Łukasz Lipiński[1], Michał Bernardelli[2]

Detekcja wyłudzeń klików przez boty internetowe

1. Wstęp

Liczba botów w Internecie oraz zadań, które wykonują bądź mogłyby wykonywać, rośnie w bardzo szybkim tempie. Szacuje się, że ponad połowa ruchu internetowego jest generowana przez boty[3], przy czym algorytmy determinujące zachowanie botów są na tyle zaawansowane, że ich odróżnienie od realnych ludzi jest często bardzo trudne. Mimo że każdy z botów ma swoją specyfikę, to można wskazać pewne cechy wspólne, jeżeli nie dla wszystkich, to dla wielu automatycznych programów wyręczających człowieka. Intuicyjnie bowiem można stwierdzić, że zadaniem bota jest wykonać pracę, która wcześniej należała do człowieka. Różnica jest taka, że bot wykonuje ją na ogół szybciej, sprawniej, bez przerw i bezbłędnie. Podstawową zatem kwestią jest to, że program komputerowy ma imitować zachowanie człowieka lub czynności przez niego wykonywane, ale wykonując te zadania bardziej efektywnie.

Boty udające ludzi generują największą część, bo prawie jedną czwartą[4], ruchu internetowego spośród wszystkich rodzajów botów. Znanych jest wiele obszarów, w których udawanie ludzi jest popularnym działaniem. Jednym z klasycznych już przykładów jest tworzenie sztucznych tożsamości w celu wpłynięcia na opinię innych osób. Zakres możliwości jest tutaj praktycznie nieograniczony. Może być to zarówno wychwalanie zalet jakiegoś produktu lub usługi, jak i alternatywnie – krytykowanie produktów lub usług konkurencji. Oba te podejścia mogą wyraźnie poprawić statystyki sprzedaży (kosztem konkurencji).

Kontrowersyjnym przykładem jest zastosowanie botów w polityce. Ten sposób kreowania opinii publicznej był wykorzystywany m.in. w Turcji i Meksyku[5]

---

1 Cloud Technologies.
2 Szkoła Główna Handlowa w Warszawie, Kolegium Analiz Ekonomicznych.
3 https://www.incapsula.com/blog/bot-traffic-report-2016.html (odczyt: 06.01.2019).
4 Dane z 2016 r. Ibidem.
5 http://www.techinsider.io/political-bots-by-governments-around-the-world-2015-12 (odczyt: 06.01.2019).

przez boty działające na Twitterze. Istnieją jednak dowody[6] na wykorzystanie tej formy walki politycznej w USA, Chinach, Syrii, Argentynie, Hiszpanii, Egipcie, Rosji, Australii czy Tybecie. Problem ten dotyczy również Polski. Według raportu *Computational propaganda in Poland: False Amplifiers and the Digital Public Sphere*[7] podczas wyborów parlamentarnych w 2015 r. prawie jedna trzecia aktywności politycznej na Twitterze była generowana przez boty.

Jednym z ważniejszych pod względem zasięgu zastosowań botów udających ludzi jest sztuczne generowanie ruchu mające na celu poprawę statystyk oglądalności, co z kolei przekłada się na realne pieniądze pochodzące np. z reklam. Zyski dla autorów botów oznaczają wymierne straty w branży reklamowej. Z jednej strony są to nakłady finansowe poniesione na reklamę, która nie będzie miała żadnego przełożenia na poprawę sprzedaży czy popularności marki, a z drugiej strony nieprawdziwe dane mogą wpłynąć na podjęcie błędnych decyzji związanych z krótko- i długoterminowymi planami rozwoju firmy. Co więcej, w dłuższym horyzoncie można się spodziewać regresu tej branży ze względu na niechęć reklamodawców do tej nieskutecznej (przez wielość sztucznego ruchu) formy reklamy.

