明治大学総合数理学部

2019 年度

卒　業　研　究

匿名加工情報公表サイト調査 (2) 加工対象情報の抽出

学位請求者　先端メディアサイエンス学科

金子侑紀

目次

第 1 章　はじめに　2

第 2 章　抽出システムの開発　3
2.1　抽出システムの評価　5

第 3 章　考察　7

第 4 章　おわりに　8

参考文献　10

付録 A　AI スピーカーレプリカの Rasberry Pi 上への試験実装　11
A.1　はじめに　11
A.2　AI スピーカーレプリカの実装　11
A.3　実験　11
A.4　おわりに　14

参考文献　16

第 1 章

はじめに

個人情報取扱事業者が個人情報データベースを匿名加工し，作成した匿名加工情報を第三者へと提供する場合には，あらかじめ，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を公表するとともに，提供先に対し，匿名加工情報である旨を明示しなければならない (改正個人情報保護法第 37 条).

しかし，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は義務付けられているが，届出や申請の必要はない. 匿名加工情報の全貌を知るためには各社の公表ページから，個人に関する情報の項目を手作業で収集する他はない.

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，5 パターンの正規表現を用いて提供項目とその手法についての自動取得を試みる. 前者のクローリングについては [3] で報告し，本稿では，後者の自動取得について述べる．本論文では匿名加工取扱事業者のサイトを匿名加工情報公表サイトと定義する.

第 2 章

抽出システムの開発

抽出システムの構成を図 2.1 に示す．クローラーが取得した匿名加工情報公表サイトの HTML ファイルから提供する項目と手法を抽出する．抽出した結果は匿名加工情報公表企業及び項目一覧 DB として N 行 $\times 4$ 列 (企業名，個人に関する情報の項目，情報の提供手法，URL) の形式で表で管理する．N は企業数である．

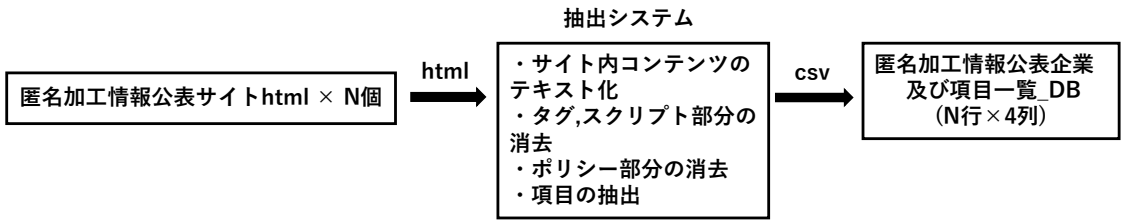

図 2.1　システム構成図

表 2.1　匿名加工情報公表サイト抽出結果

企業名	個人に関する情報の項目	提供手法
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

…<h2 class="mMH01">匿名加工情報の提供について</h2>
当組合では、保健事業や疫学調査等のために匿名加工情報を継続的に作成し、電子的な通信手段もしくはDVD等の物理媒体を用いてレセプト分析業者に提供します。作成及び提供する匿名加工情報に含まれる情報の項目は、性別、生年月、医療保険の資格情報（加入時期、脱退時期、本人・家族区分等）、診療報酬明細書の受診履歴、健診の受診履歴です。なお、個人を特定できる情報は含まれておりません。 …

入力

情報の項目
性別、生年月、医療保険の資格情報（加入時期、脱退時期、本人・家族区分等）、診療報酬明細書の受診履歴、健診の受診履歴

提供方法
電子的な通信手段もしくはDVD等の物理媒体

出力

図 2.2　抽出システム処理例

抽出システムの処理の例を図 2.2 に示す．入力文章にタグの処理とヘッダ，フッタなどの対象外の部分を取り除く処理を施した後，情報の項目，提供方法について正規表現による抽出をする．情報の項目を抽出する正規表現パターンは 5 種類用意する．正規表現の例を表 2.3 に示す．

2.1　抽出システムの評価

2.1.1　実験方法

2019 年 11 月 18 日に，クローラーを動作させて収集した 321 件のファイルに対して，本システムによる抽出を試みる．

表 2.2　抽出目標項目

項目	例
個人に関する情報の項目	性別，年代
提供方法	電子メールによる提供

2.1.2　実験結果

本システムにより，情報の項目と提供方法の抽出結果を表 2.4 に示す．取得したデータが過剰である例を図 2.3 に示す．

表 2.3　正規表現の例

正規表現	件数
(次のとおり｜下記｜以下)+([\s\S]+)\n[\s\S]+ 提供 [の]*方法	44
DPC+([\s\S]+)\n[\s\S]+ 提供 [の]*方法	14
情報 [の]*項目 ([\s\S]+)\n[\s\S]+ 提供 [の]*方法	4
項目は ([\s\S]+) です	3
([\s\S]+)(上記項目｜提供 [の]*方法｜第三者に提供)	60

