自然言語文からのプライバシー情報検知システム

〇渡辺夏樹* 水谷桂子* 吉浦裕*

* 電気通信大学大学院 電気通信学研究科

背景：Web上のコミュニケーション

- Blog、SNS等のメディアが注目
  - ヒューマンコミュニケーションの活性化
    - 新しいつながりの開拓、既存のつながりを強める
  - プライバシー情報の漏洩（SNS利用者への調査＊）
    - SNS利用には危険が潜むと考えている(51%)
    - プライバシーが保護されない（37%)
    - 子供の安全が保護されない（32%)
  - 誹謗中傷等の不適切な表現

＊「シノベイトによるSNSに関する意識調査」/シノベイト株式会社

従来技術

- 開示制御
  - ルールの設定が容易ではない
    - 話題の事前予測ができない
    - 語句の予想ができない
- mixi
  - 日記、閲覧者ごとにユーザが開示制御
    ⇒手間がかかる
  - 日記全文が見れなくなる

⇨ コミュニケーションの阻害

研究目的

BlogやSNSに投稿された日記から
プライバシー情報を検知

プライバシー情報漏えいに対する言い換えや
ユーザへの警告が可能

特定ユーザに対する追跡調査

- 対象
  - 50代男性 mixiユーザ 2005年から1年半の日記
- プライバシーの記述
  - 直接

    日記中に直接表現されているもの

    Ex)実は多摩市の住民です。⇒ 住所:多摩市
  - 想起

    日記中の表現から連想・検索・推論により認識できるもの

    Ex)昭和58年生まれ ⇒ 25歳

    H科 ⇒ 人間コミュニケーション学科

プライバシー情報	直接		想起		合計
	本人	閲覧者	本人	閲覧者	
病歴	14	2	1	0	17
電話やメールの通信履歴	3	0	1	0	4
アドレス帳の内容	12	0	3	0	15
家族に関する具体的な事柄	26	3	6	0	35
学歴職歴	39	1	11	0	51
身長や体重	3	0	0	0	3
電話やメールの	3				6
家族構成	3				1
氏名	5				6
住所	7				3
生年月日	13				23
職業	86	9	42		137
	7	5	2		14
		0	0		1
	91	6	5	0	102
知人や友人に関する具体的な事柄	8	0	2	0	10
商品やサービスの購入記録	10	1	0	0	11
合計	366	37	105	1	509