Jednym ze sposobów pomiaru skuteczności reklamy jest zliczanie kliknięć użytkowników, utożsamiane z zainteresowaniem klienta reklamowanym produktem. Rozliczenia typu PPC (ang. *pay-per-click*) powodują opłacalność tworzenia botów, które nie tylko wchodzą na strony internetowe, ale też wykonują symulację akcji kliknięcia wyświetlanego banera reklamowego, co należy zaklasyfikować jako oszukany klik (ang. *click fraud*). Identyfikacja takich botów jest kwestią kluczową dla firm operujących na rynku reklamy internetowej[8]. Wykrycie botów udających ludzi jest tym trudniejsze, im lepiej naśladują cechy i zachowania ludzkie. Imitacja człowieka, choćby wyjątkowo doskonała, powinna jednak przynajmniej raz na jakiś czas wzbudzać podejrzenia. Eksploracja miejsc potencjalnych różnic pomiędzy zachowaniem maszyny a człowieka powinna zatem pozwolić na ich rozróżnienie. Istnieje wiele technik wykorzystywanych w tym celu, przy czym część z nich bazuje zarówno na zaawansowanych narzędziach

---

6 https://medium.com/@erin_gallagher/articles-about-bots-trolls-around-the-world-f-0c563b037d2 (odczyt: 06.01.2019).

7 http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Poland.pdf (odczyt: 06.01.2019).

8 C. Kintana, D. Turner, J. Pan, A. Metwally, N. Daswani, E. Chin, A. Bortz, *The Goals and Challenges of Click Fraud Penetration Testing*, International Symposium on Software Reliability Engineering, 2009. Por. też M. Bernardelli, *Cheater detection in Real Time Bidding system – panel approach*, „Roczniki" Kolegium Analiz Ekonomicznych SGH, z. 39, Oficyna Wydawnicza SGH, Warszawa 2015, s. 11–23.

informatycznych, jak i na metodach matematycznych[9]. Opis wielu stosowanych podejść oraz ich potencjał został przedstawiony m.in. w raporcie ABT Shield *How to identify bots using mathematical modelling and IT tools*[10]. Do szerokiego spektrum stosowanych metod z pewnością należą modele ekonometryczne. Problematyka detekcji sztucznego ruchu jest jednak o tyle trudna, że brakuje na ogół zbiorów referencyjnych, a nawet jeżeli istnieją, to w zmieniającej się rzeczywistości internetowej szybko się dewaluują. Ponieważ większość metod ekonometrycznych jest zaliczana do procedur uczenia nadzorowanego[11], to w przypadku poruszanego problemu nie można ich zastosować. Istotne wydaje się zatem opracowanie takich algorytmów, których skuteczność nie opiera się na wcześniejszej nauce na zbiorach uczących.

Celem badania było przedstawienie propozycji efektywnego sposobu znajdowania reguł pozwalających na detekcję botów wyłudzających kliki. Na rozwiązanie składa się zestaw metryk charakteryzujących zachowanie człowieka lub bota podczas przeglądania strony internetowej, opracowanie reguł pozwalających na identyfikację sztucznego ruchu, jak również wykazanie istnienia zależności parametrów i reguł od urządzenia wykorzystywanego przez użytkownika do wyświetlania stron internetowych. Okazuje się bowiem, że różnice w urządzeniach mają zasadniczy wpływ nie tylko na zachowanie botów, ale też na zachowanie ludzi. Inaczej bowiem wyświetlają się strony internetowe w telefonach komórkowych, a inaczej w laptopach czy komputerach stacjonarnych. Często też urządzenia te są wykorzystywane do innych celów. Integralną częścią badania była analiza empiryczna na faktycznych danych pochodzących z Internetu. Zaproponowana metoda wykorzystuje panelowy charakter danych, ponieważ w przypadku każdego wejścia na stronę kluczowa jest chronologia zarejestrowanych działań wykonywanych przez użytkownika. Uzyskane w wyniku analizy reguły mogą być stosowane w praktyce do rozróżnienia botów od ludzi

[9] H. Saputra, E. Adi, S. Revina, *Comparison of Classification Algorithms to tell Bots and Humans Apart*, „Journal of Next Generation Information Technology" 2013, no. 4, s. 23–32; H. Xu, D. Liu, A. Koehl, H. Wang, A. Stavrou, *Click Fraud Detection on the Advertiser Side*, w: *Computer Security – ESORICS 2014*, red. M. Kutyłowski, J. Vaidya, Springer, 2014; R. Oentaryo, E. Lim, M. Finegold, D. Lo, F. Zhu, C. Phua, E. Cheu, G. Yap, K. Sim, N. Nhut, K. Perera, B. Neupane, M. Faisal, Z. Aung, W. Woon, W. Chen, D. Patel, D. Berrar, *Detecting Click Fraud in Online Advertising: A Data Mining Approach*, „Journal of Machine Learning Research" 2014, no. 15, s. 99–140.

