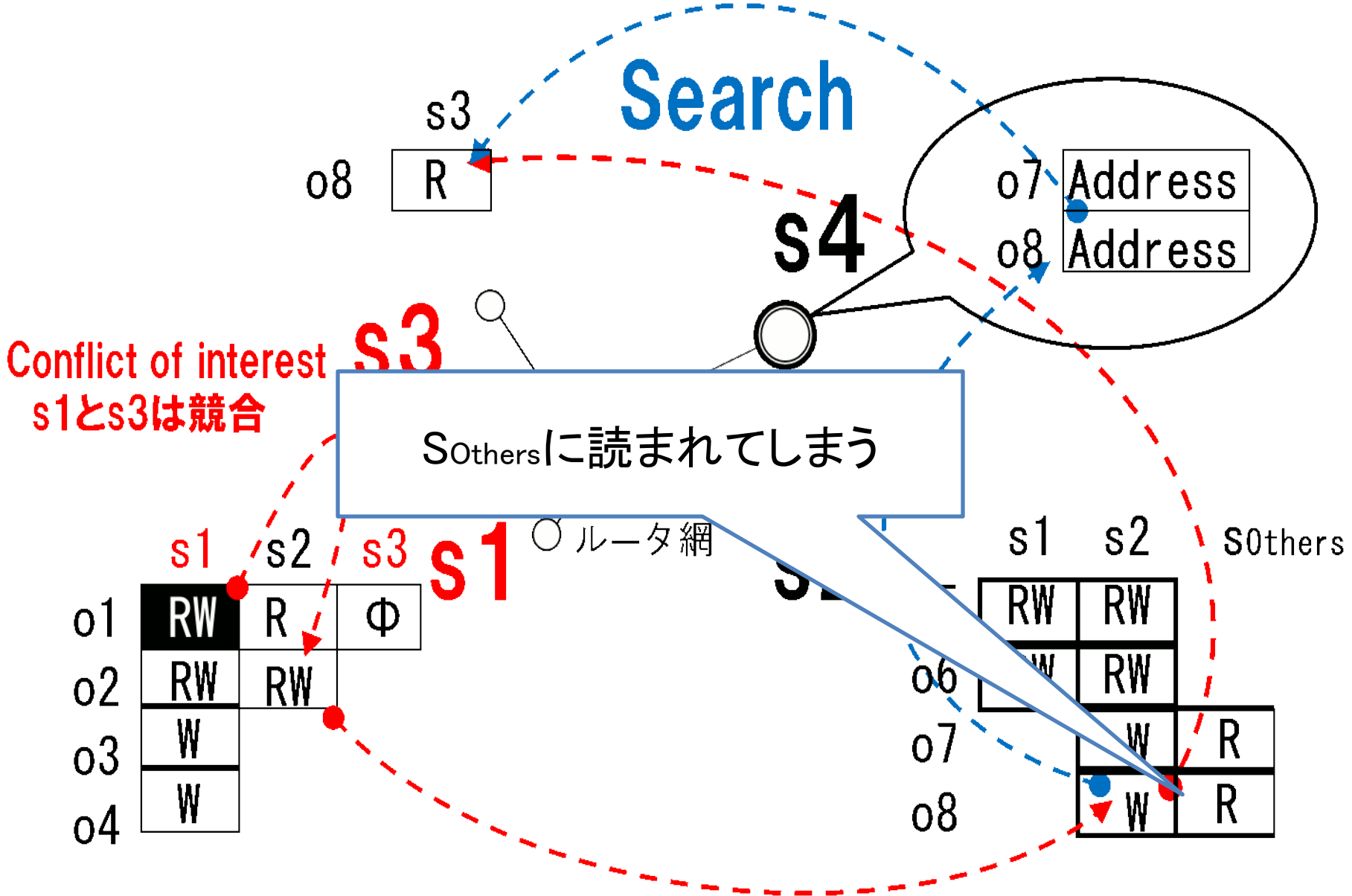
検索エンジンとCovert Channel



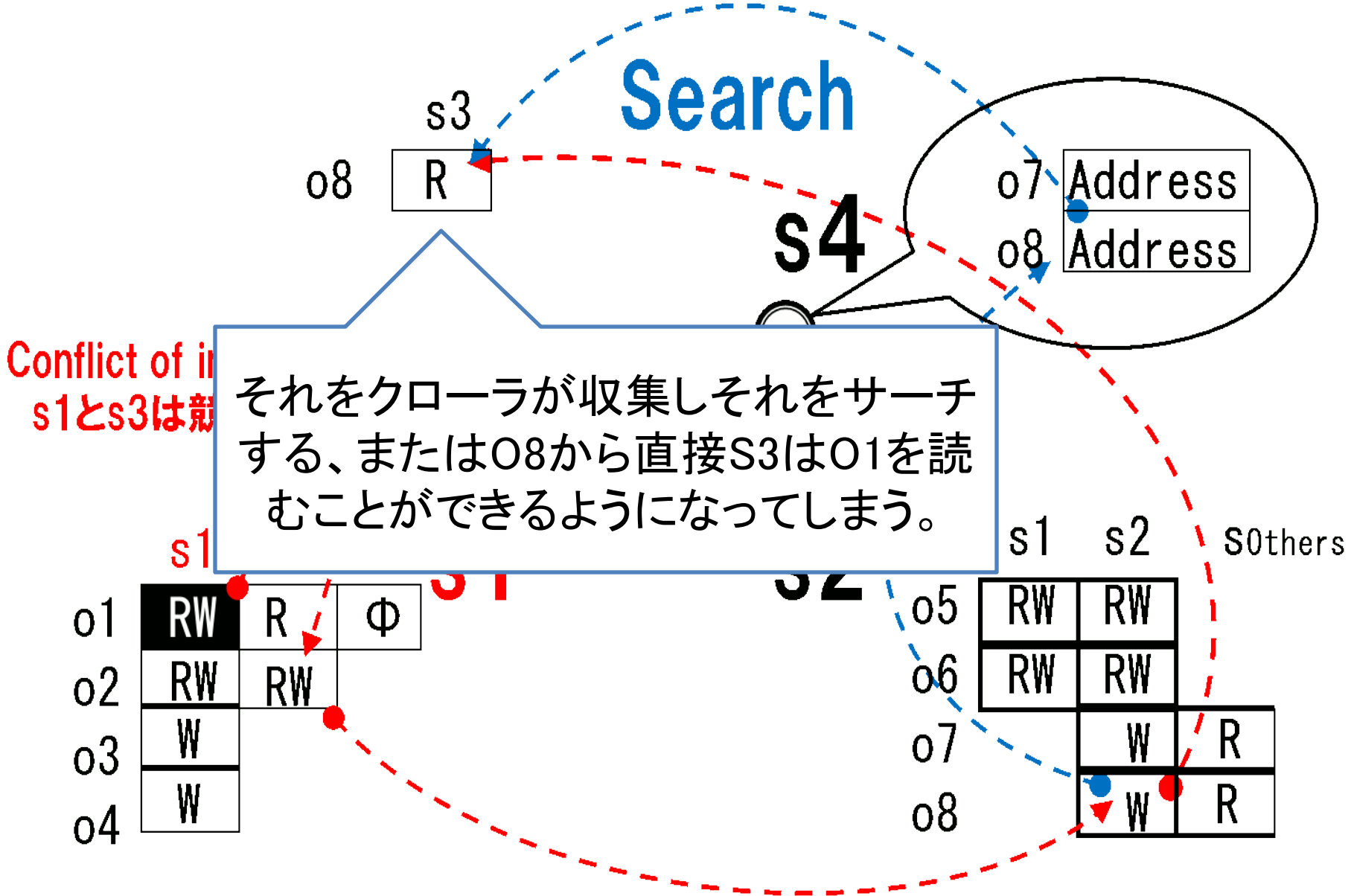
検索エンジンとCovert Channel



検索エンジンとCovert Channel

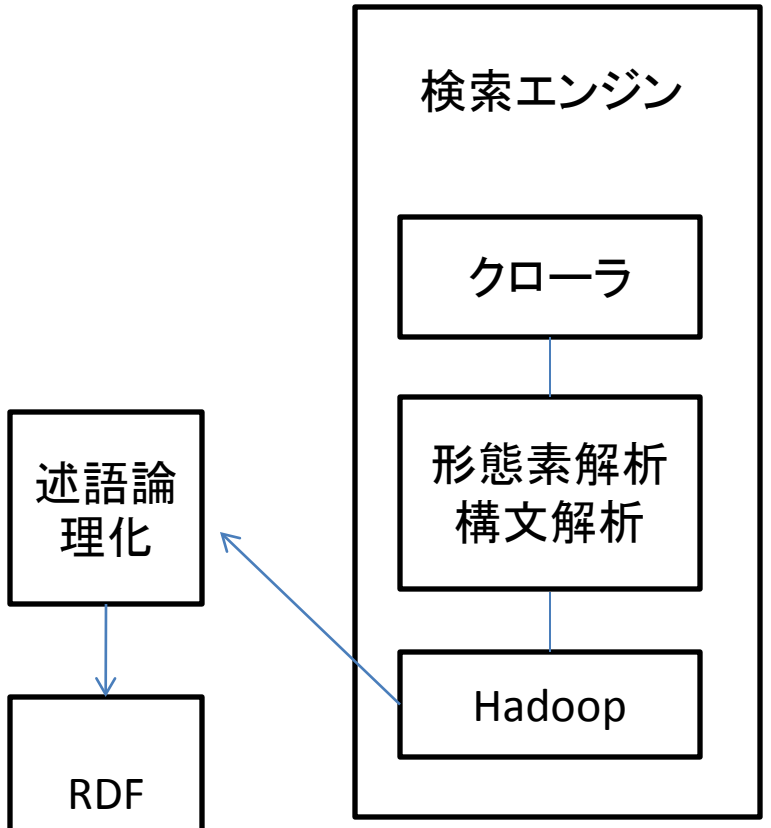


検索エンジンとCovert Channel



研究の概略

全体図



	S1
コンテンツ	R
電番	R

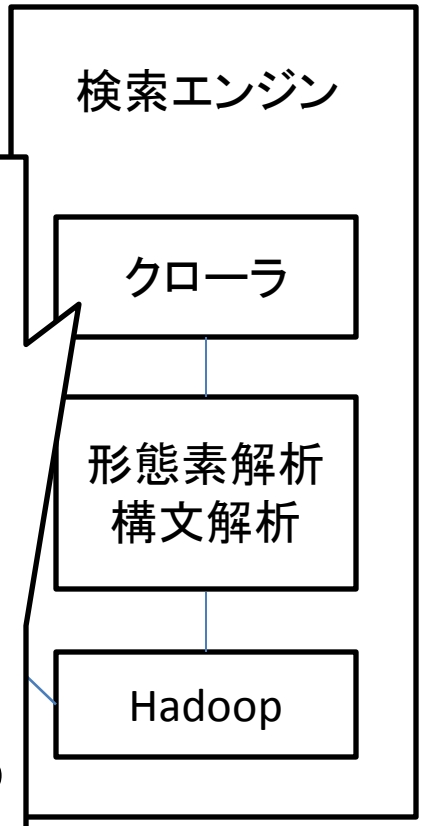
ACL

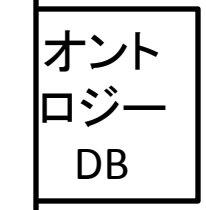
	S1
電番	×R

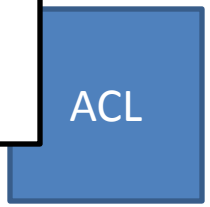
ACL

研究の概略