想起された情報
20.8%

調査対象全体
75.8%

直接書かれていた情報
79.2%

本研究が目指す処理

原文

地下鉄の駅

東京の市

で会社の近くの初台のマンションで暮らす
は多摩市の住民です。

危険度：高

月曜から土曜まで会社の近くの初台のマンションで暮らす
単身赴任者は、実は多摩市の住民です。

危険度：中

月曜から土曜まで会社の近くの渋谷区のマンションで暮らす
単身赴任者は、実は多摩市の住民です。

技術の要件

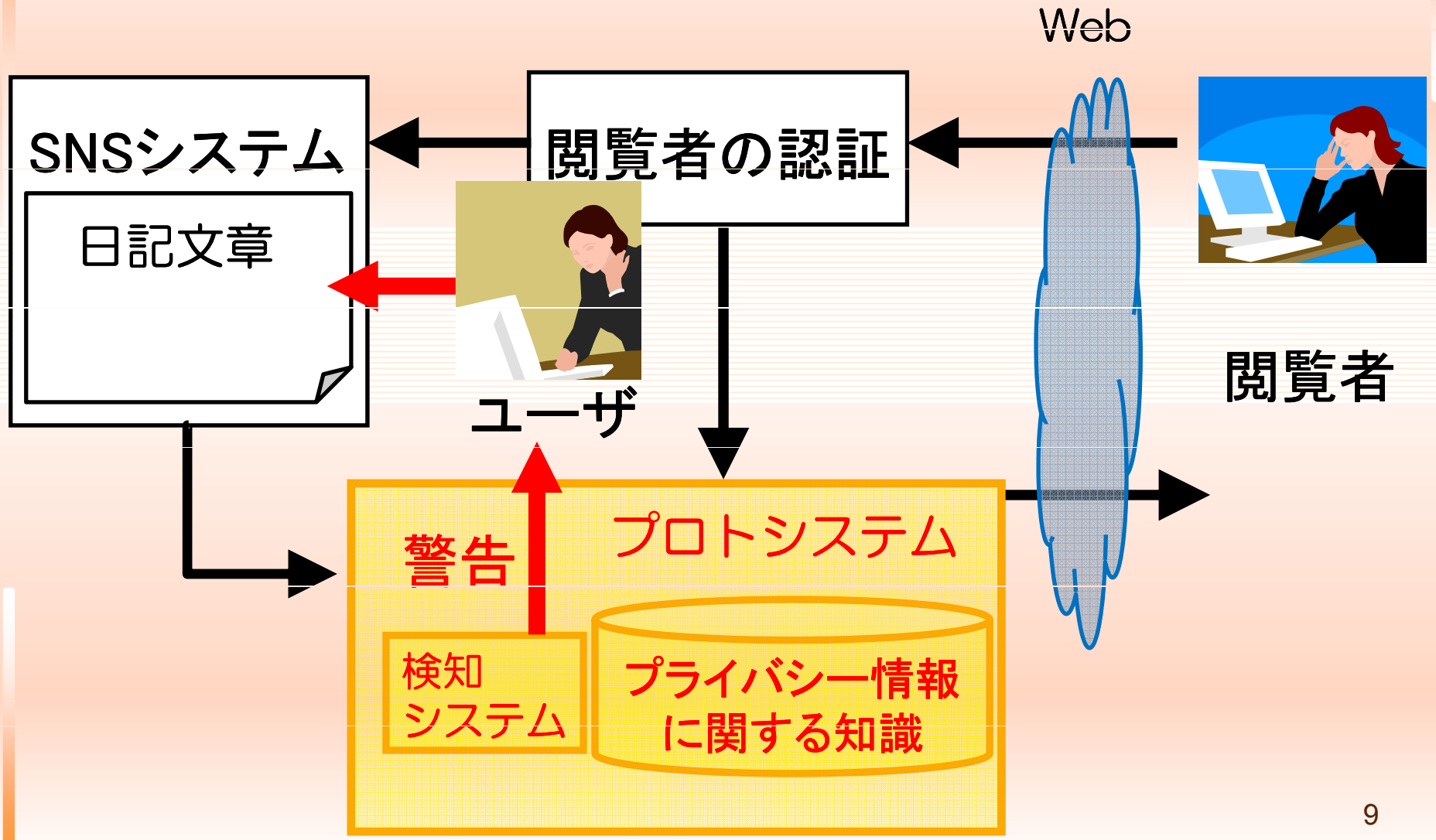
- コミュニケーション中の文章をチェック

  プライバシー情報を洩らす語句を検知

1．検知の要件

- どんな情報が洩れるか言葉から推測
- ユーザ負担を最小限に
- どの程度プライバシーか推定

本日発表のプロトシステム



プライバシー情報に関する知識

特徴語

NGワード