[10] M. Bernardelli, *How to identify bots using mathematical modelling and IT tools*, ABT Shield, January 2019.

[11] Por. też: P. Efthimion, S. Payne, N. Proferes, *Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots*, „SMU Data Science Review" 2018, vol. 1, no. 2, Article 5.

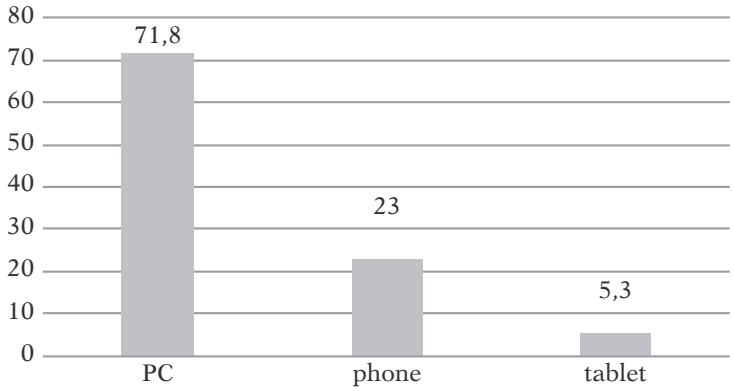
i pozwalają na skuteczne odwzorowanie ich zachowania w Internecie. Należy jednak podkreślić, że zachowania te, m.in. ze względu na postęp technologiczny, zmieniają się w czasie. Nie pozwala to tym samym na stwierdzenie, że podane w artykule reguły są ustalone, a wręcz wymusza cykliczne powtarzanie tego typu analiz, przy czym metodologia analizy wydaje się w znacznym stopniu uniwersalna (w przeciwieństwie do samych reguł).

Praca składa się z czterech części. W części drugiej, po niniejszym wstępie, przedstawiono charakterystykę danych wykorzystanych w badaniu, ze szczególnym uwzględnieniem różnic pomiędzy urządzeniami wykorzystywanymi do przeglądania stron internetowych. W trzeciej części opisano zastosowaną metodologię identyfikacji botów w tych odsłonach stron internetowych, gdy zarejestrowano chociaż jeden klik. Zamieszczone zostały również wyniki w postaci zestawu przykładowych reguł pozwalających na wspomnianą identyfikację. Artykuł zakończony jest wnioskami, w których wskazano na możliwe kierunki dalszych badań.

2. Charakterystyka danych

Do analizy wykorzystano dane z marca 2019 r. obejmujące 583 365 zapisów wejść użytkowników na europejskie strony internetowe, w trakcie których zarejestrowane zostało wykonanie co najmniej jednego klika. Każda odsłona strony internetowej była traktowana jako szereg czasowy akcji wykonanych przez użytkownika, przy czym zachowano również charakterystykę użytkownika reprezentowaną przez typ urządzenia (*devicetype*), z którego korzystał. Wyróżniono trzy typy urządzeń: komputer stacjonarny lub laptop (PC), telefon (phone) oraz tablet. Procentowy udział wymienionych urządzeń w ogólnej liczbie zarejestrowanych wejść na stronę zakończonych kliknięciem podano na rysunku 1. Okazuje się, że przeważająca liczba użytkowników korzysta z komputerów stacjonarnych lub laptopów, a zdecydowanie najmniej osób używa tabletów. Obserwacje te w dużym stopniu pokrywają się z wynikami innych badań. Przykładowo, według pomiarów statcounter.com[12] wykonanych od kwietnia 2018 r. do kwietnia 2019 r. na rynku europejskim udział odsłon internetowych z urządzeń stacjonarnych i laptopów w stosunku do tych wykonanych z telefonów i tabletów wynosi 56,33% do 38,40% i 5,27%.

[12] http://gs.statcounter.com (odczyt: 03.05.2019).



Rysunek 1. Podział ruchu internetowego ze względu na typ urządzenia wykorzystywanego przez użytkownika (na podstawie odsłon, które zakończyły się kliknięciem, dane w proc.)

Źródło: obliczenia własne.

