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

小野 敦樹 †

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室 †

1 はじめに

2017 年 5 月 30 日に全面施行された改正個人情報保護法によって，中小企業をはじめとする全ての事業者が個人情報保護法の対象となった．また，匿名加工情報という新たな情報の類型が定義されたことにより，一定の条件の下で，本人の同意がなくても第三者提供や目的外利用が可能となった．しかし，匿名加工情報を第三者提供をする際に，匿名加工情報取扱事業者は提供する情報に含まれる個人に関する情報の項目の公表が義務付けられているが，届け出は義務ではないため，現在個人情報保護委員会は匿名加工情報公表ページから，個人に関する情報の項目の公表を約 300 社分手作業で取得を行っている．

そこで本研究では，匿名加工情報公表ページのクローリングを行い，提供項目とその手法について，API を用いて自動取得を試みる．しかし，本研究で使用した API は，1 度の動作で検索結果上位 100 件のデータのみしか取得ができない制限がある為，検索対象の業種を絞り込むシードキーワードの導入をし，API とシードキーワードの組み合わせによって効果的に取得した．

2 手動での匿名加工情報公表サイト調査

自動クローラーシステムの開発にあたり，手動により，匿名加工情報公表サイトと個人に関する情報の項目を取得した．

2.1 データセットの作成

Google 社の検索システムを用いて，2019 年 5 月から 2019 年 8 月までに公表されていた，匿名加工情報公表サイトと，個人に関する情報の項目を手動取得し，データセットを作成した．作成したデータセットの一部を表 3 に示す．匿名加工情報公表サイトの検索には「匿名加工情報 + 作成 + 提供 + 提供方法-法律事務所-個人情報保護委員会」の検索語を用いた．

2.2 データセット作成による調査結果と考察

2019 年 5 月から 2019 年 8 月までに公表されていた匿名加工情報公表サイト数は表 1 の通り．

表 1 手動での匿名加工情報公表サイト調査結果

期間	匿名加工情報公表サイト数
2019/05-2019/08	308

本調査により取得した 308 企業を，日本標準産業分類を基とした 26 業種に分類をした．匿名加工情報公表サイト企業業種別数を表 2 に示す．上位 5 業種に，医療業 (病院)，健康保険組合，小売業 (薬局) が含まれていることから，匿名加工情報公表データには医療データが多く含まれていることが分かる．

表 2 匿名加工情報公表サイト企業業種別数

医療業 (病院)	58	職業紹介, 労働者派遣業	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, 介護事業	5	弁護士会	1
教育, 学習支援業	4	保険労務士法人	1

2.3 手動での取得時間の推定

2019 年 11 月，被験者例より，手動クローラーで 30 件の匿名加工情報公表サイト取得に必要な時間を計測し，1 件あたりの取得時間を推測する．結果は表 6 に示す．

3 自動クローラーシステムの開発

3.1 システム構成

システムの全体構成を図 1 に示す．(1)Googel Custom Search API を使用し，キーワードによる検索を行う．(2) 検索結果上位 100 件のサイトタイトル及び URL を json

†Atsuki Ono, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

表 3　データセットの例

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,…	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	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
新横浜障害年金相談センター	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

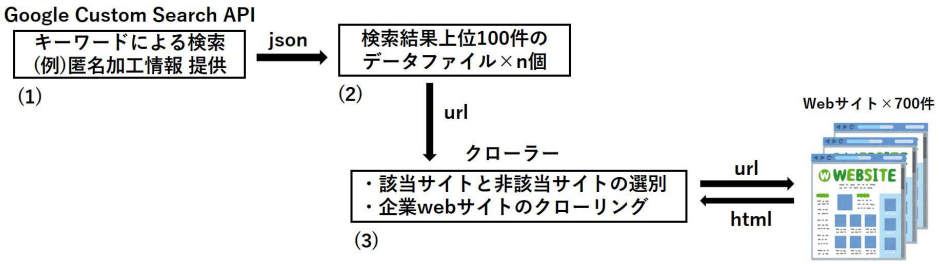

図 1　システム構成図

ファイルとして取得する．(3) サイトタイトルに特定のキーワードを含むサイトの html を保存し，テキスト化を行う．キーワードには手動調査結果に基づき，匿名加工情報公表サイトタイトルに含まれるであろう単語，「匿名加工情報」,「個人情報保護」,「プライバシーポリシー」をシードキーワードとして選定した．

3.2　Googel Custom Search API

Google Custom Search API とは，Google 社が提供をしている API サービスで，検索プログラムから検索結果を取得し，表示をする Web サイト，及びアプリケーション開発に使用することができる．概要を表 4 に示す．また，本 API は Web 検索結果を json 形式で取得する．

表 4　Google Custom Search API

Operation	Description	REST HTTP mapping
list	keyword，excludeTerms	GET

4　自動クローラーシステムの精度評価

4.1　実験方法

1．2019 年 11 月 18 日，病院，薬局，健康保険関連，生命保険，銀行，年金，その他の 7 分類に検索ターゲットを分け，計 700 件のデータを取得する (図 1(2) より，本実験の場合 $n = 7$)．検索に使用したキーワードを表 7 に示す．

2．図 1(3) による，700 件の取得データの，匿名加工情報公表サイト，非匿名加工情報サイトへの自動選別システムの識別精度を評価する．また，自動クローラーでのデータ取得時間を計測する．

4.2　実験結果

自動クローラーにより，計 321 社の匿名加工情報公表サイトの取得ができた．内訳は表 5 の通りである．321 社のデータを 2 章の手動クローラーで取得したデータと比較をしたところ，新たに 210 社の匿名加工情報公表サイトを取得することができ，111 社のサイトが重複していた．

