Tor ネットワークのクローラシステム (2) CAPTCHA 自動解析

梶間 大地 †　　鳥居 洸希 †　　菊池 浩明 †

明治大学総合数理学部 †

1　はじめに

近年，セキュリティへの意識から匿名通信システム Tor(The Onion Routering) 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多く，これらの違法商品販売サイトの調査を試みている． しかしダークウェブの違法商品販売サイトの多くは CAPTCHA を採用しているため機械的な調査を行うことができなかった [1].

そこで本研究では，OCR を用いて Tor ネットワーク内に存在する違法商品販売サイトを調査を行う．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する．システム全体と人間が CAPTCHA と解く操作のみを行う半自動クローラについては [2] 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．

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 1 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を行いたいサイトの 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．

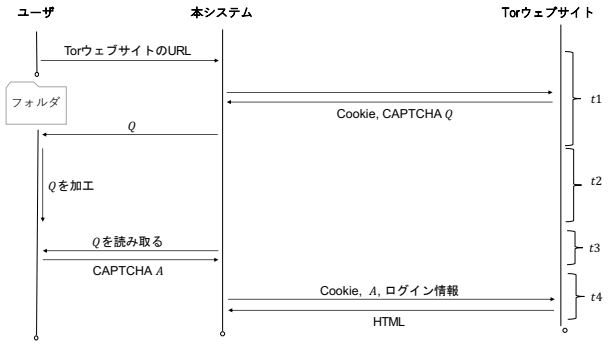


図 1　システム構成図

3　CAPTCHA の機械解読

3.1　CAPTCHA の成功率

表 1 に CAPTCHA の解決の成功率を示す．

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[3] 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 2　加工後の CAPTCHA 画像の例

Develop of Tor crawler system(2) – Automated image CAPTCHA responder

†Daichi Kajima, Hiroki Torii, Hiroaki Kikuchi, School of Interdisciplinary Mathematical Science, Meiji University.

表 1　CAPTCHA 解読の成功率

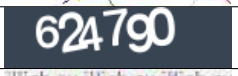
URL	主要カテゴリ	全体	正答	誤答	認識なし	閾値	CAPTCHA 画像の例
apollonujscjrlng.onion	Credit Card	1000	0	722	278	120	
deluxedzn6h572qp.onion	内容不明	1000	17	983	0	120	
27kaqicipyhous2p.onion	Credit Card	1000	653	347	0	なし	
abyssopyps3z4xof.onion	Drugs	1000	0	467	533	120	
yohoho7do5u6d35g.onion	Drugs	1000	303	697	0	なし	
torcvvq44o7ofjuu.onion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

表 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

3.3　処理時間の評価

表 2 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．

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 3 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．

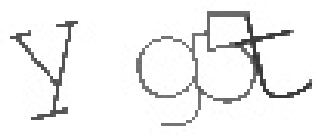


図 3　考察 1 の例

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

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．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 回全国大会 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 (2019 年 12 月参照)