- Web上を自動的に巡回してWebページを収集する検索ロボットプログラムのこと。
- 既知のHTML文書の新しいコピーを要求
- 文書中に含まれるリンクをたどる
- 別の文書を収集する。という動作を繰り返す。
- 新しい文書を見つけた場合はデータベースに登録する。





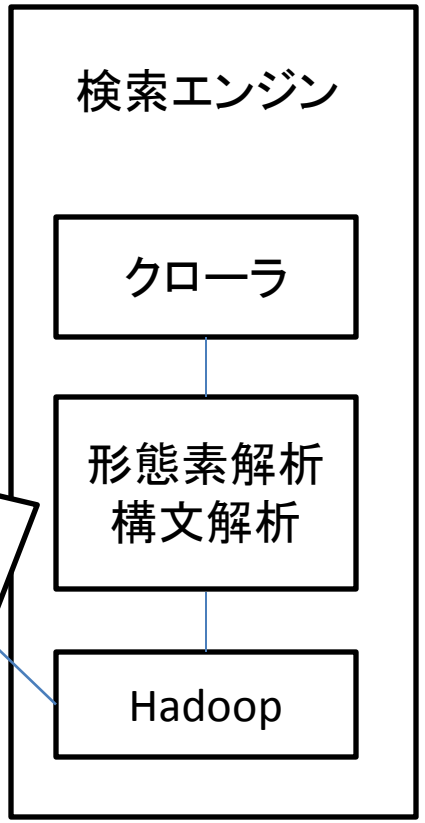


ACL

ACL

研究の概略

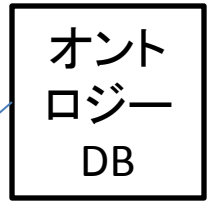
全体図



・形態素解析では単語分割、品詞タグ付けをする。単語分割とは、文中の単語を同定する作業である。例えば

子供｜の｜体力｜低下

と単語分割される。品詞タグ付けとは各単語の品詞を同定する作業である。



	S1
電番	×R

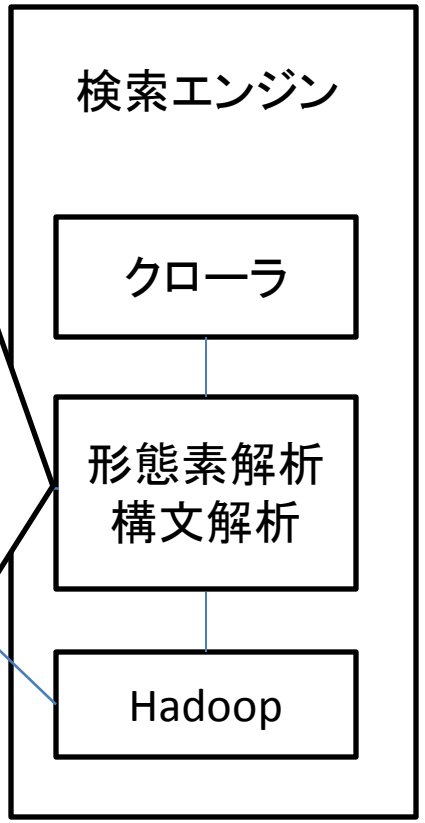