月曜から土曜まで会社の近くの初台のマンションで暮らす
単身赴任者は、実は多摩市の住民です。

種類語

具体的なプライバシー情報を表す語句

プライバシー情報の種類を表す語句

プライバシー情報に特徴的な語句

知識の不完全性

- 事前準備のコスト大
  - プライバシー情報は多種多様
  - 人によって異なる

1. ユーザ共通部分：
   インターネットコンテンツから事前学習

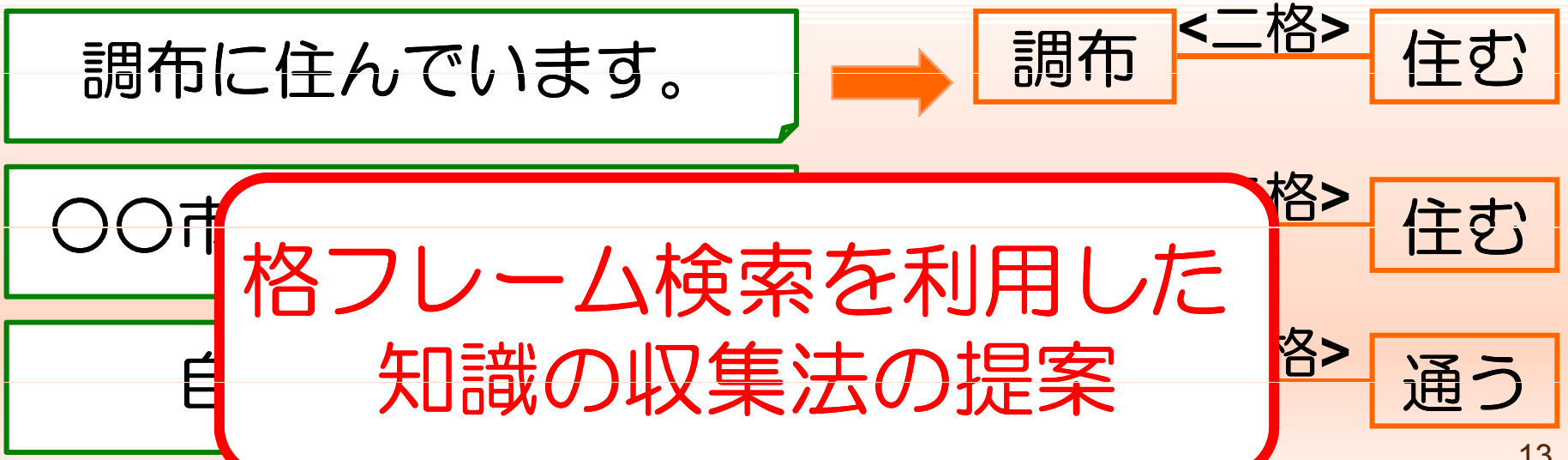
2. ユーザ依存部分：
   検知処理中にインターネットコンテンツを利用

知識の事前学習（ユーザ共通）

共通知識の事前収集の方針

- 共通知識：種類語・特徴語
  - 手動で準備した種類語・特徴語と格構造が同じ語句
    ⇒　種類語や特徴語になる可能性が高い

- 格構造とは・・・



格フレーム検索ツール (河原2006)

