匿名加工情報公表サイト調査
(1)自動クローラーシステムの開発

菊池研究室　B4小野敦樹

匿名加工とは？

・特定の個人を識別することができないように個人情報を加工すること

お客様情報
クレジット利用情報の
匿名加工

統計データ等
への利用

氏名・年齢	明治太郎・(22歳)
カード番号	1234-1234-1234-1234
住所	東京都中野区中野4-21-1
利用日	2019年4月20日
利用店舗	明大マート

匿名加工

氏名・年齢	削除・(20代)
カード番号	削除
住所	東京都
利用日	2019年4月
利用店舗	コンビニエンスストア

研究背景と問題点

・匿名加工情報を作成･提供する際に，事業者は
　個人に関する情報の項目(氏名，カード番号，住所等)の公表が義務

・個人情報保護委員会は約300社の公表サイトを手作業で収集している

⇒①全国すべての事業者を網羅することが困難

⇒②新たな公表サイトの収集が困難

匿名加工情報の作成・第三者提供について

三井住友海上火災保険株式会社

匿名加工情報の作成・第三者提供について

当社は、当社が保有する当社従業員に係る以下の情報について、特定の個人を識別すること及び作成に用いる個人情報を復元することができないよう適切な保護措置を講じたうえで匿名加工情報として作成し、健康経営に関する研究・分析のために第三者に提供いたしますので、公表します。

個人に関する情報の項目

1. 健康診断情報および問診情報

   生年、性別、健診年月、健診項目データ（数値等）、問診項目データ（病歴、生活習慣等）

2. ストレスチェック情報

   加工を施した属性情報（勤務地域、所属部支店、役職、職種）、ストレスチェック回答および結果、医師面接状況

提供の方法

パスワードにより保護された電子ファイルを外部記憶媒体で手交

(例)三井住友海上火災保険株式会社(https://www.ms-ins.com/privacy/anonymous.html)

研究目的

匿名加工情報公表サイトのクローリングを行い，

匿名加工情報公表サイトの自動取得を試みること

① 手動での匿名加工情報公表サイト調査

匿名加工情報公表サイトから個人に関する情報の項目を手動取得し，データセットの作成をした。

（データセット例）

企業名	個人に関する情報の項目	提供手法	URL
三井住友海上	生年月日，性別，…	電子ファイルを外部記憶媒体	https…com
日立総合病院	病名，薬剤情報，…	提供先指定サーバーにアップロード	https…com
株式会社スギ薬局	処方日，調剤日，…	電子メールによる送信	https…com
新横浜障害年金センター	障害状態区分，年金受給額	第三者が利用できるようにサーバーにアップロード	https…com

調査方法

期間：2019年5月～2019年9月

検索プロバイダ： Google

調査結果

期間	匿名加工情報公表サイト数
2019年5月～2019年8月	308

308件取得推定時間
13時間10分32秒

匿名加工情報公表サイト作成企業308社を
日本標準産業分類を基とした26業種に分類.

医療業(病院), 健康保険組合, 小売業(薬局)
含まれることから, 医療データが多い.

医療業（病院）	58	労働派遣業	4
情報通信業	45	年金相談センター	4
健康保険組合	34	不動産業	4
小売業（薬局）	29	一般社団法人	3
製造業	21	運輸業	3
保険業	19	卸売業	3
情報サービス業	17	健康保険協会	3
サービス業	14	公益社団法人	2
金融業	14	特別民間法人	2
小売業	12	建設業	1
医療業（製薬）	5	信用格付け機関	1
社会保険, 社会福祉	5	弁護士会	1
教育, 学習支援	4	保険労務士法人	1

手動サイト調査による困難点

①網羅性に欠ける

②サイト収集に多大な労力(時間)を要する

③業種分類にも多大な労力(時間)を要する

②自動クローラーシステムの開発

・Googleが提供をしているAPIサービス，検索結果のサイトタイトルとURLを取得.


```
if __name__ == '__main__':
 target_keyword = '匿名加工情報 公表'
 exclude_keyword = '法律事務所 個人情報保護委員会'
 print('検索キーワード:  ' + str(target_keyword))
 print('検索除外キーワード:  ' + str(exclude_keyword))
 getSearchResponse(target_keyword)
```

(プログラム一部例)

Google Custom Search API

・Googleが提供をしているAPIサービス，検索結果のサイトタイトルとURLを取得.


```
if __name__ == '__main__':
 target_keyword = '匿名加工情報 公表'
 exclude_keyword = '法律事務所 個人情報保護委員会'
 print('検索キーワード:  ' + str(target_keyword))
 print('検索除外キーワード:  ' + str(exclude_keyword))
 getSearchResponse(target_keyword)
```

(プログラム一部例)

Google Custom Search API の問題点(1)

(1)API仕様により，上位100件の検索結果のみしか取得をすることができない.

