4X-05

クラウド型CAPTCHAに対するbotによるアクセスの分析

寺田 健太 *　　岡部 寿男 †　　松本 悦宜 ‡
京都大学　　Capy 株式会社

1　はじめに

近年、Webサイトに対するbotによる不正アクセスが増加傾向にある。botによる不正アクセスの手口の典型として、リスト型攻撃によるパスワードクラッキングが挙げられる。また、不正アクセスとは言えないものの、転売目的でbotを用いてチケット買占めを行うような悪質なアクセスも見られるようになってきている。

botによる自動アクセスを防ぐ手段にCAPTCHA（Completely Automated Public Turing test to tell Computers and Humans Apart）がある。これはアクセスを行っているのが人なのかbotなのかを判別するためのテストである。しかし、botの高度化に伴って、CAPTCHAの複雑化が必要となり、それに伴ってユーザの利便性が低下する問題がある。

そのため、リスクベース認証の考え方で、アクセス元のIPアドレスやブラウザの種類等の情報を基にbotであることが疑われるアクセスに対しては難しいCAPTCHAを課すことが行われている。[1]

本研究では、Capy株式会社が提供するクラウド型CAPTCHAサービスであるパズルCAPTCHAへの実際のアクセスを分析し、リスクベース認証に有効な特徴量を検討した。ある日本国内向けのWebサイトに対するアクセスログについて、そのうち実際のインシデントからbotであると判断されたアクセスを分析したところ、海外のISPからのアクセスで、かつブラウザが送出するHTTP Accept-Languageが中国語であるという特徴が見つかり、それを手掛かりに他のアクセスについても精査したところ同様に疑わしいアクセスを多く見つけることができた。

2　Capy パズル CAPTCHA

Capy社のパズルCAPTCHAはクラウド型のサービスでCAPTCHAに関するシステムはWebサイトとは独立のクラウド上のサーバに存在する。実装はWebサイトにCAPTCHAに関するスクリプトを追加することで行われる。

パズルCAPTCHAにおいてユーザによる通信はCAPTCHAに関するスクリプトの含まれるページのリクエストを行うget-js、パズルCAPTCHAで用いる画像のリクエストを行うget-image、パズルCAPTCHAの回答のポストを行うverifyの3手順からなり、Webサイトや CAPTCHA 生成/照合サーバと通信を行う。[2]

3　データセット

今回分析に用いたデータセットは、ある顧客のサービス上におけるパズルCAPTCHAへのアクセスログ622077件にbotによるアクセスかどうかを判断したフラグを付与したものである。この判断はCapy社が通常取得しているログに加えて、導入元自身による調査も含め、Capy社により行われている。なおアクセス期間は2019年3月1日9時0分0秒から2019年4月1日8時59分59秒までである。このサービスは日本国内向けのもので、botによる不正なアクセスによって金銭的被害が想定されるものである。

データセットに含まれる情報のうち、本研究ではアクセス時刻、アクセス元のIPアドレス、HTTP User-Agent、HTTP Accept-Language、パズル認証の成否、パズル認証の回答時間、ならびにis_botフラグを用いた。is_botフラグは先述したbotによるアクセスであると判断されたアクセスログに立てられるフラグである。

またアクセス件数やそのグラフについてはget-imageのアクセスのみをカウントすることにする。

4　データの分析

4.1　Bot と判断されたアクセスの分析

まずis_botフラグがTrueのアクセスログ（426件）について、HTTP Accept-Languageを分析したところ、HTTP Accept-Languageが“zh-CN,zh:q=0.9”または“zh-CN,zh:q=0.8”のいずれかに限られることが判明した“zh”は中国語、“zh-CN”は中国語（中国本土）を表す言語コードである。またIPアドレスについてwhoisサービスで調べた結果、香港にある特定のISPに割り当てられているアドレスからのアクセスであることがわかった。これは国内向けの日本語でのサービスへのアク

Analyzing bot access to cloud-based CAPTCHA
* KENTA TERADA, Kyoto University
† YASUO OKABE, Kyoto University
‡ YOSHINORI MATSUMOTO, Capy Japan Inc.