ACL

研究の概略

- 構文解析では主に係り受けを解析する。先程の例を係り受けでは

子供→体力　体力→低下

と表される。
形態素解析、構文解析両方から検索することによって検索の精度が上がる。



オントロジーDB

コンテンツ	R
電番	R

ACL

ACL

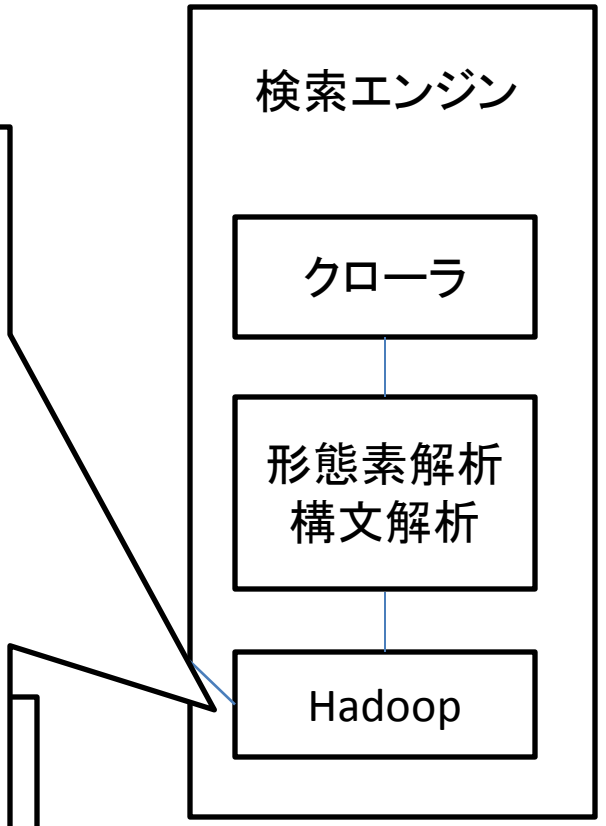
ACL

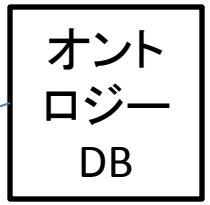
	S1
電番	×R

ACL

研究の概略

- Hadoopは，Google検索システムにおいて大量の「メタ言語のインデックス」を整理分類する。
- インターネット内に散らばったリソースのファイル名，ファイル内容の語を収集分析する。
- インデックスとしてまとめる機能Mapフェーズと Reduceフェーズの2つから成り，計算処理を分散して行う。





