明治大学総合数理学部

2020 年度

卒 業 研 究

国内の匿名加工情報を一覧する利活用ポータルサイト「匿名加工情報目録」の開発

学位請求者 先端メディアサイエンス学科

梶間大地

目次

第 1 章

はじめに

1.1　研究背景

自社の有する顧客情報などの個人情報を匿名加工してビックデータビジネスに利活用する動きが盛んである [1]．その一方，取扱事業者が匿名加工情報を作成する際には届出は不要で，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に過ぎないため，利活用者が望む匿名加工情報を探し求めることが困難であった．

そこで金子らは GoogleSearchAPI と正規表現を利用して，匿名加工情報を公表している事業者を収集するクローラシステムの開発を行った [2]．開発したクローラシステムを利用して 321 件の事業者を収集した．また藤田らは匿名加工情報のカタログを開発している [3]．しかし，[2][3] では事業者ごとに公表形式がまちまちであり，業種ごとの取得項目や提供方法のトレンドを知ることはできなかった．

そこで，本研究では，ウェブで公開されている匿名加工情報のページを収集してデータベースに格納して，検索などのサービスを実現したポータルサイトの開発をする．前述の問題点に対して，SQL のあいまい検索などを工夫した点に新規性がある．本稿では，開発サイトの設計と実現した機能について報告する．

第 2 章

ポータルサイトの開発

2.1　概要

図 2.1 にシステム構成図，表 2.1 にポータルサイトに実装されている機能を示す．

ポータルサイトの開発には PHP ver 5.3.3 と MySQL ver 5.1.73 を用いた．ポータルサイトは情報一覧ページ，情報を追加するページ，情報を集計した統計情報ページの 3 つで構成されている．

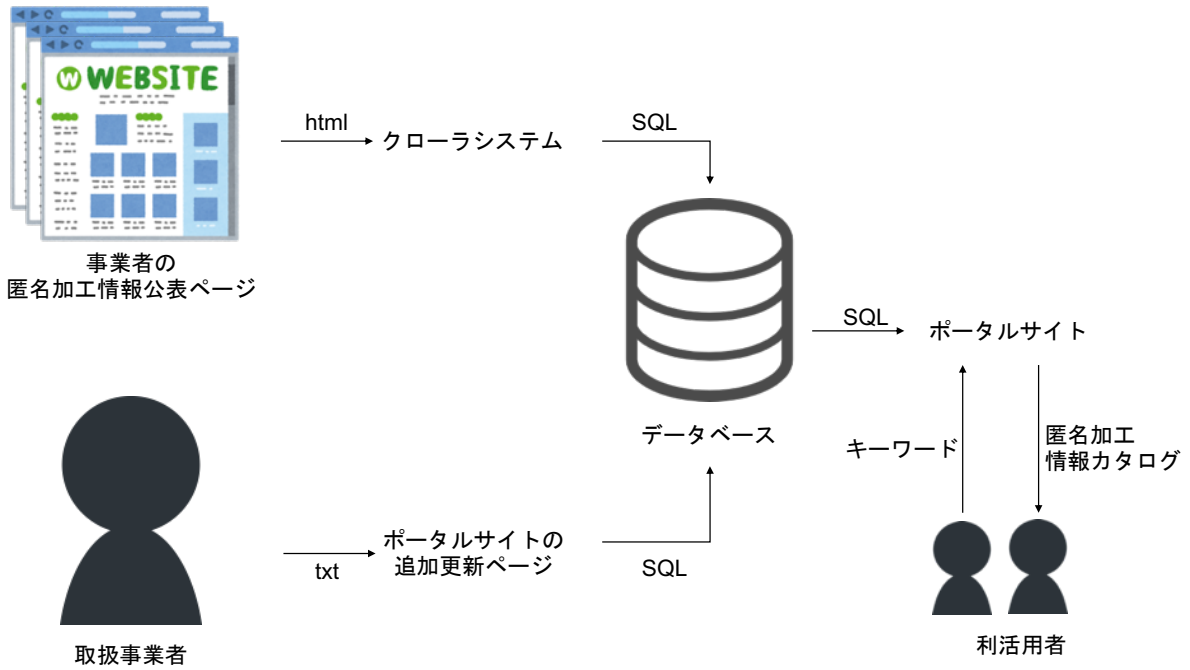


図 2.1　システム構成図

表 2.1　ポータルサイトの機能

機能	概要
情報一覧	登録されている企業名，業種，公表されている URL を示す． 詳細から提供項目，提供方法を確認する．
追加	新たに情報を追加する．
検索	企業名，業種，提供項目，提供方法を選択して登録されている情報に検索をかける．
編集	既に登録されている情報を編集する