表 2.4　匿名加工情報公表サイト抽出結果

適切に取得できた	取得したデータが過剰	取得できなかった	計
125	27	167	319

…サイト内検索(中略)取り組み
匿名加工情報の作成及び第三者提供について
匿名加工情報の取り組み
匿名加工情報に関する問合せ窓口
匿名加工情報の取り組み
DPC制度の導入の影響評価及び今後のDPC制度の見直しを図る目的で、厚生労働省が収集し管理する情報となるデータ（DPCデータ）を作成しております。また、審査支払機関への請求のため診療に係る費用を診療報酬明細書（レセプト）として作成しております。

DPCデータは、診療録からの情報および診療報酬明細書からの情報で構成されており、レセプトデータは、医療機関情報・保険者情報・診療行為情報・医薬品情報・特定器材情報等から構成されております。
DPCデータ並びにレセプトデータを利活用することで、医療の質向上および病院経営の改善に役立てる事が可能になるため、匿名加工後のデータを第三者へ提供しております。第三者提供するこれらのデータは氏名、住所、電話番号は含みません。なお、地域傾向や受診年齢層等を分析する必要があるため、郵便番号（上3桁のみ）、生年月日(生年月及び入院時年齢に変換を行い100歳以上は100歳に一括り)、各種保険証に関する情報については保険者番号（健康保険事業の各運営主体を指す番号）のみを含みます。

当院は上述の通り、診療情報から匿名加工情報を作成（毎月継続）し、第三者に提供しております。
匿名加工情報の提供の方法

図 2.3　取得したデータが過剰である例

第 3 章

考察

取得できなかった公表サイトの理由を表 3.1 に示す．

クローラー側から渡されたファイルはすべて HTML 形式だったが，PDF が多くあった．そこで，PDF 形式に変更して処理を行ったところ，75 件全ファイルのデータが空であり，取得できていなかった．

PDF 以外のファイルについて，文字コードが間違って保存されているものが多く存在した．Python の requests ライブラリは HTTP レスポンスヘッダの content-type が指定されていないファイルは defalt の ISO8859-1 形式で保存されるため，読み込めないすべてのファイルに対して強制的に変換を行っている．25 件中 2 件がサーバーエラー，23 件が ISO8859-1 形式への変換後も問題が解消されなかったファイルである．文字コードが原因のものは適切な変換手法が見つかれば項目を抽出できる可能性がある．

取得したデータが過剰であるファイルは，ヘッダやフッタ部分が特徴的で削除しきれなかったものや，匿名加工情報についての記載がプライバシーポリシーの一部として記載されていた．

抽出対象が含まれていないファイルとは法令で定める基準に従う旨が記載されており，個人に関する情報の項目や提供方法が明記されていないファイルである．

表 3.1　取得に失敗した公表サイトの内訳

理由	件名	例
PDF の破損	75	抽出不可
サーバーエラー・ファイル破損	25	500 等，ページ取得に失敗
提供していないと明記しているサイト	5	現在、匿名加工情報は取扱っておりません。
公表サイトでない	3	匿名加工情報作成ツールの記事
項目と方法は別ファイルに記載	11	弊社が作成する匿名加工情報に含まれる個人に関する情報の項目、弊社が第三者に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は、こちらをご覧ください。
抽出対象が含まれていない	42	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。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。
抽出失敗	8	抽出不可

第 4 章

おわりに

本研究では匿名加工情報取扱事業者の公示ページに含まれる情報の自動抽出を実現した．匿名加工情報取扱事業者の公示ページにはパターンがあり，正規表現を使用することで 319 件中 125 件の公表サイトから情報を抽出した．

今後は，クローラー側と協力して取得するファイルを増やすとともに，より高度な抽出手法を検討する必要がある．

謝辞

本研究を行うにあたり，多くの方より御指導いただきました．特に，多大なる御指導を受け賜りました，明治大学総合数理学部先端メディアサイエンス学科，菊池浩明教授に深く感謝申し上げます．共に研究に取り組んでくださった小野くん，予備実験等に協力してくださった菊池研究室の皆様並びに先端メディアサイエンス学科の方々に深く感謝の意を表するとともに，謝辞とさせていただきます．

参考文献

[1] 濱田，他，“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”，CSS 2018，pp935-940，2018 情報処理学会 2018.