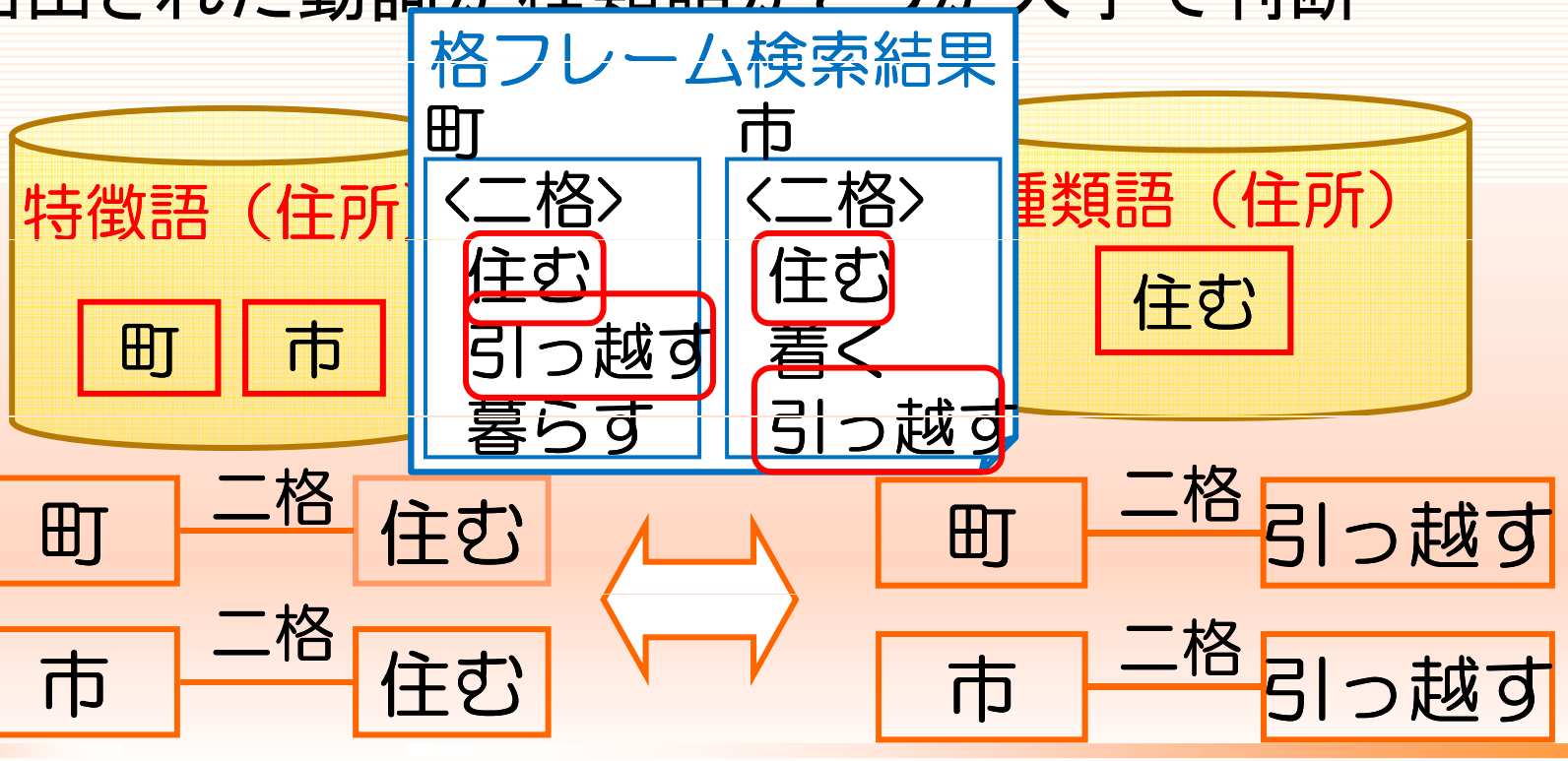
- 格フレームに該当する語句を抽出するツール
  - コーパス：Web上の5億の自然言語文

- 格フレーム検索ツールにより抽出できる語句
  - ある語句を格としてとりうる動詞 ⇒種類語の収集
  - ある動詞の格に該当する語句 ⇒特徴語の収集

- たとえば…
  - 「大学」を検索すると「通う(ニ格)」「入学(ヘ格)」
  - 「住む」を検索すると「調布(ニ格)」「一人(デ格)」

種類語の収集

1. 既存知識に対する格を格フレーム検索で決定
2. 割出した格と同じ格の動詞を特徴語に対する格フレーム検索結果から抽出
3. 抽出された動詞が種類語かどうか人手で判断



実例(住所に関する種類語の収集)

- 手動で準備した住所に関する知識
  - 特徴語：市・区・町・村
  - 種類語：住む

二格

- 格フレーム検索により格を決定する
- 検索結果

	市	区	町	村
総数	11196	3243	8299	3851
二格の語句数	1650	632	1470	851
抽出した種類語	50	38	53	40

実例(住所に関する種類語の収集)

- 手動で
  - 特徴
  - 種類
- 格フレ
- 検索結

うまれる	帰郷	産まれる	生まれ育つ	通学
暮らす	帰国	住	生む	通勤
移住	帰省	住い	生れる	定住
移築	帰属	住まい	生育	越す
移転	居る	住みつく	生活	転勤
育つ	居座る	住み込む	すむ	転校
引っ越し	居住	住む	属	転入
引っ越す	居着く	住める	属す	届け出る
引越	健在	所在	属する	納付
永住	在る	所属	滞在	赴く
転居	在学	常駐	誕生	暮らし
下宿	在勤	新築	駐在	暮らす
帰り着く	在住	生きる	駐屯	里帰り
帰る	在職	生まれ	駐留	
帰還	在籍	生まれる	通う	

計73語

二格の

抽出し

実例(所属に関する種類語の収集)

