ボットアクセスの観測と検証

Observed Bot's Accesses and the Verification

藤田 吾矢・ネットワーク分科会・情報セキュリティ大学院大学

Many bots are crawling on websites of the Internet.
They are crawling in order to collect information of webpages all over the world automatically and they are running all day.
Some are search engine's robots, others are bots for an archive like *Internet Archive*.
Therefore, we observed accesses at the website of Okubo Lab, Institute of Information Security for 2 month.
In this study, we discuss its access analyses and consider current bot's crawling.

1. 研究の背景と目的

インターネット上にはWebサイトの情報収集を目的に、多くのボットが巡回をしている。検索エンジンへのインデックス化を目的としたrobotや、Webサイトのアーカイブ化を行うInternet Archiveをはじめ、多種多様である。

今回は、2016年12月, 2017年1月の2ヶ月間にかけて、情報セキュリティ大学院大学の大久保研究室のWebサイトを活用し、観測されたアクセス記録から、ボットのクローリング動向を調査し、Webサイトへの攻撃手法について検証を行った。本稿で、「アクセス数」とはindex.php へのページビューを指すことにする。

表1 2ヶ月間で観測したボットとそのアクセス回数
(index.php における観測、UserAgentの値・ホスト名から解析)

ボット名	アクセス回数	ボット名	アクセス回数
Hatena Antenna/0.5	253	MJ12bot/ v1.4.7	3
bingbot/2.0	120	Googlebot/2.1 (*)	2
Googlebot/2.1	51	DeuSu/5.0.2	2
ltx71(amazonaws.com)	17	Microsoft Office Protocol Discovery	2
Yahoo! Slurp	15	Daum 4.1	1
360Spider(*)	8	SMTBot/1.0	1
DotBot/1.1	7	ichiro/3.0	1
Java/*	5	AhrefsBot/5.2	1
SemrushBot/ 1.*~bl	3	AppEngine-Google	1
ndl-japan-warp/0.1	3		

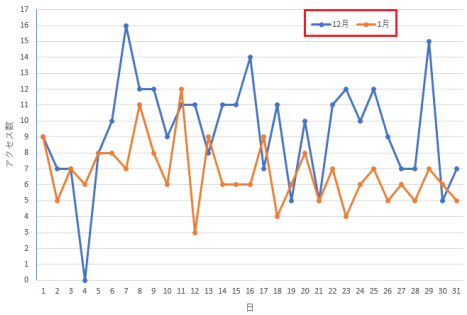

図2 表1で示したボットの日計アクセス回数 (12/4は観測ミス)

2. ボットの観測結果

Webページはインターネット上に公開後、直ちにボットによる監視を受ける。今回、観測したボットによるアクセス数の集計結果を表3に示す。2016年12月, 2017年1月の合計アクセス数のうち、それぞれ約61.0%, 約56.1%がボットによるアクセスであったことが確認された。

ただし、観測を行ったindex.php において、.htaccess, robot.txt, METAタグ等を用いたボットの巡回拒否は行っていない。インターネット上の情報は、あらゆるアプリケーションから収集されており、一度公開を行った情報の完全な削除は極めて困難であると再認識した。

表3 2016年12月, 2017年1月に観測したアクセス数の集計

	2016年12月	2017年1月
一般アクセス数	185	162
ボットアクセス数	289	207
合計アクセス数	**474**	**369**

3. ボットによる情報収集の効果

ボットによるクローリングで、インターネット上に公開されている情報を短時間で発見することが可能になった。2ヶ月間の観測であらゆるボットを確認できたが、世界中のアプリケーションにより情報が収集され、ビッグデータが形成されている。

一方、情報の削除が困難となり、あらゆる論争も呼んでいる。2014年には、Googleによる検索で、自身の逮捕歴に関する報道がヒットすることはプライバシーの侵害であるとして訴えが起こされている。2018年1月31日、最高裁判所は検索結果の削除を認めない決定を下しているが、「忘れられる権利」を主張する動きもみられる。

4. ボットのクローリング拒否

Webページへのクローリングを拒否する手法として3点、Webアプリケーションベースで優先度の高い順に挙げる。いずれもUserAgentベースに指定を行う。ボット名は、ワイルドカードによる指定も可能である。

[1] .htaccess を用いた方法 (設置したディレクトリ以下の階層に有効)

例) SetEnvIf User-Agent "Googlebot" shutout # Googlebotの巡回を拒否
allow from all

[2] robot.txt を用いた方法 (設置したディレクトリ以下の階層に有効)

例) User-agent: ia_archiver # Internet Archiveの巡回を拒否
Disallow: /

[3] METAタグ を用いた方法 (設置したWebページベースに有効)

例) <!-- robotに対して、インデックス作成・リンク先の巡回を禁止 -->
<META name="robots" content="noindex, nofollow">

図4 robot.txt により巡回の拒否されたクローラー
https://archive.org/ (2017.02.08 21:20)

5. クライアント情報

UserAgent, リファラURLをはじめとしたクライアント情報は、クライアントが自己申告している値に過ぎない。したがって偽造が可能であり、PHPコードを混入させたPoPs(Points of Presence)攻撃を行うボットも存在する。

また、Windows, Linux等を名乗るボットも存在し、完全なクローリングの拒否は困難であると考えられる。ボットの巡回頻度, SEOによるボットのアクセス数の変化については未検証であり、あらゆるWebサイトで観測を行う必要がある。