⇒少なくとも300件以上の匿名加工情報公表サイトが存在している.

💡問題点の解決法

(1)複数のシードキーワードの導入

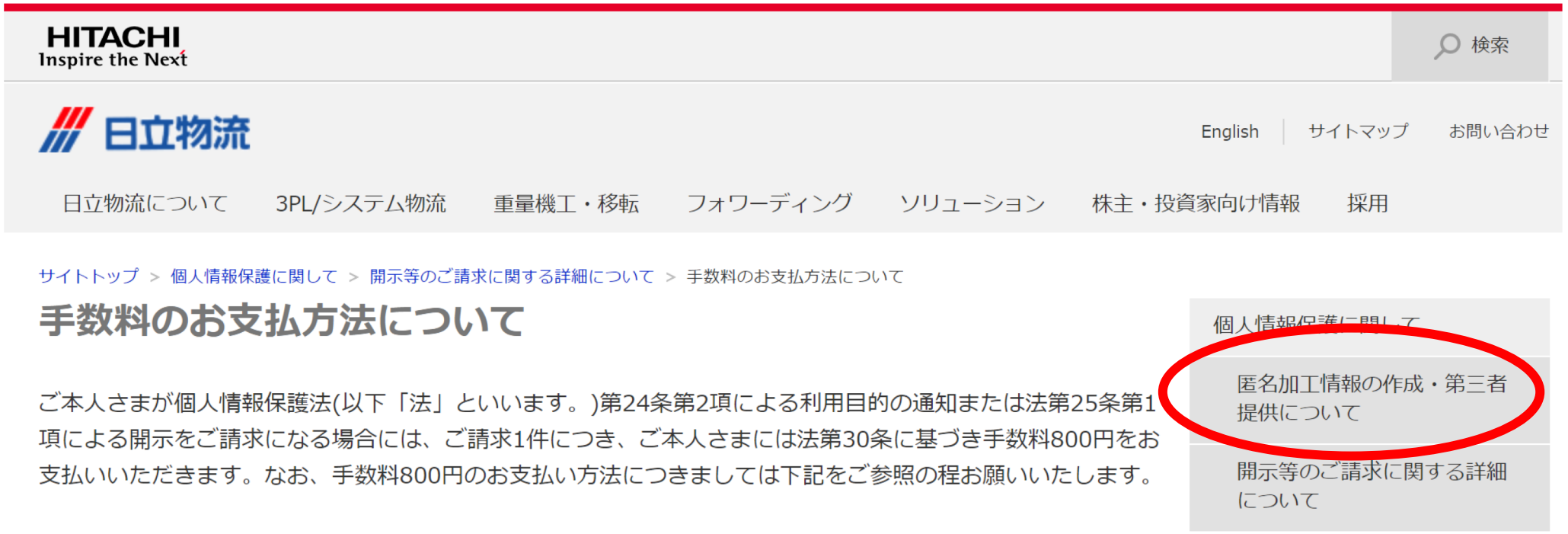
業種	共通キーワード	シードキーワード
病院	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	病院
薬局	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	調剤
健康保険関連	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	健康保険組合 or 健康保険協会
生命保険	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	生命保険
銀行	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	銀行
年金関連	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	年金
その他	匿名加工情報 作成 提供 –法律事務所 -個人情報保護委員会	-病院 –健康保険組合 –銀行 –年金 –生命保険

⇒結果的に業種分類の手間も軽減される。

Google Custom Search API の問題点(2)

(2)検索キーワードを絞っても，非該当サイトが抽出されてしまう。

(例)日立物流:手数料のお支払方法について



💡問題点の解決法

(2)匿名加工情報公表サイトのサイトタイトルに含まれる

「匿名加工情報」

「個人情報保護」

「プライバシーポリシー」

いずれかの単語を含むサイトのみを抽出

⇒98.5%(321サイト/326サイト)の精度で該当サイトの抽出が可能

システム構成図

Google Custom Search API

キーワードによる検索
(例)匿名加工情報 提供

(1)

json

検索結果上位100件のデータファイル×7個

(2)