	S1
電番	×R

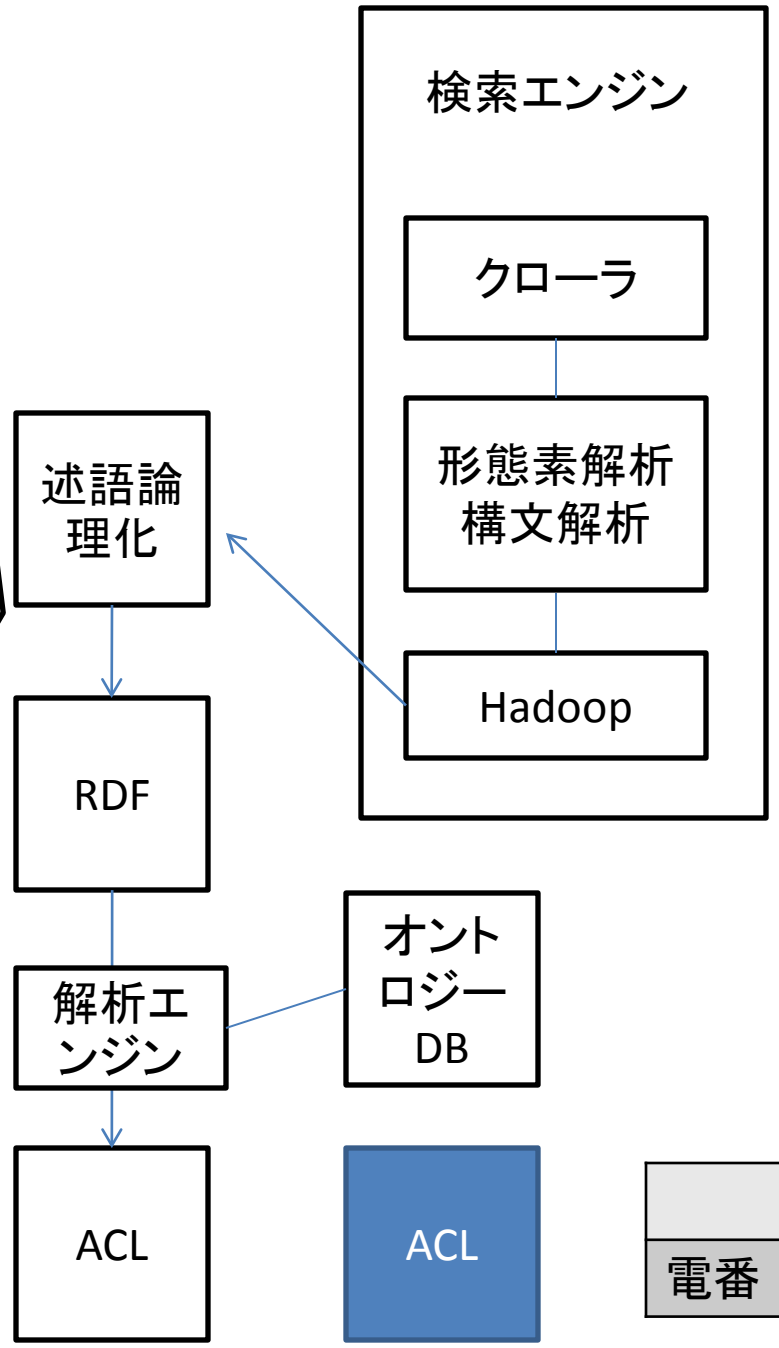
ACL

研究の概略

主語Xを変えるとそれに応じた命題P(X)が得られるのだが、変数Xを具体的に指定することなしに

P(X)