- 手動で準備した所属に関する知識
  - 特徴語：大学・高校・学校・会社
  - 種類語：通う

二格

- 格フレーム検索により格を決定する
- 検索結果

	大学	高校	学校	会社
総数	7225	2190	6304	11470
二格の語句数	1044	321	1287	2013
抽出した種類語	48	31	51	59

実例(所属に関する種類語の収集)

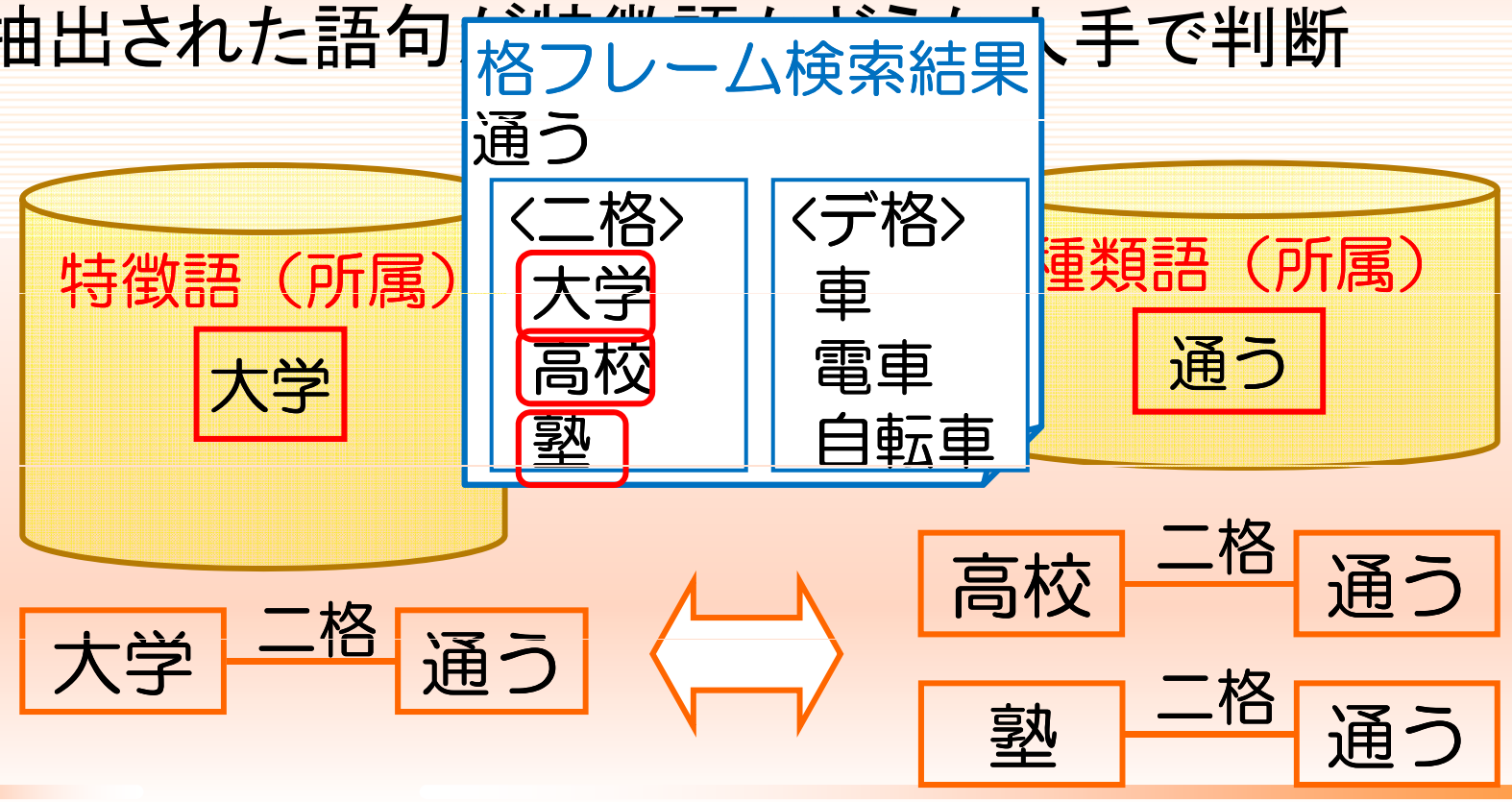
-
-
-
-
- 抽

異動	合格	従属	着任	転任	配属	エントリー
移籍	採用	出勤	通い	登校	赴く	かよう
解雇	居着く	出向	通い詰める	登録	赴任	かよう
入社	在職	出社	通う	働く	復学	つくす
入学	在籍	出席	通える	内定	復帰	つとめる
帰属	残留	出張	通ったりする	馴染む	復職	なじむ
帰着	受かる	出入り	通わせる	入り	編入	はいる
在学	受験	所属	通学	入る	奉公	リストラ
勤	就学	常駐	通勤	入院	奉仕	
勤め	就業	進学	転勤	入会	奉職	
勤める	就職	属す	転校	入校	遊学	
勤務	就任	属する	転出	受かる	落ちる	
雇う	就労	滞在	転職	引っ越す	落ち着く	
雇用	修学	遅れる	転籍	入隊	留まる	
貢献	従事	遅刻	転入	派遣	留学	

計98語

特徴語の収集

1. すでにある知識に対する格を格フレーム検索で決定
2. 割出した格と同じ格の単語を種類語に関する格フレーム検索結果から抽出
3. 抽出された語句が特徴語かどうか人手で判断



実例(所属に関する特徴語の収集)

- 手動で準備した所属に関する知識
  - 種類語:通う
  - 特徴語:大学

  } ニ格

- 格フレーム検索により格を決定
- 検索結果

	通う
ニ格の語句数	1125
抽出した特徴語	80

実例(所属に関する特徴語の収集)