[2] 匿名加工情報 - 個人情報保護委員会，(https://www.ppc.go.jp/personalinfo/tokumeikakouInfo/，2019 年 12 月 11 日参照).

[3] 小野敦樹，“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”，明治大学菊池研究室 2019 年度卒業論文，2019.

付録 A

AI スピーカーレプリカの Rasberry Pi 上への試験実装

A.1　はじめに

Amazon Echo，Google Home，LINE/Clova など AI スピーカーが普及しつつある．一方で，物理的な障害による音声データの流出や超音波スピーカーによる操作 [1] などの危険性が叫ばれている．改造などのハードウェア上の問題 や，マルウェアの感染などによるソフトウェア上の問題による攻撃はあまり検討されていない．

そこで，本研究では，AI スピーカーへのマルウェアの感染を仮定し，市販されている AI スピーカーの通信を観測し，AI スピーカーの状態別の通信量を調査する．Raspberry Pi 3 B+(以下 Raspberry Pi) に Google Assistant SDK(以下 SDK)[2] をインストールし，Google Home の動作を模倣する AI スピーカーを作成し，研究室での AI スピーカー利用のログを取得する．

A.2　AI スピーカーレプリカの実装

Raspberry Pi に SDK をインストールし，Google Home と互換の AI スピーカーを作成する．これを AI スピーカーレプリカと呼ぶ．SDK はマイクとスピーカーがある端末に Google アシスタントを組み込むことができる．ユーザが Google Home を使用するための「OK，Google」(以下ホットワード) の呼びかけを行うと SDK が Google Home 用サーバとの通信を行う．SDK に含まれる hotword.py は AI スピーカーレプリカの動作を記述することができる．hotword.py を編集し，ホットワード発話時に内容を json 形式で外部ファイルに保存している．

マイクはサンワサプライ社の MM-MC23 を用いる．システム構成図，システム外観を図 2.1，図**??**に示す．

A.3　実験

A.3.1　実験目的

1. AI スピーカーの状態を 3 つに分け，各状態での通信量を調査する．(実験 1)
2. AI スピーカーレプリカのログから研究室内での AI スピーカーの使われ方を調査する．(実験 2)

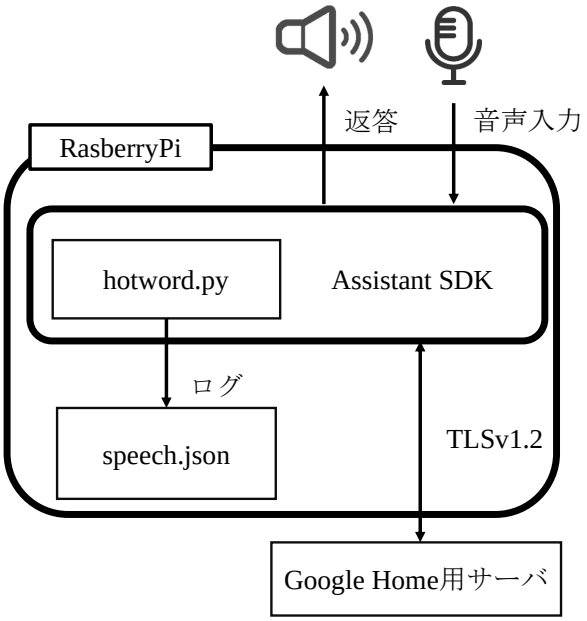

図 A.1　システム構成図

図 A.2　システム外観

A.3.2　実験内容

実験 1

ミュート状態の当否とホットワード発話の有無によって，ミュート時，ホットワード発話時，ホットワード非発話時の 3 状態に区分した．ミュート状態でホットワードの呼びかけが行われないことをミュート時，ミュート状態が解除されホットワードの呼びかけが行われないことをホットワード非発話時，ミュート状態が解除されホットワードの呼びかけが行われる状態をホットワード発話時とする．