という命題を考えるのが述語論理である
ここでは意味まで考慮したマッチングを行うために述語論理化します



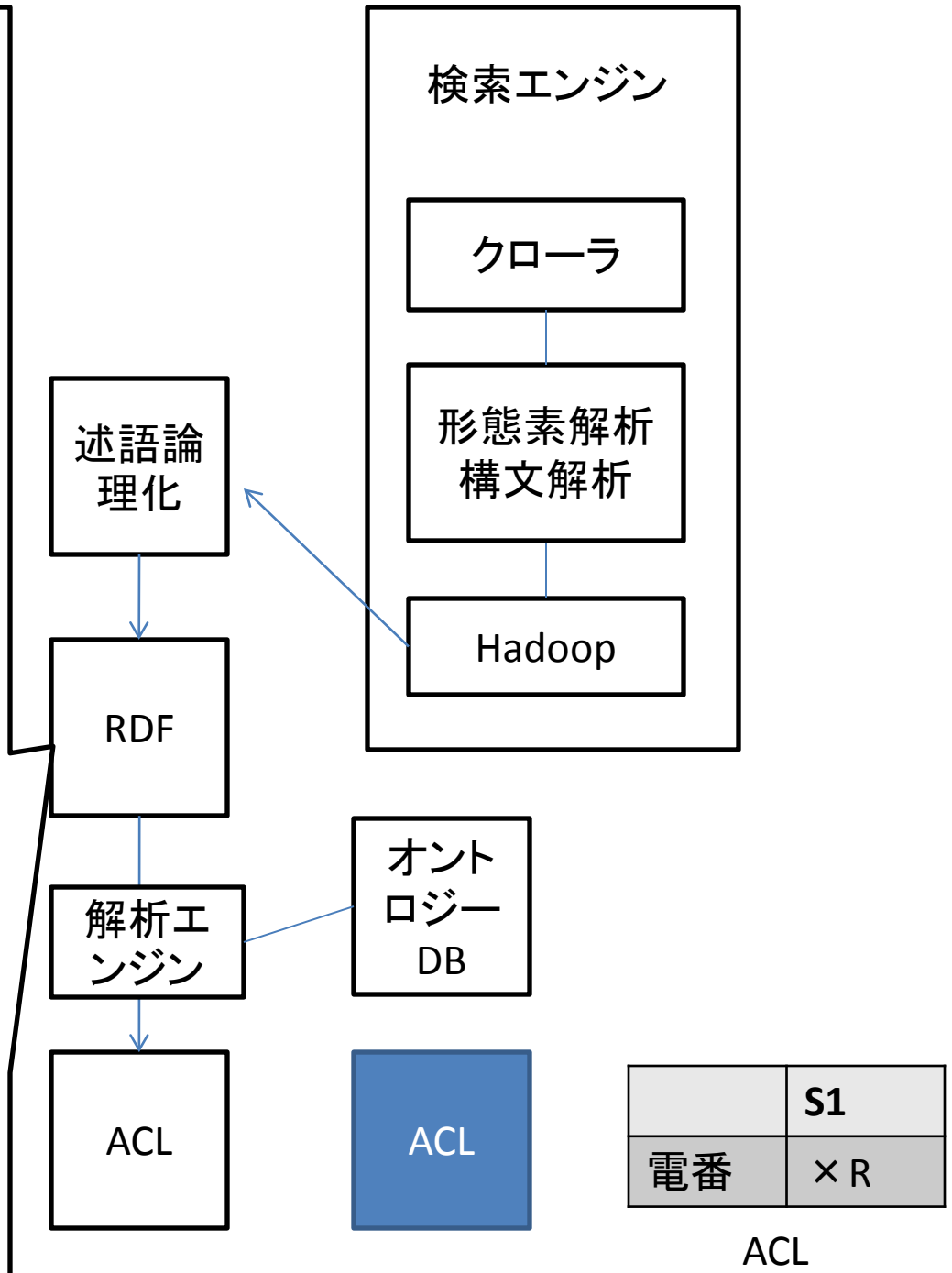
	S1
電番	×R

ACL

・RDF は，リソースを表記する枠組みで，SPO のトリプルでリソースとオブジェクト間の関係を厳密に記述することができる。

例http://www.kanzaki.com の作者は神崎正英です。といった文があった場合。

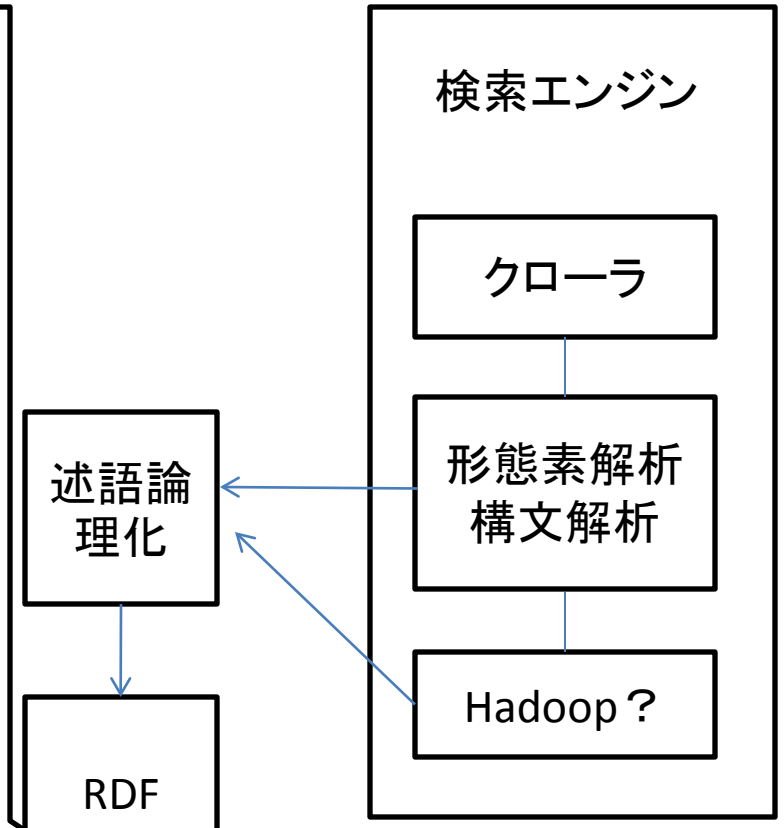
主語（Subject）	リソース	http://www.kanzaki.com
述語（Predicate）	プロパティ	作者
目的語（Object）	プロパティの値	神崎正英



	S1
電番	×R

ACL

・Ontology にはさまざまな定義があるが，工学的には「概念化の明示的な記述」とされている.
・その目的は，自然言語によって記述されたメタデータに存在する曖昧さを排除することである.
・それによりコンピュータが知識を意味論的に扱うことを可能とすることである.