2.2　データベース

表 2.2 にデータベースの設計，図 2.2 に登録されている情報の一例を示す．公表している取得項目，提供方法は区切り文字を設定することで複数の情報をひとつのカラムに保存している．

表 2.2　データベースの設計

カラム名	概要
id	各データの固有番号
url	事業者の匿名加工情報公表ページの URL
conpany	事業者名
industry	事業者の業種
item	公表している取得項目
method	公表している提供方法
hash	パスワードをハッシュ化したもの

図 2.2　データベースに登録されている情報

2.3　匿名加工情報一覧ページ

図 2.3 に開発したポータルサイト*の情報一覧ページを示す．ここでは，匿名加工情報公表ページへのリンク，事業者の業種と詳細ページへのリンクを表示している．詳細ページから匿名加工情報の項目と提供方法を確認することができる．

図 2.3　匿名加工情報一覧ページ

2.4　情報追加更新ページ

図 2.4 に開発したポータルサイトの情報追加ページを示す．情報を追加するページでは，誰でも匿名加工情報を提供している企業の情報を追加することができる．ポータルサイトの利用者の登録は不要とし，登録する際には登録データに対してパスワードを設定している．また，エージェントによる「荒らし」等への対策するため，Google 社の提供する reCAPTCHA[4] を導入している．

*https://windy.mind.meiji.ac.jp/~kajima/2020

図 2.4　情報追加ページ

2.4.1　ユースケース「統計情報の集計」

統計情報ページでは，任意のキーワードを与えてデータベースに格納されている業種ごとの事業者一覧と件数を提供する．本機能を用いてポータルサイト内に格納されている取得項目と提供方法を集計した結果を表 2.3 に示す．業種分類は，日本標準産業分類を基にした独自分類 [2] を採用した．登録されている情報は [2] で収集されたものである．

表 2.3　業種別の全てのデータ数と特定のキーワードを含む数

業種	数†	氏名	健診	メール	サーバ	外部媒体	セキュリティ
医療業 (病院)	89 (6)	1	0	0	64	62	3
小売業	11 (5)	1	0	2	3	2	1
小売業 (薬局)	56	25	0	2	4	1	47
複合サービス業	50 (43)	1	0	1	1	0	0
情報サービス業	31 (8)	2	3	5	6	3	5
健康保険組合	19 (1)	0	12	0	0	1	3
保険業	15 (9)	0	2	2	0	4	1
金融業	15 (5)	0	0	0	2	0	3
情報通信業	9 (6)	0	0	1	1	1	0
サービス業	9 (4)	0	1	1	1	1	1
医療業 (製薬)	7(1)	0	1	2	4	1	2
製造業	6 (2)	0	1	0	2	1	1
運輸業	5 (4)	0	0	0	1	1	0
年金保険センター	3	0	0	0	1	1	0
電気・ガス・熱供給・水道業	3 (3)	0	0	0	0	0	0
不動産業	2	0	0	0	1	1	0
健康保険協会	2	0	1	0	0	0	0
教育，学習支援業	2	0	0	0	2	1	1
映像・音声・文字情報制作業	2	0	0	2	1	0	0
化学工業	2	0	0	0	0	0	0
社会保険，社会福祉，介護事業	5 (1)	0	0	0	2	1	0
一般社団法人	1	0	0	0	0	0	1
未分類	18 (8)	0	2	1	2	2	4
合計	378	28	23	19	98	84	73

() 内は取得項目，提供方法のどちらも欠損しているデータ数

第 3 章

おわりに

本研究では，国内の匿名加工情報取扱事業者の公表情報を収集して，検索などの機能を加えたポータルサイトを開発した．データの追加，更新等を適時行うことを今後の課題とする．

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．また，研究に使用するデータを提供してくださった小野さん，金子さん，研究室の皆さん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] 国内ビッグデータ活用事例 - 総務省 (https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc254330.html,2020 年 12 月参照)

[2] 金子侑紀, 小野敦樹, 伊藤聡志, 菊池浩明, 服部充洋, 飯田泰興, 藤田真浩, 山中忠和, “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”, 情報処理学会第 82 回全国大会, pp.3_447-3_448, 2019.

[3] 藤田真浩, 飯田泰興, 服部充洋, 山中忠和, 松田規, 伊藤聡志, 菊池浩明, ”匿名加工情報取扱事業者による公表情報を利用した匿名加工カタログの提案と実装”, コンピュータセキュリティシンポジウム, 2020.