幼稚園	短大	歯科	教会	屋敷	クラブ
薬局	大学校	施設	機関	駅	クラス
役所	**大学院**	市場	眼科	院	キャンパス
分校	大学	産婦人科	館	医大	カレッジ
部	食堂	産院	学部	医院	オフィス
病院	床屋	高等	学舎	センター	
美容院	小学校	高専	学校	ゼミナール	
農園	女学校	高校	学級	セミナー	
内科	書店	校舎	学科	ゼミ	
道場	所	校	学園	スクール	
店舗	塾	工房	学院	スイミング	
店	獣医	工場	外科	ジム	
中学校	耳鼻科	劇場	会所	サロン	
中学	耳鼻咽喉科	銀行	会社	サークル	
中	寺	教室	会館	クリニック	

計80語

検知時の利用（ユーザ個別）

未登録NGワードの検知

ユーザが登録していないNGワードに対応

月曜から土曜まで会社の近くの初台のマンションで暮らしています。

種類語

Web知識の利用

初台の属性が地名

住所＝初台

住所を例としてWeb知識の利用を検討

知識ソースの比較

- 未登録NGワードの検知の実現

⇨ 適したWeb知識を調査

比較対象
- Wikipedia（集合知）
- Googleマップ（検索サイト）
- マピオン（商用サイト）
- JUMAN（形態素辞書）

比較項目
- 知識の網羅性
- 更新速度
- 研究利用の権利処理

知識の網羅性

○Wikipedia ○Googleマップ ○マピオン ○JUMAN	渋谷,上原,鶯谷町,初台,千駄ヶ谷,広尾, 本町,円山町,桜ヶ丘町,笹塚,猿楽町
○Wikipedia ○Googleマップ	恵比寿,松涛,神南,富ヶ谷
×Googleマップ ○マピオン ×JUMAN	本町(東京),神山町(東京,徳島), 神宮前(愛知),青葉台(宮城,東京,埼玉)
×Wikipedia ×Googleマップ ○マピオン ×JUMAN	神山町(北海道,大阪,兵庫) 青葉台(大阪,島根,愛媛,山口), 北沢(愛媛)