Rejestrowane akcje zostały ograniczone do trzech rodzajów: klików (ang. *clik*), przewijania strony (ang. *scroll*) oraz akcji związanych z czasem przebywania na stronie. Czas wystąpienia konkretnej akcji był mierzony z dokładnością do milisekundy, a reprezentowany w postaci znacznika czasu, tzw. timestamp. Pomijano miejsca dokonania klika oraz wszelkie inne działania użytkownika poza przewijaniem strony. Statystyki liczby zarejestrowanych klików w trakcie odsłony strony internetowej podano w tabeli 1. Zdecydowanie przeważają odwiedziny, w trakcie których zarejestrowano zaledwie jeden klik. Zdarzają się jednak przypadki, w których liczba klików jest liczona w dziesiątkach. W analizowanym zbiorze danych największa liczba zarejestrowanych klików była równa 68.

Podczas przewijania strony internetowej możliwe jest rejestrowanie, do którego jej fragmentu dotarł użytkownik. W badaniu rejestrowano następujące etapy przewijania strony: enter, scroll_2_8, scroll_4_8, scroll_6_8 oraz scroll_8_8, przy czym enter oznacza wyświetlenie strony, oznaczenie scroll_X_Y zaś należy interpretować w następujący sposób: użytkownik przewinął stronę do X/Y jej długości, przykładowo scroll_6_8 oznacza, że wyświetlonych zostało 75% zawartości strony internetowej. Dodatkowo, rejestrowano akcje czasowe związane z długością wyświetlania strony, np. użytkownik oglądał stronę przez 30 sekund, 60 sekund itd. W badaniu kolejne etapy przewijania strony oraz akcje czasowe zostały ze sobą utożsamione i zliczane po prostu jako działanie użytkownika na stronie różne od klika. W analizie wykorzystano zmienne powiązane z akcjami klikowymi (nazywanymi w dalszej części artykułu klikami) oraz pozostałymi akcjami (nazywanymi po prostu akcjami). Kluczowa w kontekście

modelowania była chronologia działań użytkownika na stronie. Statystyki liczby zarejestrowanych akcji na stronie internetowej w trakcie jej wyświetlenia podano w tabeli 2. Największa liczba zarejestrowanych akcji na stronie w danych wykorzystanych w badaniu to 24. Należy przy tym podkreślić, że zazwyczaj strony są przewijane w jednym kierunku (z góry do dołu) i rejestrowane jednokrotnie. Stąd liczba zarejestrowanych akcji nieklikowych, nawet jeżeli użytkownik przewinie stronę do jej końca, jest ograniczona.

Tabela 1. Procentowy podział danych względem liczby zarejestrowanych klików z podziałem na typ urządzenia

Liczba klików	Ogółem	PC	Phone	Tablet
1	95,022	95,135	94,767	94,588
2	3,181	2,873	4,084	3,443
3	0,593	0,611	0,517	0,687
4	0,527	0,619	0,248	0,482
5	0,216	0,209	0,186	0,443
6	0,082	0,092	0,048	0,085
7	0,070	0,085	0,029	0,046
8	0,067	0,080	0,028	0,059
9	0,058	0,071	0,020	0,046
10+	0,185	0,225	0,074	0,124
Suma	100,000	100,000	100,000	100,000

Źródło: opracowanie własne.

Tabela 2. Procentowy podział danych względem liczby zarejestrowanych akcji (nieklikowych) na stronie internetowej z podziałem na typ urządzenia

Liczba akcji	Ogółem	PC	Phone	Tablet
1	36,04	33,75	45,15	35,26
2	10,64	11,25	8,50	9,87
3	10,17	11,20	6,42	9,17
4	13,49	13,36	12,60	18,68
5	13,45	12,55	16,96	13,45
6	1,82	2,12	0,73	1,68
7	2,42	2,85	0,94	1,62
8	3,70	4,13	2,20	3,02
9	4,35	4,51	3,75	4,24
10+	3,92	4,28	2,74	3,02
Suma	100,000	100,000	100,000	100,000

Źródło: opracowanie własne.

Na podstawie szeregu czasowego reprezentującego konkretną odsłonę strony internetowej utworzono następujące zmienne, wykorzystane w modelowaniu zjawiska klikalności:

- *nclicks* – liczba zarejestrowanych klików podczas wyświetlania strony internetowej;
- *nscrolls* – liczba zarejestrowanych akcji na stronie internetowej;
- *scrolls_per_click* – liczba zarejestrowanych akcji na stronie internetowej na jeden klik;
- *click_response_time* – różnica czasu pomiędzy pierwszym klikiem a czasem wejścia na stronę internetową;
- *engage_time* – różnica czasu pomiędzy ostatnią zarejestrowaną akcją a czasem wejścia na stronę internetową;
- *mean_time_between_scrolls* – średni czas pomiędzy kolejnymi akcjami na stronie internetowej, jeżeli zarejestrowane zostały co najmniej dwie akcje;
- *mean_time_between_clicks* – średni czas pomiędzy kolejnymi kliknięciami użytkownika na stronie internetowej, jeżeli zarejestrowane zostały co najmniej dwa kliki;
- *std_time_between_scrolls* – odchylenie standardowe czasu pomiędzy kolejnymi akcjami na stronie internetowej, jeżeli zarejestrowane zostały co najmniej dwie akcje;
- *std_time_between_clicks* – odchylenie standardowe czasu pomiędzy kolejnymi kliknięciami użytkownika na stronie internetowej, jeżeli zarejestrowane zostały co najmniej dwa kliki.

Statystyki opisowe dotyczące każdej z wymienionych zmiennych z podziałem na typ urządzenia podano w tabeli 3.

Tabela 3. Statystyki opisowe zmiennych wykorzystanych w analizie z podziałem na typ urządzenia

nclicks			
device	PC	phone	tablet
mean	1,12	1,08	1,11
std	0,85	0,54	0,71
min	0,00	0,00	0,00
25%	1,00	1,00	1,00
50%	1,00	1,00	1,00
75%	1,00	1,00	1,00
max	68	31	23

nscrolls			
device	PC	phone	tablet
mean	4,11	2,69	3,52
std	4,57	3,92	4,29
min	0,00	0,00	0,00
25%	0,00	0,00	0,00
50%	3,00	1,00	1,00
75%	7,00	5,00	6,00
max	24	18	18

cd. tabeli 3

click_response_time			
device	PC	phone	tablet
mean	6,76	1,97	3,07
std	34,61	15,91	24,23
min	0,00	0,00	0,01
25%	0,25	0,17	0,26
50%	0,62	0,38	0,53
75%	2,09	1,00	1,30
max	1503,18	1120,34	1194,71

engage_time			
device	PC	phone	tablet
mean	1,93	1,28	1,84
std	13,27	8,92	14,31
min	0,00	0,00	0,00
25%	0,00	0,00	0,00
50%	0,00	0,00	0,00
75%	1,29	0,59	1,16
max	1108,62	973,56	746,94

mean_time_between_clicks			
device	PC	phone	tablet
mean	8,80	3,53	4,70
std	48,15	20,02	24,91
min	0,00	0,00	0,00
25%	0,04	0,03	0,04
50%	0,37	0,19	0,48
75%	2,72	1,47	3,12
max	940,28	558,85	525,42

mean_time_between_scrolls			
device	PC	phone	tablet
mean	1,09	1,14	1,18
std	7,36	7,34	8,45
min	0,00	0,00	0,00
25%	0,23	0,29	0,26
50%	0,46	0,55	0,52
75%	0,89	1,04	0,97
max	1029,66	973,56	629,33

std_time_between_clicks			
device	PC	phone	tablet
mean	2,77	1,07	1,36
std	16,89	10,06	7,73
min	0,00	0,00	0,00
25%	0,00	0,00	0,00
50%	0,00	0,00	0,00
75%	0,32	0,00	0,22
max	433,06	558,10	227,08

std_time_between_scrolls			
device	PC	phone	tablet
mean	0,80	0,68	0,84
std	6,78	5,48	8,31
min	0,00	0,00	0,00
25%	0,05	0,06	0,06
50%	0,18	0,19	0,19
75%	0,44	0,41	0,44
max	522,22	375,68	333,29

Źródło: opracowanie własne.

Z danych umieszczonych w tabeli 3 można wyciągnąć wiele wniosków, m.in. te wynikające z porównania średnich. Ograniczając się do trzech najbardziej znaczących różnic, należy zwrócić uwagę na fakt, że różnica czasu pomiędzy pierwszym klikiem a czasem wejścia na stronę internetową (*click_response_time*) dla użytkowników używających PC jest w porównaniu z tabletami średnio ponad dwa razy większa, a w porównaniu z telefonami ponad trzykrotnie większa. Odpowiada za to w głównej mierze konstrukcja wyświetlania stron internetowych na poszczególnych urządzeniach.