リピータハブとミラーリングの設定を行ったルータを用いてそれぞれの状態におけるトラフィック観測を行

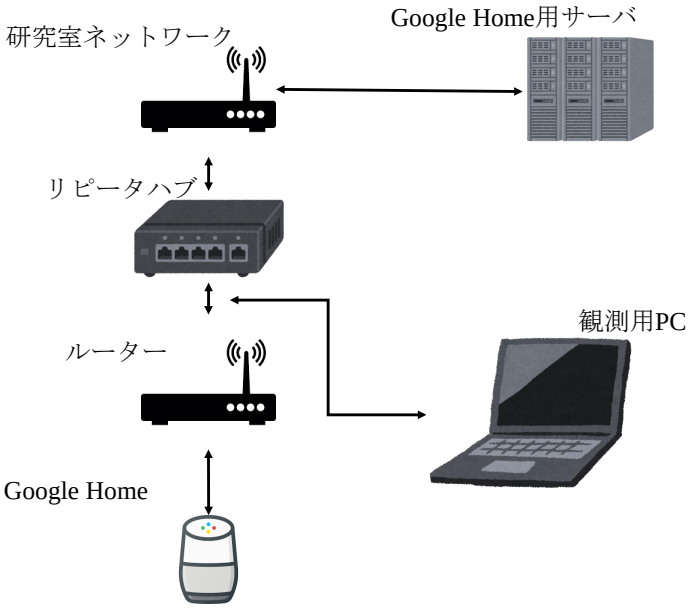

図 A.3　ネットワーク構成図

う．実験日時は 2018 年 7 月 11 日 13 時 21 分から 15 時 5 分まで，実験場所は本学 1005 実験室である．ネットワーク構成図を図 A.3 に示す．観測用 PC 上で WireShark を実行し，各状態の 5 分間のトラフィックを取得した．ホットワード発話時は平均 20 秒間隔で Google Home に「今何時」の呼びかけを行う．

実験 2

実装したシステムを用いて運用実験を行う．AI スピーカーレプリカを設置し，利用ログを取得する．実験期間は 2018 年 11 月 15 日から 12 月 18 日まで，実験場所は本学 1005 実験室である．

A.3.3　実験結果と考察

実験 1

図 A.4 にミュート時のトラフィックとホットワード非発話時の累積トラフィックを示す．x 軸は実験開始時刻からの経過時間，y 軸は 1 秒あたりの累積通信量とした．ミュート時 (図 A.4 : 点線) とホットワード非発話時 (図 A.4:実線) を比較すると，ホットワード非発話時のトラフィックが多いことがわかる．ミュート時が最大で 6976bps，ホットワード非発話時が最大で 70488bps であった．ここから，ホットワード発話が行われているか判断するためのデータが送信されていると推測する．

実験 2

実験期間中，ホットワード発話は 107 回行われた．発話の時間分布を図 A.5 に示す．発話内容に対して MeCab を用いて単語の出現頻度分析を行った．出現頻度の高い単語を表 A.1 に示す．研究室に人が集まる 12 時と 20 時にホットワード発話が多い．出現頻度の高い単語の内，タイマーはカップ麺を食べる際に使われ