7業種×100件

url

クローラー

・該当サイトと非該当サイトの選別
・企業webサイトのクローリング

(3)

前スライドの3ワードによって選別

url

html

Webサイト×700件

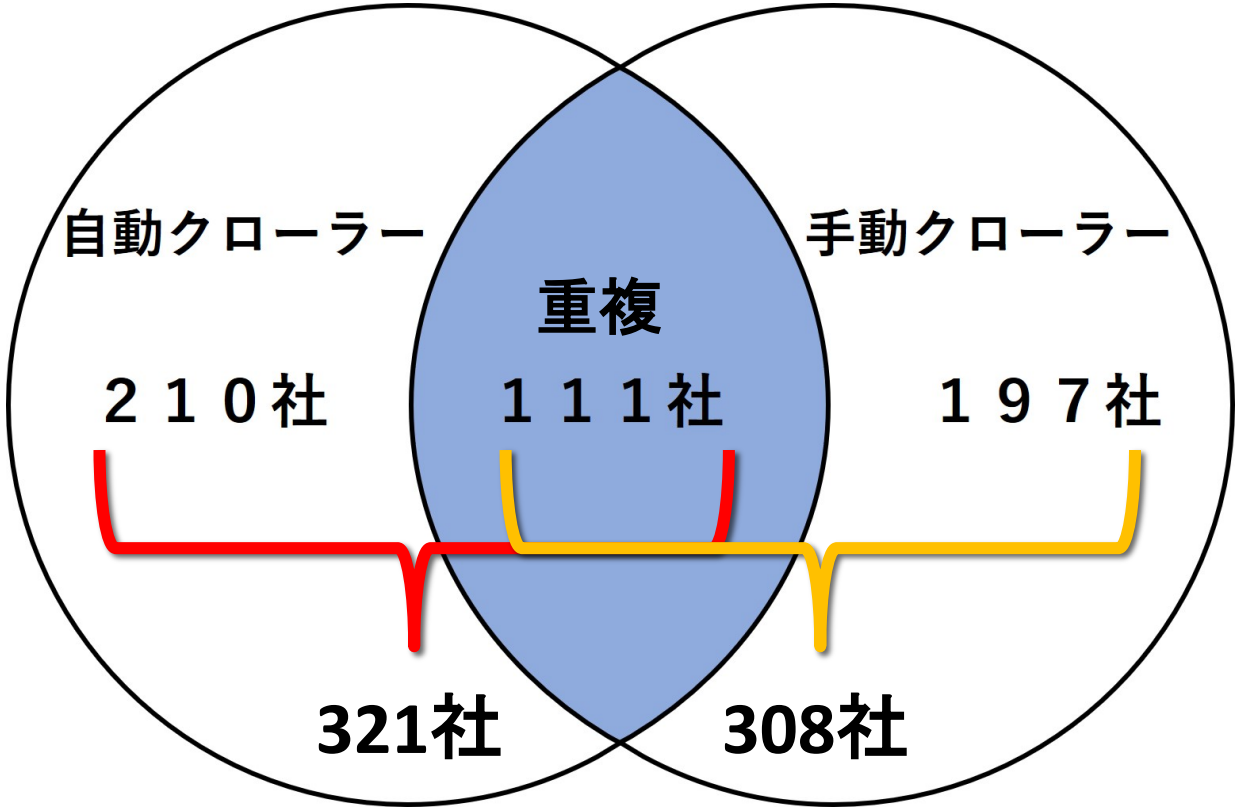
③自動クローラーシステムの精度評価

自動クローラーによる匿名加工情報公表サイト取得結果

業種･団体	重複データ	新データ	合計	手動取得データ
病院	28	48	76	58
薬局	15	58	73	29
健康保険関連	8	78	86	37
生命保険	3	0	3	5
銀行	2	2	4	4
年金関連	0	4	4	6
その他	55	20	75	169

自動クローラーと手動調査との比較

自動クローラー 手動調査取得データの差異



自動クローラー手動調査データ取得時間

自動クローラー 平均取得時間/件
4.02秒/件
手動調査平均取得時間/件
2分34秒/件

④まとめ

①網羅性に欠ける

⇒手動より多い321件のデータを取得

②サイト収集に多大な労力(時間)を要する

⇒手動と比較し，1件あたり2分30秒速く収集が可能

③業種分類にも多大な労力(時間)を要する

⇒シードキーワードにより業種分類の手間軽減