Drugi z wniosków dotyczy różnicy czasu pomiędzy ostatnią zarejestrowaną akcją a czasem wejścia na stronę internetową (*engage_time*). Z danych wynika, że średnio najkrócej są przeglądane strony w telefonach komórkowych. W przypadku telefonów najmniejsze jest też odchylenie standardowe. Biorąc pod uwagę celowość przeglądania stron internetowych w telefonach, które są w głównej mierze używane do łączenia z mediami społecznościowymi, przeglądania witryn sklepowych czy czytania wiadomości na portalach internetowych, ale raczej nie do pracy *sensu stricto*, należy stwierdzić, że taka obserwacja wydaje się mieć logiczne uzasadnienie.

Ostatni z wniosków dotyczy prawidłowości dotyczących kolejnych kliknięć na stronach internetowych, a dokładniej wielkości *mean_time_between_clicks* oraz *mean_time_between_scrolls*. Okazuje się, że największe średnie odległości czasowe pomiędzy klikami występują w przypadku użytkowników korzystających z PC – są one ponad dwukrotnie większe niż w przypadku użytkowników telefonów komórkowych. Tymczasem w przypadku analogicznych odległości pomiędzy akcjami (nieklikowymi) nie zarejestrowano tak dużych różnic w zależności od wykorzystywanego urządzenia.

Dane z tabeli 3 pozwalają na odczytanie zależności od urządzeń wykorzystywanych do łączenia z Internetem dla poszczególnych zmiennych. Nie dają one jednak możliwości wnioskowania na temat wzajemnych powiązań pomiędzy tymi zmiennymi, a nade wszystko nie dają ścisłych reguł pozwalających na odróżnienie użytkowników rzeczywistych od botów internetowych. Kompleksowa próba modelowania w przypadku zagadnienia bez nadzoru, jakim jest przedstawiony w pracy problem, została przedstawiona w kolejnej części artykułu.

3. Analiza empiryczna

Postawiony w artykule problem wykrywania wyłudzeń klików na stronach internetowych sprowadza się do określenia zależności, które pozwolą na identyfikację tego typu zdarzeń. Ze względu na brak wartości referencyjnych rozwiązanie tego problemu opierało się na modelowaniu bez nadzoru. Dlatego została pominięta analiza wykorzystująca modele ekonometryczne. W zamian zastosowano algorytmy grafowe, a konkretnie – rozwiązanie oparte na lasach, znane pod nazwą Isolation Forest lub w skrócie iForest. Metoda ta jest pod względem ideologii podobna do często używanej metody lasów losowych (ang. *Random*

Forest[13]), ale w zamierzeniu służy do wykrywania obserwacji odstających w wielowymiarowych przestrzeniach. Większość istniejących podejść do modelowania identyfikacji anomalii w danych opiera się na mniej lub bardziej zaawansowanym profilowaniu sytuacji uznawanych za mieszczące się w normach, a następnie znajduje te sytuacje, które nie pasują do utworzonego profilu. Tymczasem iForest podaje *explicite* reguły wyróżniające wartości odstające[14]. Zakładając, że takimi odstającymi od normy obserwacjami są wizyty na stronach generowane przez boty internetowe, otrzymujemy efektywny sposób detekcji wyłudzeń klików. Ze względu na specyfikę algorytmu jako obserwacje odstające zostaną potraktowane również anormalne zachowania ludzi, przy czym zazwyczaj odróżniają się one w nieco inny sposób niż obserwacje będące wyłudzeniami.

Implementacja algorytmu iForest z biblioteki Python sklearn została wykorzystana do rozpoznania zestawu czynników, które mogą pomóc w identyfikacji prób wyłudzeń klików, przy czym celem było poznanie reguł obejmujących całe zestawy zmiennych, a nie badanie każdej zmiennej z osobna. Badanie dotyczyło stworzenia dwóch modeli: jednego bez uwzględniania urządzenia, jakim posługuje się użytkownik korzystający z Internetu, a drugie z uwzględnieniem danych na temat używanych przez użytkowników urządzeń. Motywacją do wykorzystania dodatkowych danych jest analiza przedstawiona w poprzedniej części. W wyniku działania algorytmu iForest zidentyfikowano odpowiednio 3734 reguły dla danych bez informacji o urządzeniach użytkowników oraz 4629 reguł dla danych zawierających te informacje. Nie sposób wymienić i opisać wszystkie uzyskane reguły, stąd zdecydowano się na przedstawienie kilku przykładowych, mając na uwadze prezentację potencjału stojącego za tym podejściem, jak również pozwalającego na określenie znacznie lepszych, konkretnych zasad filtrowania ruchu internetowego z uwzględnieniem specyfiki urządzeń (PC, phone, tablet). Część z reguł jest w dużym stopniu redundantna, ale każda z nich łatwa do przetłumaczenia na język naturalny.