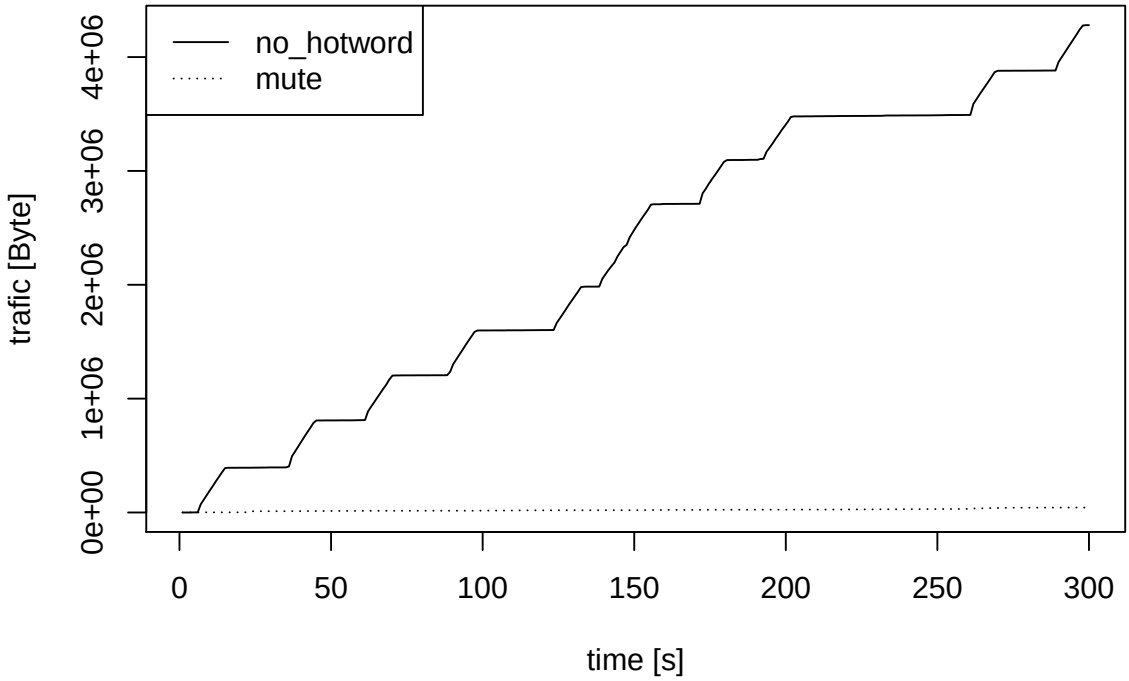

図 A.4　ミュート時及びホットワード非発話時のトラフィック累積

表 A.1　出現頻度の高い単語

単語	出現回数
天気	30
今日	24
明日	13
タイマー	12
何	9
分	8

ている．天気は性能を調べる目的で発話されることがあり，頻度が高くなったと思われる．

A.4　おわりに

GoogleHome のトラフィックについて，ホットワード発話時はホットワード非発話時と比較して約 10 倍のパケットが流れていることが示された．本研究では発話ログの取得を実現した．本実験で個人情報を含むような発話は行われなかった．研究室という共有空間において消極的になっていると予測される．サービスの充実によって個人情報を含む発話が増えれば，予定や連絡先情報など，発話ログの収集から漏洩する恐れがある．

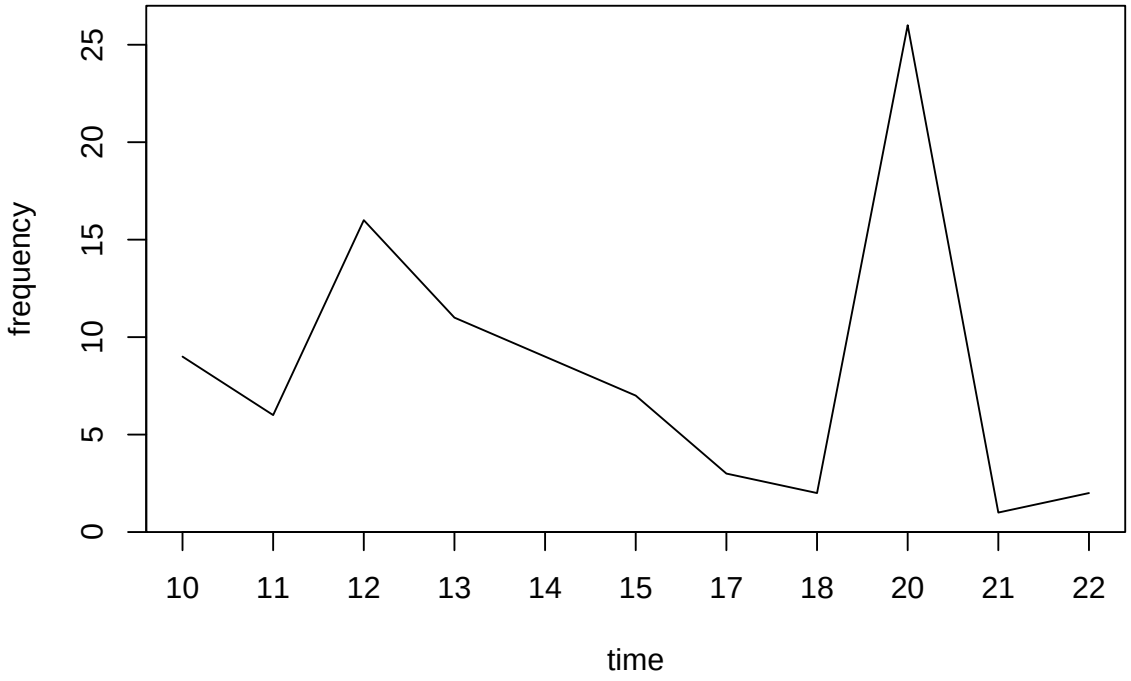

図 A.5　時間分布

参考文献

[1] 飯島 涼， 南 翔汰， シュウ インゴウ， 及川 靖広， 森 達哉， “指向性スピーカを用いた音声認識装置への攻撃と評価”， Symposium on Cryptography and Information Security (SCIS 2018) ， pp. 1-8， Jan，2018.

[2] Google Assistant SDK for devices， (`https://developers.google.com/assistant/sdk/`， 2018 年 11 月 4 日参照).