匿名加工情報取扱事業者を調査するクローラーシステムの開発

金子 侑紀，小野敦樹，伊藤聡志，菊池浩明(明治大学)
服部充洋，飯田泰興，藤田真浩，山中忠和(三菱電機)

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
匿名加工例

統計データへの利用

氏名	明治太郎
年齢	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	削除
年齢	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

匿名加工情報取扱事業者の公表事例

法36条 (匿名加工情報取扱事業者の義務) 4項

個人情報取扱事業者は(中略)あらかじめ，個人に関する情報の項目及びその提供の方法について公表(中略)しなければならない.

2. 匿名加工情報に含まれる項目

【1】 お客様の個人属性情報（性別、年齢（年代）、住所（都道府県）、カード入会歴（入会年月））

【2】 カード利用情報（利用日（利用年月）、利用金額、利用店舗情報）

3. 匿名加工情報の第三者提供

当社が作成した匿名加工情報を第三者に提供する場合の目的、匿名加工情報に含まれる項目及びその提供方法は以下の通りです。

(1) 目的

当社が提携する会社で統計化情報を作成するため。

(2) 第三者に提供する匿名加工情報に含まれる項目

上記2【1】【2】の項目

(3) 提供方法

データファイルを暗号化し、セキュリティが確保された手段で提供を行います。

(例) JCB.Co., Ltd (https://www.global.jcb/ja/policy/privacy/personal.html)

公表サイト調査による困難点

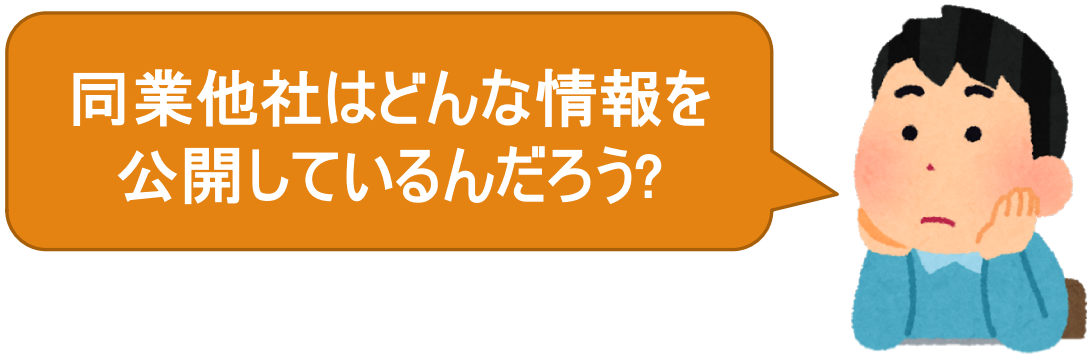
1．網羅性に欠ける

2．サイト収集に多大な労力(時間)を要する
- 予備調査で308社分の収集に14日かかった

3．業種分類にも多大な労力(時間)を要する
- 308社を26業種に分類するのに12日かかった

研究目的

1. 公表サイトを自動取得
2. 匿名加工情報の自動分類
3. 個人に関する情報の項目と提供方法の抽出



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)クローラー準備

json

検索結果上位100件の
データファイル×7個

7業種×100件

url

(2) クローラーシステム

・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

正規表現

・個人に関する情報
・提供方法

(3) 情報抽出

Webサイト×700件

html

(1) 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)

問題点：API仕様により，上位100件の検索結果のみしか取得できない

💡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

⇒課題2：業種分類も実施される。

（3）正規表現による情報抽出

	正規表現パターン
1	(次のとおり|下記|以下)+([¥s¥S]+)¥n[¥s¥S]+提供[の]*方法
2	DPC+([¥s¥S]+)¥n[¥s¥S]+提供[の]*方法
3	情報[の]*項目([¥s¥S]+)¥n[¥s¥S]+提供[の]*方法
4	項目は([¥s¥S]+)です
5	([¥s¥S]+)(上記項目|提供[の]*方法|第三者に提供)