Przykładowe reguły podano w tabelach 4 i 5. Porównując reguły w wersji z uwzględnieniem urządzenia do reguł z wersji bez jego uwzględnienia, można zauważyć pewną prawidłowość, a mianowicie jeżeli nie mamy informacji o urządzeniu, to reguły są na tyle restrykcyjne, że przypuszczalnie pozwolą na odfiltrowanie niewielkiego odsetka ruchu. Interpretacja reguł jest zgodna z intuicją,

[13] T. Shi, S. Horvath, *Unsupervised learning with random forest predictors*, „Journal of Computational and Graphical Statistics" 2006, vol. 15(1), s. 118–138.

[14] F. Liu, K. Ming Ting, Z. Zhou, *Isolation Forest*, 2008 Eighth IEEE International Conference on Data Mining, 2009, s. 413–422.

ale prawdopodobieństwo, że spełniony będzie komplet określonych przez regułę warunków, jest znikome. Diametralnie inną sytuację mamy w momencie, gdy znamy urządzenie, którym posługuje się użytkownik. Możemy wówczas znacznie bardziej precyzyjnie określić, czy ruch jest sztuczny, czy też pochodzi od botów, imitujących ludzi. Zdecydowana większość reguł przypisuje sztuczny ruch takim użytkownikom, którzy wyświetlają ją na ułamek sekundy i zdążą w tym czasie wykonać klik. Kolejna, zdroworozsądkowa prawidłowość to przypisanie wyłudzenia obserwacji, przy której klik wystąpił nadzwyczaj szybko, przy czym sam czas wyświetlania strony nie musi być krótki.

Tabela 4. Opisy przykładowych reguł identyfikacji anomalii przy odsłonach stron internetowych w wersji bez uwzględnienia typu urządzenia

Lp.	Opis reguły
1.	brak akcji, 1 klik wykonany w czasie od 4,03 do 4,12 sekundy
2.	8 akcji, 1 klik, czas wyświetlania strony poniżej 1,04 sekundy
3.	wyświetlenie strony pomiędzy 7,2 a 10,15 sekundy, pierwszy klik w czasie poniżej 1,54, średni czas pomiędzy kolejnymi akcjami poniżej 1,28
4.	brak akcji, pierwszy klik wykonany po 46,66 sekundy lub później
5.	1–2 kliki, z czego pierwszy wykonany po 1,88 sekundy lub później, 7–8 akcji na każdy klik wykonywanych w podobnych odstępach czasu (z dokładnością do 0,2 sekundy)

Źródło: opracowanie własne.

Tabela 5. Opisy przykładowych reguł identyfikacji anomalii przy odsłonach stron internetowych w wersji z uwzględnieniem typu urządzenia

Lp.	Opis reguły
1.	PC, 1 klik w czasie poniżej 0,35 sekundy, 16 zarejestrowanych akcji, czas odsłony co najwyżej 4,73 sekundy
2.	PC, 1–2 kliki, z czego pierwszy po ponad 77 sekundach, co najmniej 3 akcje wykonywane z odchyleniem standardowym co najwyżej 0,61 sekundy
3.	telefon, 1 klik, 6–8 akcji, średnio czas pomiędzy kolejnymi akcjami od 1,03 do 1,6 sekundy
4.	telefon, 1 klik, brak akcji, czas do pierwszego klika od 3,41 do 8,31 sekundy
5.	tablet, 1 klik, co najmniej 1 akcja, czas wyświetlenia strony co najwyżej 0,09 sekundy
6.	tablet, 1 klik po mniej niż 0,19 sekundy, brak akcji

Źródło: opracowanie własne.