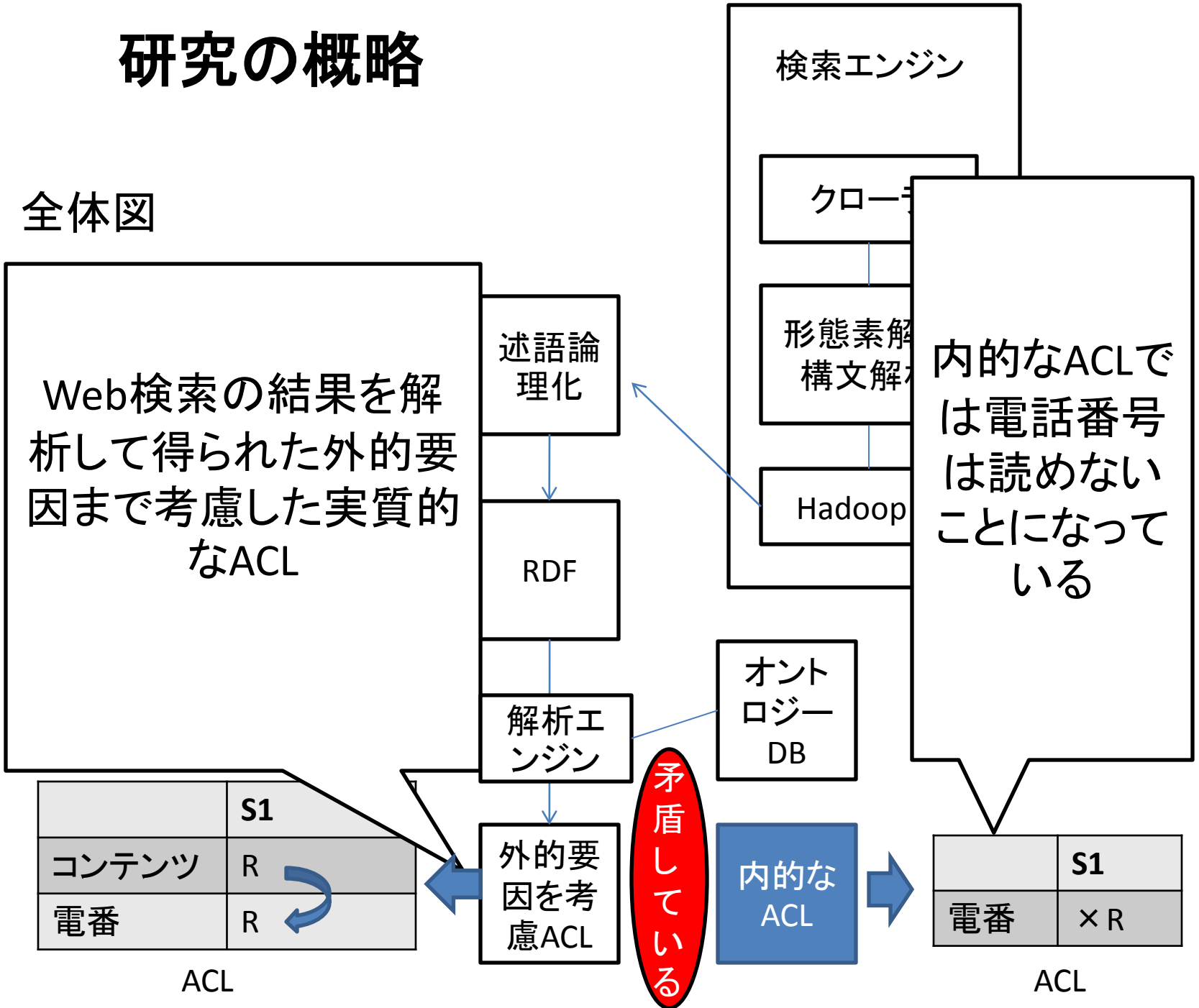


	S1
電番	×R

ACL

研究の概略

全体図


検索エンジン
クローラ
形態素解析
構文解析
Hadoop
述語論理化
RDF
解析エンジン
オントロジーDB
Web検索の結果を解析して得られた外的要因まで考慮した実質的なACL
内的なACLでは電話番号は読めないことになっている
外的要因を考慮ACL
矛盾している
内的なACL
S1
コンテンツ R
電番 R
ACL
S1
電番 ×R
ACL