マピオン＞Wikipedia＞Googleマップ＞JUMAN

知名度大

知名度小

全項目の比較

	知識の網羅性	更新速度	研究利用の権利処理
マピオン		○	×
Wikipedia	○	○	○
Google マップ		○	×
JUMAN	×	×	○

オンライン

オフライン

NGワードを想起させる語句の検知

個人写真撮影の追加撮影会が決定
平成17年11月14日、15日　10時から19時　リサージュ1階
撮影対象は卒業研究着手者。
アルバム購入は卒業、修了が必須条件ではありません。
ちなみに、H科は昼夜合わせて52人が撮影済みらしいです。

NGワードを

個人写真撮影の追加
平成17年11月14日
撮影対象は卒業研究
アルバム購入は卒業、
ちなみに、H科は昼

- 「リサージュ1階」から「電気通信大学」を想起

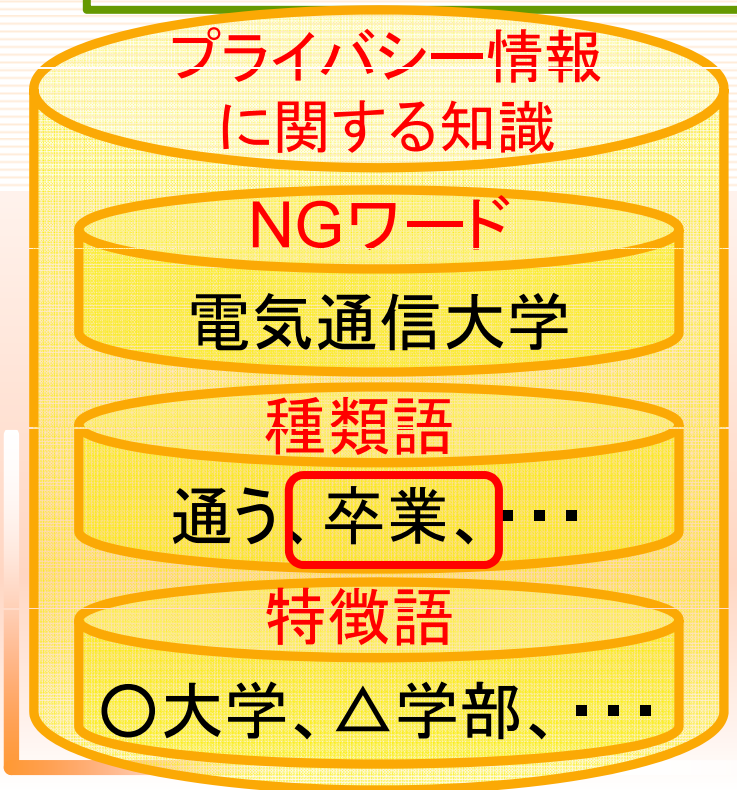
⇒文章中に直接記述されていないプライバシー情報を発見

- 所属を例としてNGワードを想起させる語句（例:リサージュ1階）を自動的に検知

起動方法

種類語

個人写真撮影の追加撮影会が決定
平成17年11月14日、15日　10時から19時　リサージュ1階
撮影対象は卒業研究着手者。
アルバム購入は卒業、修了が必須条件ではありません。
ちなみに、H科は昼夜合わせて52人が撮影済みらしいです。



