F-021

ウェブアクセスログ分析のためのボット判別手法の提案
Bot Detection for Web Log Analysis

田中 孝昌†1　新堀 秀和†1　李 石映雪†1　野村 眞平†1
Takamasa Tanaka　Niibori Hidekazu　Li Shiyingxue　Shimpei Nomura
河島 宏樹†2　津田 和彦†3
Hiroki Kawashima　Kazuhiko Tsuda

1. はじめに

近年、ユーザー毎のウェブアクセスログを分析することで、ユーザーの行動パターンを抽出し、AI に学習させることで、ユーザー毎に適したコンテンツを自動で配信することは、一般的な機能となっている[1][2]。一方、ウェブアクセスログには、ボットによる閲覧情報が含まれる。ボットには DDos 攻撃やコンテンツの不正な大量抽出などを目的とした悪意のあるボットが存在する。さらには、ボットが自身の属性をユーザーに見せかけるという偽装を行う事も少なくない。そこで本研究では、ボットによるウェブアクセスログを分析する対象から除外するため、ユーザーとボットのウェブアクセスログを判別する手法を提案する。

2. ボット判別モデルの構築

2.1 モデル構築に用いる実験データ

本研究では、不動産広告ウェブサイトのウェブアクセスログを用いて、ユーザーまたはボットを判別するモデルの構築を行う。

2.2 判別するウェブアクセスログの単位

モデルで判別を行う単位は、同一 Cookie からの 1 回のサイト訪問を単位とする。サイト訪問は、セッションという言葉で定義する。当該 Cookie の最初のページビューからセッションは開始され、直前のページビューから 30 分以内のページビューは同一セッションとして扱う。30 分を超過してページビューが発生した場合は、別セッションとして扱う。

2.3 ボットとユーザーの識別ラベル

ボットによるセッションであることを示すラベルとして、ウェブページへのアクセス時に、ウェブページに実装された javascript の動作有無の情報を用いる。今日のウェブサイトでは、Javascript はページ動作の主要機能を担っている場合が一般的であり、それは実験データの提供サイトにも当てはまる。更に、多くのボットは、高速にサイトの情報を収集するために、ウェブページに実装された javascript を動作させない技術方式を採ることが一般的である。尚、javascript の動作有無の情報取得は、ウェブサイトにアクセスするクライアント側の通信状況や、次ページへの遷移タイミングなどにより欠損が生じるため、完全ではない。また、javascript を動作させる技術方式を採るボットは存在する。しかし、その数は相対的に少ないため本研究の対象外とする。

2.4 モデル構築に用いるセッションの特徴量

実験データの基礎分析をもとに、モデル構築に用いるセッションの特徴量を検討する。基礎分析の結果は誌面の都合のため省略する。

セッションのアクセス元を示すユーザーエージェントは、ユーザーによるアクセスの場合は、一般的なウェブブラウザの名称が含まれることが多い。その一方で、ボットによるアクセスの場合は、OS やプログラミング言語の名称が含まれることが多く、ボットの判定に有効な特徴量として活用できる。但し、ユーザーエージェントはテキスト情報であり、モデルの特徴量として用いるためには、何らかの変換処理が必要になる。本研究では、テキスト情報の一般的な変換処理である bag-of-words 表現を採用する。

本研究に用いた実験データには 4,930 個のユーザーエージェントが含まれ、691 個の単語を抽出できた。尚、ユーザーエージェントに含まれる数値、記号は除外し、アルファベットは、全て小文字に変換する。変換前後のユーザーエージェントの例を以下にしめす。

変換前：
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_4) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/12.1 Safari/605.1.15

変換後：
mozilla / macintosh / intel / mac / os / x / applewebkit / khtml / like / gecko / version / safari

図 1 ユーザーエージェントの変換の例

また、その他の有効な特徴量として、サイト内での振る舞いに関する情報を検討する。例えば、サイトに訪問する時間帯は、ユーザーは日中に訪問しやすい一方で、ボットは深夜から早朝にかけてのアクセスが多い。また、リファラーと呼ばれるサイトに流入する直前のページ情報は、ユーザーであれば検索サイトやウェブ広告ページであることが多いが、ボットはリファラー自体が取得できないことが多い。その他にも、ページビューの数が異常に多い、ページビュー間の時間間隔、ページビューの対象ページの種類など、ボットのサイト内の振る舞いに関する 14 個の特徴量を作成した。

†1 株式会社リクルート住まいカンパニー Recruit Sumai Company Ltd.
†2 株式会社野村総合研究所 Nomura Research Institute, Ltd.
†3 筑波大学ビジネス科学研究科 University of Tsukuba Business Science Department

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

2.5 ユーザーエージェントのみを用いたロジスティック回帰モデル

ユーザーエージェントによるボットの説明力を評価するために、ユーザーエージェントのみを用いたロジスティック回帰モデルを構築する。691 単語分の特徴量を投入し、L1 正則化を行うことで、判別に有効な単語の絞り込みを行う[3]。また、正則化項の係数は、モデルの AUC と偏回帰係数が 0 ではない特徴量の数を確認しながら決定する。