調査方法

期間：2019年11月18日

検索プロバイダ： Google, 700件

手作業での評価

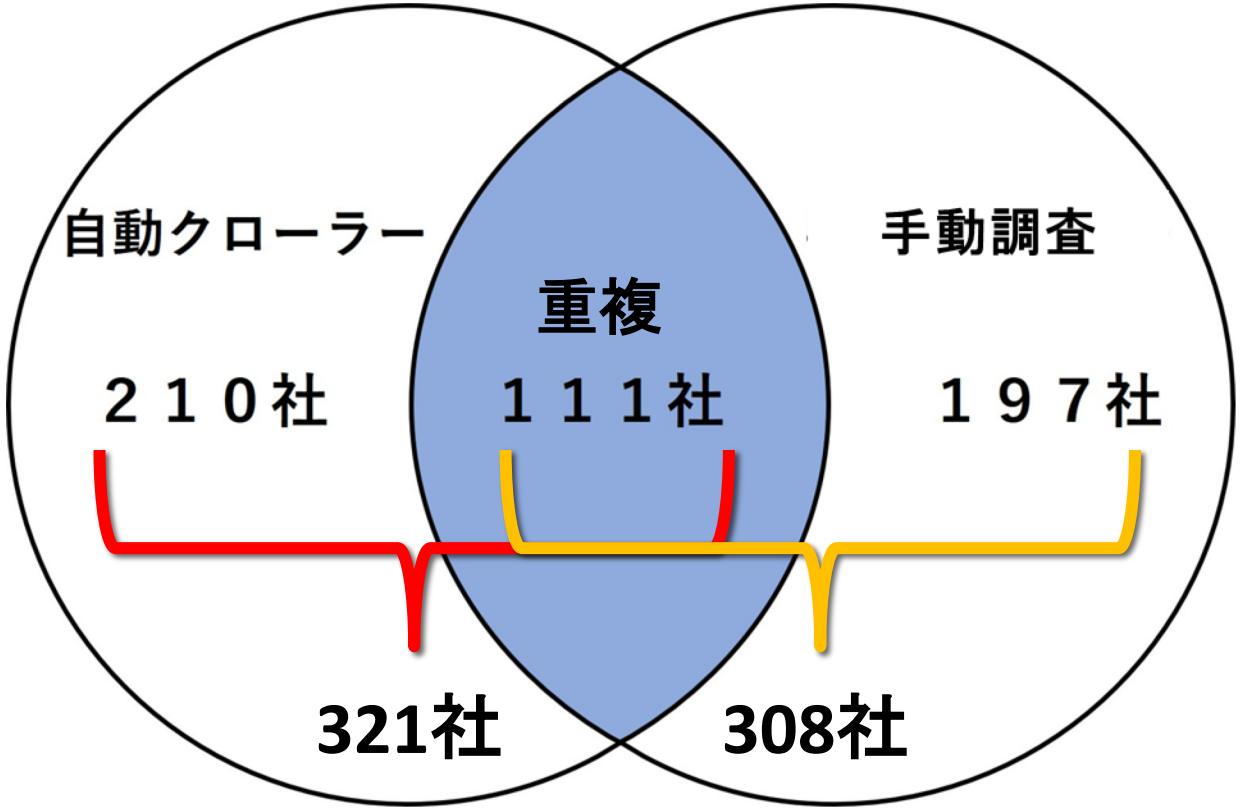
期間	匿名加工情報公表サイト数
2019年5月～2019年8月	308

308件取得推定時間
13時間10分32秒

企業名	個人に関する情報の項目	提供手法	業種
兵庫医科大学病院	郵便番号，生年月日， 保険者番号…	提供先が運用管理するサーバへのアップロード or 外部記録媒体を郵送	病院
マツモトキヨシ	会員番号，年齢，性別， 購入履歴…	電子メールによる送信	薬局
伊藤忠連合保険組合	氏名，生年月日，年齢， 被保険者記号，医師の氏名， …	セキュリティが確立された方式	健康保険
青森銀行	債務者番号，業種，生年月日 財務情報を含む信用情報…	セキュリティが確保された電子媒体	銀行
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	年金関連
公文教育研究会	公文式学習履歴、配偶者の有無、職業…	電子ファイル or サーバーへのアップロード	その他
SHARP	性別，年代，広告識別情報 位置情報…	パスワード等によりアクセス制御をしているサーバー	その他

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

自動クローラーシステムの収集結果

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

情報抽出結果の一部

企業名	個人に関する情報の項目	提供方法
株式会社日本医薬総合研究所	年齢(生年)，性別，処方せん情報，調剤情報，各種アンケート回答	電子メール，CD-ROM，USB等の外部記憶媒体，HTTPS
セキ薬局	氏名，生年月日，被保険者記号番号，公費受給者番号，医師氏名，処方日，調剤日，性別，生年，処方・調剤履歴	セキュリティが確立された転送方式
イオン銀行	性別，年代，申込手段，当行普通預金口座の有無，現在の借入の有無，契約から初回借入までの経過日数，現在の返済実績等	パスワードで保護し，CD-ROMで手交

正規表現による抽出精度

	正規表現パターン	件数
1	(次のとおり|下記|以下)+([¥s¥S]+)¥n[¥s¥S]+提供[の]*方法	44
2	DPC+([¥s¥S]+)¥n[¥s¥S]+提供[の]*方法	14
3	情報[の]*項目([¥s¥S]+)¥n[¥s¥S]+提供[の]*方法	4
4	項目は([¥s¥S]+)です	3
5	([¥s¥S]+)(上記項目|提供[の]*方法|第三者に提供)	60
計		125

抽出成功	125
抽出したデータが過剰	27
抽出失敗	**167**
計	319

例1)抽出に失敗

具体的にはオーディエンスデータから個人を特定できる

特異なデータ、個人を特定できる可能性のあるデータ等を排除または置き換え、同じ属性を持つデータが同一データセット内においてデータ件数に応じ複数件以上ある状態になるまで匿名加工する等を実施します。

株式会社マイデータ・インテリジェンス

https://www.meyportal.com/policy/user_protect

例2)抽出対象が含まれていない

(1) 当行は、匿名加工情報を作成した場合には、匿名加工情報に含まれる個人に関する情報の項目を公表いたします。

(2)当行は、匿名加工情報を第三者提供する場合には、

提供する匿名加工情報に含まれる個人に関する情報の項目および提供の方法について公表するとともに、提供先に、提供される

情報が匿名加工情報である旨を明示いたします。

株式会社 阿波銀行 http://www.awabank.co.jp/policy/privacy/

⇒匿名加工情報を取り扱っているのか不明瞭

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

第36条違反の可能性がある公表サイトを42件発見した