Covert Channelの検出手順

1 ・クローラで収集された情報のタグを取り除き形態素解析、構文解析を行う。
形態素解析，構文解析を行うことで，検索の精度を上げることができる。

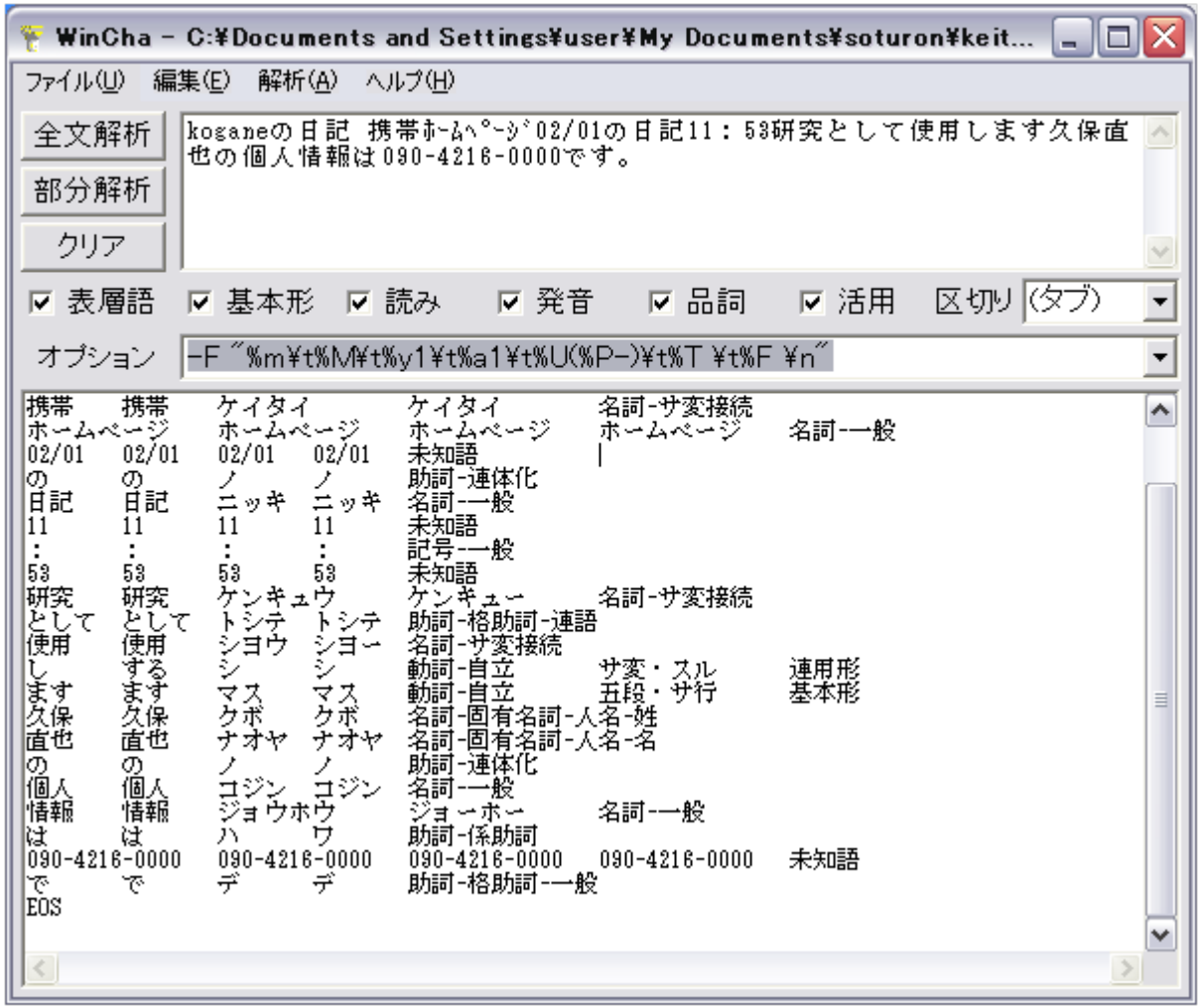
・ここではHTMLファイルからタグ部分を取り除き、検索対象となる文章を取りだし形態素解析を行う。

<html> <head> <meta http-equiv="Cache-Control" content="no-cache"> <meta http-equiv="Content-Type" content="text/html; charset=Shift_JIS"> <meta Name="keyword" CONTENT=""> <title>koganeの日記 携帯ホームﾍﾟｰｼﾞ ﾌｫﾚｽﾄ</title> </head> <body bgcolor="#ffffff" text="#000000" link="#ff3333" vlink="#0000ff" alink="#00ff00"><tt>
 02/01の日記
  11：53 研究として使用します
――――――――
 久保直也の個人情報は090-4216-0000です。

 口 日記を書き直す
 口 この日記を削除
 <hr size="1" color="#000000">[戻る]
 <hr size="1" color="#000000"> <img src="/img/ad/papyless/pc/9-3105-200-46806001.gif"border=0>

 声優＋ｱﾆｿﾝｼﾝｶﾞｰ
両方学べる学校に行く
  <hr size="1" color="#000000"> (C)フォレストページ
  </tt> </body> </html>

・先程のHTMLファイルから
Koganeの日記携帯ホームページ02/01の日記11:53研究として使用します久保直也の個人情報は090-4216-0000です。
という文を取りだし形態素解析を行う



２　・形態素解析，構文解析された情報を述語論理化RDF化する.
　　述語論理化して,RDF化しなければ意味まで考慮したマッチングが
　　取れないためRDF化を行う.