2.6 ユーザーエージェントとサイト内の振る舞いを用いた木構造モデル

ユーザーエージェントとサイト内の振る舞いを用いてボットの判別を行う。ユーザーエージェントに含まれる単語を用いた特徴量と、振る舞いに関する 14 個の特徴量を合わせて木構造モデルである LightGBM を用いる[4]。尚、単語は全 691 単語を投入したモデルと L1 正則化で有効性が確認できた単語に絞り込んだモデルの 2 つを構築する。

3. ボット判別モデルの評価

3.1 判別精度の定量評価

実験データの 80%を学習データ、20%を検証データとして、分割する。学習データから生成したモデルに対して、検証データへの当てはまり精度を、AUC、Accuracy、Precision、Recall の 4 つの指標で評価する。評価結果を表 1 に示す[5]。

表 1　3 つのモデルの精度の比較

モデル	特徴量	AUC	Accuracy	Precision	recall
1. ロジスティック回帰	UA 691 単語	0.933	0.902	0.997	0.813
2.LightGBM	UA691words+ 振る舞い	0.990	0.965	0.963	0.969
3.LightGBM	UA17words+ 振る舞い	0.989	0.964	0.963	0.968

先ず、モデル 1 の AUC,Accuracy は 3 つの中で最も低いものの、90%を超える判別精度が確認できており、ユーザーエージェントの bag-of-words 表現が有効に機能していると言える。また、サイト内の振る舞いを特徴量として投入したモデル 2 と 3 は、投入しないモデルより高い判別精度が確認できており、有効な特徴量が作成できたと言える。また、モデル 2 はユーザーエージェントに含まれる全 691 単語、モデル 3 は、モデル 1 で判別への有効性が確認できた 17 単語を特徴量として投入している。モデル 2 と 3 の Accuracy の差は、二項検定を行い、統計的に有意な差がないことを確認している。少量の単語数で同性能が確認できたことから、モデル 3 を最も優れたモデルと言える。

3.2 モデルの解釈

モデル 1 では、L1 正則化を行うことで、691 単語から、偏回帰係数が 0 ではない単語を 17 個に絞り込むことができた。単語の抜粋を図 2 に示す。また、正則化を行う際に、正則化係数を 3 パターンで試行した。図 3 に、正則化係数を大きくすることで、0 ではない偏回帰係数を持つ単語が絞り込まれていく過程を示す。表 2 に、正則化係数ごとの単語数と AUC を示す。AUC が同程度であることから 17 個の単語が重要であると解釈した。

ボット傾向のある単語：ubuntu/apple/go/linux

非ボット傾向のある単語：like/android/win/gecko

図 2 偏回帰係数が 0 ではない単語の抜粋

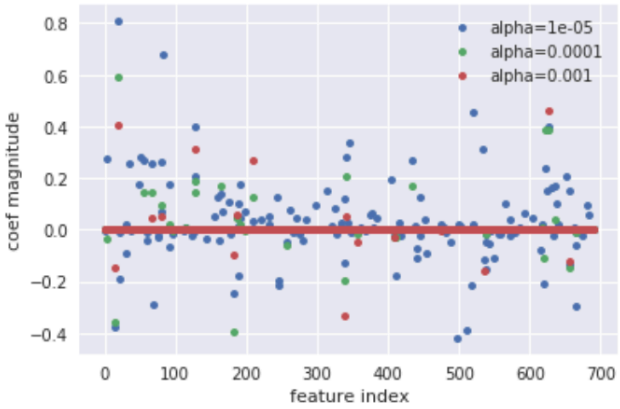

図 3 正則化項の係数 α と偏回帰係数の関係

表 2 正則化項の係数と有効な単語数および AUC の関係

正則化項の係数	10^{-5}	10^{-4}	10^{-3}
偏回帰係数が 0 ではない単語数	151	35	17
AUC	0.934	0.934	0.933

また、モデル 2 と 3 に投入したサイト内の振る舞いを示す特徴量は、ページビューの数、間隔の平均値と分散、サイトに訪問した時間帯などに、高い重要度が確認できた。

4. おわりに

本研究では、ウェブアクセスログ分析のノイズとなるボットからのアクセスを判別する手法を提案した。今後の課題として、javascript を稼働させるタイプのボットに対する対策を、分布推定による外れ値検知の手法を用いて行う。

謝辞

株式会社野村総合研究所の落合成光氏、中尾忠義氏に感謝の意を評します。

参考文献

[1] Li, Shiyingxue, et al. "Web-Scale Personalized Real-Time Recommender System on Suumo." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2017.
[2] 大塚真吾, and 喜連川優. "Web アクセスログとその利活用 (<特集> ソーシャルネットワーク時代の Web インタラクション)." *人工知能学会誌* 21.4 (2006): 410-415.
[3] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996): 267-288.
[4] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems*. (2017)
[5] 井出剛. "入門機械学習による異常検知." (2015)

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.