セスとしては極めて不審なアクセスであると言える。

また、これらのアクセスの時期は、分析の対象としている期間において、全体のアクセスの分布と比較すると、3月下旬に集中していることが言える（図1、図2のbot）。

一方、それ以外の情報については顕著な特徴は見出せなかった。

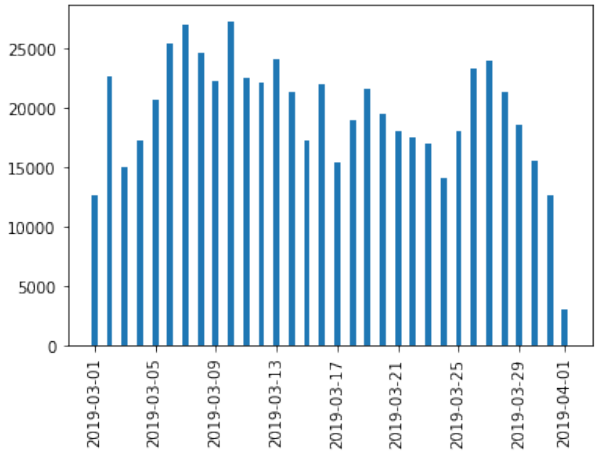

図 1: 全体のアクセスの分布

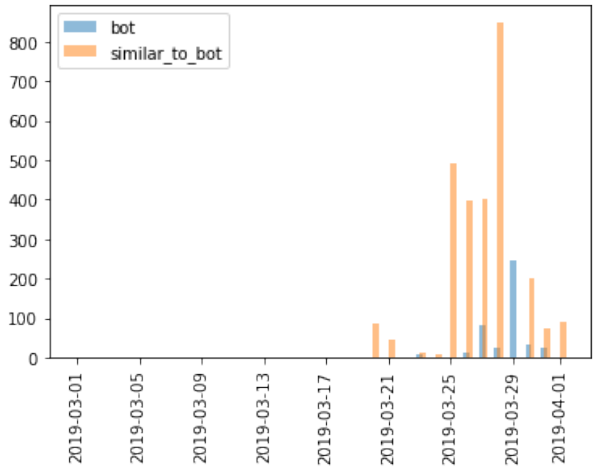

図 2: bot とそれに類似したアクセスの分布

4.2 Bot に類似したアクセスの分析

次に is_bot フラグが False であるアクセスのうち、HTTP Accept-Language が is_bot フラグが True で見られたものと同じであるアクセス（7693 件）の IP アドレスを whois サービスで調べた。その結果、is_bot フラグが True のものと同一の ISP から割り当てられている IP アドレスが見つかった。それらの IP アドレスと一致するアクセスを is_bot フラグが False のものから抽出したところ 2649 件見つかり、HTTP Accept-Language は is_bot フラグが True のアクセスと同様、すべて “zh-CN,zh:q=0.9” または “zh-CN,zh:q=0.8” のいずれかに限られていた。これらは is_bot フラグが True のアクセスと同程度に bot によるアクセスであることが疑われ、かつ同一の攻撃者グループによる同一の ISP からの攻撃である可能性が高い。またアクセス時期は is_bot が True であるアクセスに少し先立つ形で分布しており（図2の similar_to_bot）、ログ調査により bot と判断されたアクセスよりも前から攻撃が始まっていたと推定されている。

さらに、HTTP Accept-Language が、is_bot が True のものと同じで IP アドレスが上記の ISP 以外の範囲にあるアクセス（5044 件）についての分析を行ったところ、その大半（4899 件）が中国にある特定の ISP に割り当てられたアドレスからのアクセスであり、それ以外は香港の ISP から 122 件、国内の ISP のアドレスから 23 件と比較的少なかった。

5 まとめ

本研究では、パズル CAPTCHA への bot によるアクセスの実際のログを分析し、Web サイトの対象の国や言語と、アクセス元の IP アドレスが示す地域ならびにブラウザの受け入れる言語の不一致が bot の判別に有効であることを示した。また、それを用いてさらに bot によるアクセスと思われる他のアクセスを発見することができた。

今後は更に多言語かつ全世界向けの Web サイトについて同様の分析を進める予定である。

参考文献

[1] S. Wiefling, L. Lo Iacono, M. Dürmuth, “Is This Really You? An Empirical Study on Risk-Based Authentication Applied in the Wild,” In: Dhillon G., Karlsson F., Hedström K., Zúquete A. (eds) ICT Systems Security and Privacy Protection. SEC 2019. IFIP Advances in Information and Communication Technology, vol 562. Springer, Cham. https://doi.org/10.1007/978-3-030-22312-0_10

[2] T. Arai, Y. Okabe, Y. Matsumoto and K. Kawamura, “Detection of Bots in CAPTCHA as a Cloud Service Utilizing Machine Learning,” 2020 International Conference on Information Networking (ICOIN), Barcelona, Spain, 2020, pp. 584-589, doi: 10.1109/ICOIN48656.2020.9016522.