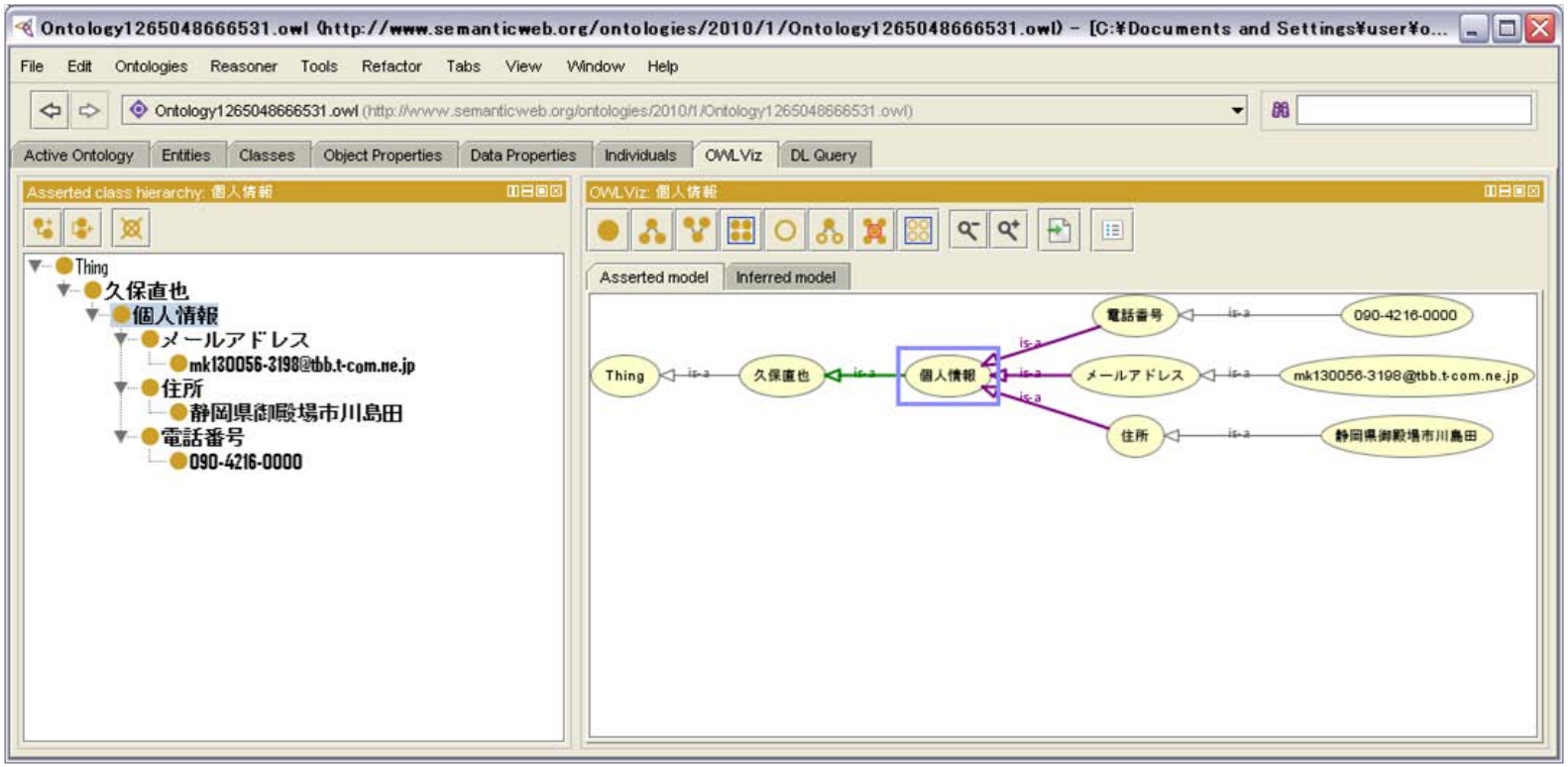
・例えば久保直也の電話番号は090−4216−0000 です，という文があった場合久保直也を主語(リソース) 電話番号を述語(プロパティ)090−4216−0000 を目的語(プロパティの値) となる.

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf=http://www.w3.org/1999/02/22-
rdf-syntax-ns#

xmlns:mr3=http://mmm.semanticweb.org/mr
3#

xml:base="http://mmm.semanticweb.org/mr3
#">
 <rdf:Description rdf:ID="久保直也">
  <mr3:電話番号>090-4216-0000</mr3:電話
番号>
 </rdf:Description>
</rdf:RDF>
```

3　オントロジーDBを記述しておく.
例えば090-4216-0000instance-of電話番号is-a 個人情報is-a 久保直也といった記述ができる. ここではオントロジーDBの記述にはprotege を使用する.



4　ACL を導く

RDF で検索された処理結果とオントロジーDBから外的要因を考慮したACL を導き出す解析エンジンによりACL を検出する.

5　Covert Channelの検出

以上により,内的なACL では読めないことになっている情報がWeb検索の結果を解析して得られた外的要因まで考慮した実質的なACL では読めると言ったような矛盾を見つけることができる.

矛盾があった場合その矛盾と経路を表示する

さらに矛盾があった場合Covert Channel の経路を表示される.

要因を考慮したACL

	S1
テンツ	R
	R

	s1	s2
電番	ϕ	RW
コンテンツ	RW	RW

4　ACL を導く

RDF で検索された処理結果とオントロジーDBがら外的要因を考慮したACL を導き出す解析エンジンによりACL を検出する.

5　Covert Channelの検出

S1 が読み書きのできない電話番号情報をS2 が読み書きすることができた場合S2 がその情報を読みS1 が読み書きのできるコンテンツにコピペしてそこからS1 が本来読み書きのできない電話番号情報を読み書きすることができるようになってしまった.といった経路を表示する.

	s1	s2
電番	ϕ	RW
コンテンツ	RW	RW

まとめ

・本稿では検索エンジンを用いたCovertChannel の検出方法を提案した。

・これにより従来のように把握したコミュニティのACLのみを用いたCovertChannel だけでは検出できないアクセス権の矛盾が存在する場合でも

・検索エンジンによって得られた情報にオントロジーを用いたセマンティックな解析手法を適用することで外的要因を考慮した場合のACL の矛盾や経路を効率よく見つけることが可能となり

・従来のCovert Channel 解析法での外的要因を考慮した場合検出できないアクセス権の矛盾が存在する, という問題点を解決することができるかもしれない.

今後の課題

- Hadoop によるRDF導出効率化
- RDFで検索された処理結果とオントロジーDBから外的要因を考慮したACLを導き出す解析エンジンの構築
- ACLの矛盾や経路を検出するのに最も適したオントロジーの記述法の検討