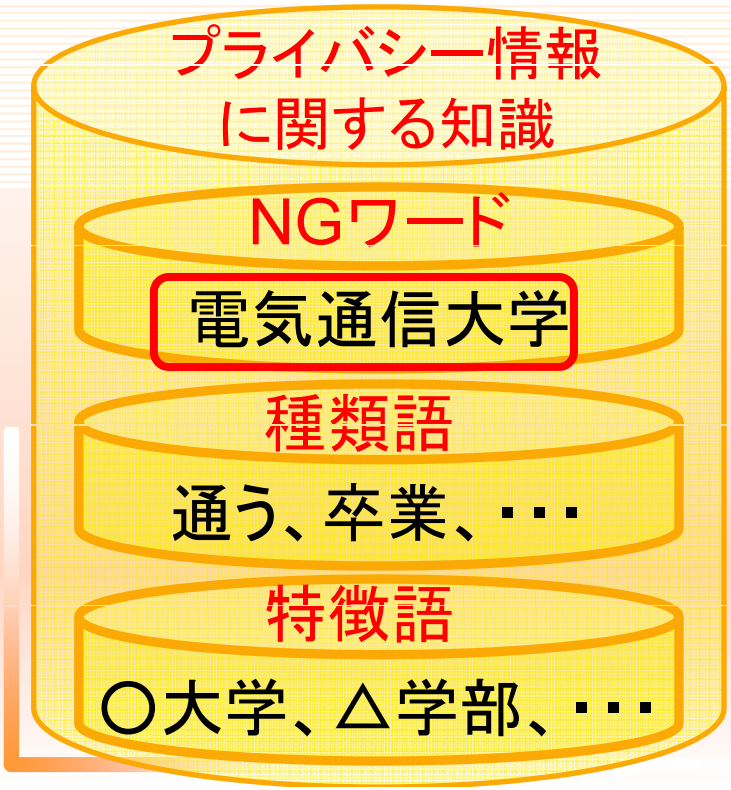
NGワード(大学名)を想起する可能性

文中の名詞を抽出

・・・ リサージュ1階 撮影対象 ・・・

Web検索

NGワードを想



ウェブ 画像 地図 ニュース 動画 Gmail more ▼

Google "リサージュ1階" 検索 検索オプション 表示設定

ウェブ全体から検索 日本語のページを検索

ウェブ

電気通信大学地域貢献部門

会場, 電気通信大学 創立80周年記念会館**リサージュ1階**. 内容, 実験:ホバークラフトを作ろう (工作教室と合同)。(資料). 空気を真下に噴出し、浮き上がって進む車を製作した。どの車も同じ動き方をしないのが見ていておもしろい。...

www.dcc.uec.ac.jp/inv80/20051.html - 16k - キャッシュ - 関連ページ

指導員募集

活動時間, 下準備のため 10時集合(創立80周年記念会館(通称 **リサージュ) 1階**) 実地指導時～4時(概ねリサージュ3階) ... 作業時間, 午前10時～午後5時、**リサージュ1階**にて. 3), 出張講座が年に何回かあります。これまでのようすはここを ...

www.dcc.uec.ac.jp/inv80/wanted.html - 8k - キャッシュ - 関連ページ

www.dcc.uec.ac.jp からの検索結果 »

[PDF] 社団法人目黒会 首都圏総支部2008年度総会および懇親会

ファイルタイプ: PDF/Adobe Acrobat - HTMLバージョン

リサージュ 1階北側前庭. 注) 総会後、午後に電気通信大学創立90周年記念 特別講演II 久夛良木健氏. (. (. 株). ソニー・コンピュータ. エンタテインメント名誉会長、. ソニー株式会社シニア・テクノロジーアドバイザー). の講演会に参加 ...

www.megrokai.or.jp/metro/info/20081120/2008MetroSoukaiAgenda_20081120.pdf - 関連ページ

事務局より

この絵は、第5回全国公募「北の大地展」においてビエンナーレ受賞作品で、創立80周年記念**サージュ) 1階**の応接室に展示しておりますので、リサージュにお越しの際は是非ご覧下さい. 会事務局 ...

www.megrokai.or.jp/megurokai/office/osirase/kaigakisou.html - 32k - キャッシュ - 関連ページ

www.megrokai.or.jp からの検索結果 »

[PDF] 11. まとめ−調布方式による地域情報化の今後のあり方−

ファイルタイプ: PDF/Adobe Acrobat - HTMLバージョン

電気通信大学**リサージュ 1 階**. ①活動報告 ②3月の IT 講座 ③その他のｲﾍﾞﾝﾄ ④調布. FM での番組

結論

- SNSの日記からのプライバシー漏えいについてを実例調査
  - 一年半の日記中に17種類、509件のプライバシー情報
- 自然言語文からのプライバシー情報検知技術
  - 検知に必要な知識のうち、ユーザ共通部分をインターネットコンテンツから事前学習
  - ユーザ依存部分は検知処理の中でインターネットコンテンツを検索し利用

ご清聴ありがとうございました