また，自動クローラーでの 321 件データ取得時間は 21 分 32 秒であった．手動クローラーとの 1 件あたり平均取得時間の比較を表 6 に示す．

表 5　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	手作業取得データ
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

表 6　自動クローラー手動クローラーデータ取得時間

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

表 7　システム構成図 (1) における検索に使用したシードキーワード

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 提供方法 -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 提供方法 -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 提供方法 -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 提供方法 -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 提供方法 -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 提供方法 -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 提供方法 -法律事務所 -個人情報保護委員会	-病院 -健康保険組合 -健康保険協会 -銀行 -年金 -生命保険

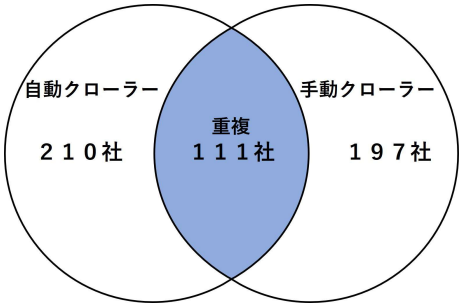

図 2　自動クローラー手動クローラー取得データの差異

表 8　自動クローラー手動クローラーによる公表サイト数

自動クローラー取得サイト数	321 社
手動クローラー取得サイト数	308 社
重複サイト数	111 社
合計サイト取得数	518 社

表 9　自動クローラー取得データ, 重複データ例

	企業名	サイト作成日
自動クローラー	(株)TOKAI コミュニケーションズ	2019 年 10 月 1 日
自動クローラー	(株) リオン	2019 年 11 月 1 日
重複データ	(株) 平和堂	2018 年 8 月 28 日
重複データ	(株) 第一生命保険	2018 年 12 月 18 日

4.3　考察

計 7 分野のうち，「病院」，「薬局」，「健康保険関連」分野では手作業取得データ数を上回るデータ数を本自動クローラーにより取得することができた. しかし，「その他」の分野において手動クローラーでは 169 社のデータを取得できたのに対し，自動クローラーでは 75 社のみしか取得できなかった．それは「その他」に分類されている企業を共通のターゲットとしての検索が困難であること，また本自動クローラーの検索システムとして使用している Google 社の CustomSearchAPI の仕様により，検索結果上位 100 サイトのみばかりしか取得出来ないことも大きな要因であることが考えられる．

さらに表 9 に自動クローラーのみ取得データ，重複データの例を示す．手動クローラーと自動クローラーによる取得データの大きな差は，手動クローラーと自動クローラーではデータの取得時期が異なるため，表 9 の例の様に自動クローラーにより新たに取得したデータの多くはサイト作成日に起因するものであると考える．

5　匿名加工情報公表サイトを作成している企業の割合調査

2 章の調査，4 章の実験により，2019 年 11 月現在 518 社の企業，団体が匿名加工情報公表サイトを作成していることが分かった．そこで，本章では日本の匿名加工情報公表サイトを作成している企業，団体についての調査をする．

5.1　実験方法

個人情報取り扱い事業者を，一般社団法人日本情報経済社会推進協会によるプライバシーマークを付与されている企業とする．

5.2　プライバシーマークとは

プライバシーマークとは，日本産業規格「JISQ15001 個人情報保護マネジメントシステム－要求事項」に適合して，個人情報について適切な保護措置を講ずる体制を整備している事業者等を評価して，その旨を示すプライバシーマークを付与し，事業活動に関してプライバシーマークの使用を認める制度である．2019 年 12 月現在プライバシーマークを付与されている事業者は 16,373 社である．

5.3　実験結果

2 章の調査，4 章の実験による，手動クローラー及び自動クローラーにより，2019 年 11 月現在 518 社の企業，団体が匿名加工情報公表サイトを作成していることが分かった. 実験結果を表 10 に示す．

表 10　匿名加工情報公表サイトを作成している企業の割合

日本の匿名加工情報公表サイト作成企業数	518 社
日本の個人情報を取り扱う企業数	16,372 社
匿名加工情報公表サイトを作成している企業割合	0.0316..

個人情報を取り扱う企業のうち約 3.2% の企業が匿名加工情報公表サイトを作成している．

6　おわりに

本研究では，4 章実験結果より，自動クローラーシステムにより計 321 社の匿名加工情報公表サイトを取得することができ，手動クローラによる取得数を上回る結果を示した．また，手動クローラーでは取得することのできなかった新たな 210 社のデータを取得できた．データ数だけではなく，手動クローラーでは，1 件のデータ取得にかかる平均時間は約 2 分 34 秒であるが，自動クローラーでは約 4 秒であった．この差も本自動クローラーシステムに大きな優位性があるといえる．検索ターゲット細分化や検索ワードの選定精度をより向上させる事で，有用性を向上する予定である．

今後は，本研究で得た匿名加工情報公表サイトのデータを基に，検索ターゲット細分化や検索キーワードの選定精度向上の為の分析を用いての新たな自動クローラーシステムの開発を試す必要がある．

参考文献

[1] 濱田，荒井，小栗，菊池，黒政, 中川, 西山, 波多野, 村上, 山岡, 山田, 渡辺’ 匿名加工再識別コンテストの設計 履歴データの一般化, 再識別’,PWS Cup 2018, pp935 - 940,2018.

[2] 小林祐貴’ 一般化匿名加工された購買履歴データの顧客・商品の RFM 分析’, 明治大学菊池研究室 2018 年度卒業論文,2018.

[3] GoogleCustomSearchAPI(https://developers.google.com/custom-search/v1/overview?hl=ja),2019 年 11 月参照

[4] 一般社団法人日本情報経済社会推進協会プライバシーマーク制度 (https://www.jipdec.or.jp),2019 年 12 月参照