[4] reCAPTCHA(https://www.google.com/recaptcha/about/,2020 年 12 月参照).

[5] 匿名加工情報 - 個人情報保護委員会 (https://www.ppc.go.jp/personalinfo/tokumeikakouInfo/,2020 年 12 月参照).

付録 A

Tor ネットワークのクローラシステム (2) CAPTCHA 自動解析

A.1　はじめに

近年，セキュリティへの意識から匿名通試飲システム Tor(The Onion Router) の利用が増えている． Tor は匿名での通信が可能なため違法商品販売等の不正な目的で利用されている場合もある．そこ で，ダークウェブ上の違法商品販売サイトの統計調査に注目が集まっている．しかしダークウェブ上 には CAPTCHA を採用ている違法商品販売サイトが多く存在するため，先行研究では，CAPTCHA を採用し ている違法商品販売サイトには効率的な調査を行うことができなかった [**?**].

そこで本研究では，Tor ネットワーク内に存在する違法商品販売サイトの調査を行う．本稿は，OCR を 用いた CAPTCHA 画像の自動解析システムの開発について報告し，システム全体と人間が CAPTCHA を解く操作のみを行う半自動クローラについては [**?**] で報告する．ただし，本研究で扱う CAPTCHA の種類はテキストベースのものに限る．

A.2　全自動クローラシステムの開発

本研究では，Tor ネットワークに Tor ネットワークにアクセスするために，Tor をインストールした Linux サーバと Python を用いた．図**??**に本研究で開発したクローラシステムの構成図を示す．

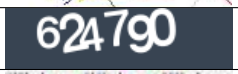
Python プログラムに対して Tor ネットワーク内のクローリングしたいサイトの URL を送る．URL を受け取っ た Python プログラムは socks5，Tor を利用してサイトにアクセスする．アクセスした際に表示された CAPTCHA 画像 Q をサーバ内に保存し，セッションの Cookie 情報を取得する．次にサーバに保存された Q を Python プログラムを用いて画像加工を行い，加工した Q をサーバに保存する．加工した Q を Tesseract OCR を利用して文字認識を行い，帰ってきた文字列 A を Python プログラムに渡す．その後，サイトにユ ーザ名，パスワードのログイン情報，保存した Cookie，A を利用してログインを試みる．サイトのユー ザ用ページにアクセスできれば終了．A が誤りの場合，プログラムに URL を送った時点からやり直す． これをユーザ用のページにアクセスできるまで繰り返す．

A.3　CAPTCHA の機械解読

A.3.1　CAPTCHA の成功率

表**??**に CAPTCHA の解決の成功率を示す.

表 A.1　CAPTCHA 解読の成功率

URL	主要カテゴリ	全体	正答	誤答	認識なし	閾値	CAPTCHA 画像の例
apollonujscjrlng.onion	Credit Card	1000	0	722	278	120	
deluxedzn6h572qp.onion	内容不明	1000	17	983	0	120	
27kaqicipyhous2p.onion	Credit Card	1000	653	347	0	なし	
abyssopyps3z4xof.onion	Drugs	1000	0	467	533	120	
yohoho7do5u6d35g.onion	Drugs	1000	303	697	0	なし	
torcvvq44o7ofjuu.onion	Credit Card	1000	596	331	73	120	

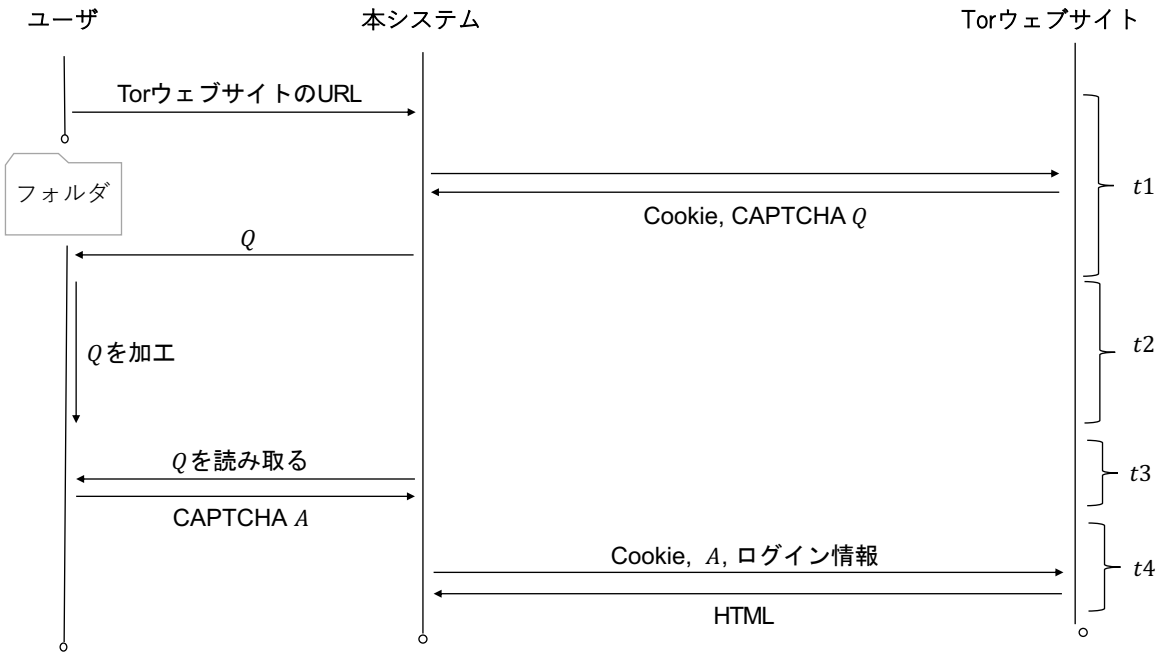


図 A.1　システム構成図

表 A.2　HTML 取得にかかった時間

	平均 [秒]	標準偏差 [秒]
$t1$	2.717	0.286
$t2$	0.013	0.002
$t3$	0.195	0.030
$t4$	2.444	0.187
全体	7.237	3.483

A.3.2　CAPTCHA 画像の加工手法

CAPTCHA 画像の解析には openCV の Python 用ライブラリを用いて画像加工を行い，google 社が開発を行った OCR(光学的文字認識) ツールである Tesseract OCR[**?**] を用いて加工後の画像の文字 認識を行った．画像の加工処理では，取得した CAPTCHA 画像を 2 値化処理を行ったのち，閾値処理によっ てノイズの除去を行った．また，CAPTCHA 画像にノイズのないものには閾値処理を行わなず 2 値化処理の みとした．

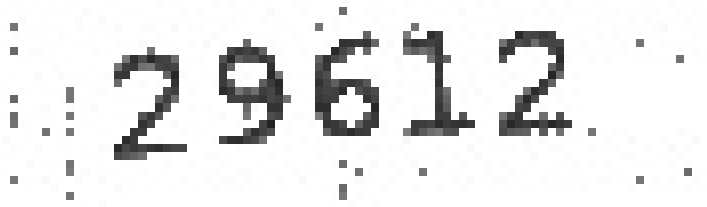


図 A.2　加工後の CAPTCHA 画像の例

A.3.3　処理時間の評価

表**??**に HTML 取得にかかった時間を計測した結果を示す．これは torcvvq44o7ofjuu.onion に合計 150 回アクセスした際のデータである．

A.3.4　考察

自動解読できなかった CAPTCHA 画像については以下のような原因が考えられる．

1. 図**??**のように読み込む文字とノイズの色が似ていたりノイズが文字の上に 存在しているため，閾値処理を行った際に読み解く文字まで欠けたり削除されてしまう．
2. 読み解く文字の歪みや重なりによって，文字認識に失敗する

A.4　おわりに

本研究では，Tor ネットワーク内に存在する違法商品販売サイトの情報を効率的に収集するための全自動クローラシステムの開発を行った．CAPTCHA 画像については平均 26.15% で解決できることがわかった．

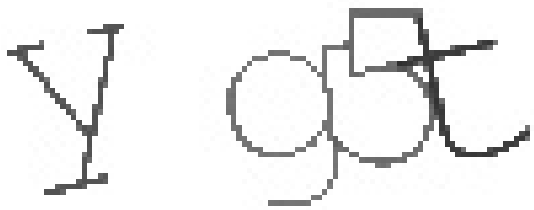


図 A.3　考察 1 の例

しかし，汎用的に用いることは出来ないこともわかった．今後は CAPTCHA 画像を加工すること以外での CAPTCHA の自動解決の方法を考案することを課題とする．

参考文献

[1] 鳥居洸希，菊池浩明．“Tor ネットワークのクローラシステムの開発と違法商品販売サイトの調査”，情報処理学会第 81 回 pp.3_435-3_436，2019．

[2] 鳥居洸希，菊池浩明．“Tor ネットワークのクローラシステム (1) 違法商品販売の調査”，情報処理学会第 82 回発表予定．

[3] Tesseract OCR，https://opensource.google/projects/tesseract