4. Wnioski

Efektywne wykrywanie wyłudzeń klików na stronach internetowych przez automatyczne programy komputerowe stanowi wartościowe narzędzie w walce z tego typu oszustwami i daje natychmiastową, wręcz wymierną korzyść w postaci oszczędności pieniędzy niewydanych na źle ukierunkowaną reklamę. Zmieniająca się szybko technologia wymusza wybór innych parametrów w regułach odróżniających wartościowy ruch w Internecie wykonywany przez ludzi od ruchu generowanego przez boty, w zależności od urządzenia wykorzystywanego do surfowania po Internecie. Parametry te mogą też ulegać modyfikacjom wraz z doskonaleniem botów internetowych w postaci lepszego dopasowania do ludzkiego zachowania. Celem badania opisanego w artykule było przedstawienie metody uczenia nienadzorowanego, która daje zestaw reguł pozwalających na efektywne wykrywanie botów wyłudzających kliki, przy założeniu, że ich zachowanie różni się w jakichś aspektach od zachowania ludzi. Można wymienić m.in. następujące korzyści z prezentowanego podejścia:

- efektywne działanie przy braku zbiorów referencyjnych;
- łatwość interpretacji uzyskanych reguł decyzyjnych;
- możliwość uogólnień w postaci dołączania innych zmiennych;
- uniwersalność zastosowanego podejścia oraz jego pełną skalowalność (na dowolne akcje);
- dostępność kalibracji parametrów w regułach w zależności od ustalonego stopnia ryzyka niewykrycia oszustwa w stosunku do nadmiernego filtrowania ruchu.

Zaprezentowane podejście wydaje się doskonałą alternatywą dla modeli ekonometrycznych, przy dużych ilościach danych oraz ograniczeniach dostępności informacji na temat faktycznej klasyfikacji analizowanego ruchu. Podejście to wykorzystuje panelowy charakter danych i wymaga rozbudowanej infrastruktury informatycznej, jeżeli miałoby być wykorzystywane w czasie rzeczywistym. W takim przypadku jednak metoda ta pozwala na natychmiastowe wręcz oszczędności w postaci lepszego sprofilowania odbiorców reklamy.

Bibliografia

Bernardelli M., *Cheater detection in Real Time Bidding system – panel approach*, „Roczniki" Kolegium Analiz Ekonomicznych SGH, z. 39, Oficyna Wydawnicza SGH, Warszawa 2015, s. 11–23.

Bernardelli M., *How to identify bots using mathematical modelling and IT tools*, ABT Shield, January 2019.

Efthimion P., Payne S., Proferes N., *Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots*, „SMU Data Science Review" 2018, vol. 1, no. 2, Article 5.

Kintana C., Turner D., Pan J., Metwally A., Daswani N., Chin E., Bortz A., *The Goals and Challenges of Click Fraud Penetration Testing*, International Symposium on Software Reliability Engineering, 2009.

Liu F., Ming Ting K., Zhou Z., *Isolation Forest*, 2008 Eighth IEEE International Conference on Data Mining, 2009, s. 413–422.

Oentaryo R., Lim E., Finegold M., Lo D., Zhu F., Phua C., Cheu E., Yap G., Sim K., Nhut N., Perera K., Neupane B., Faisal M., Aung Z., Woon W., Chen W., Patel D., Berrar D., *Detecting Click Fraud in Online Advertising: A Data Mining Approach*, „Journal of Machine Learning Research" 2014, no. 15, s. 99–140.

Saputra H., Adi E., Revina S., *Comparison of Classification Algorithms to tell Bots and Humans Apart*, „Journal of Next Generation Information Technology" 2013, no. 4, s. 23–32.

Shi T., Horvath S., *Unsupervised learning with random forest predictors*, „Journal of Computational and Graphical Statistics" 2006, vol. 15(1), s. 118–138.

Xu H., Liu D., Koehl A., Wang H., Stavrou A., *Click Fraud Detection on the Advertiser Side*, w: *Computer Security – ESORICS 2014*, red. M. Kutyłowski, J. Vaidya, Springer, 2014.

Źródła sieciowe

http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Poland.pdf (odczyt: 06.01.2019).

http://gs.statcounter.com (odczyt: 03.05.2019).

https://medium.com/@erin_gallagher/articles-about-bots-trolls-around-the-world-f0c563b037d2 (odczyt: 06.01.2019).

https://www.incapsula.com/blog/bot-traffic-report-2016.html (odczyt: 06.01.2019).

http://www.techinsider.io/political-bots-by-governments-around-the-world-2015-12 (odczyt: 06.01.2019).

* * *

Click fraud detection rules

Summary

Effective detection of clicks on websites done by automatic computer programs is a valuable tool in the fight against this type of fraud and gives immediate measurable benefit in the form of savings for poorly targeted advertising. The purpose of the study described in the article was to present the unsupervised learning method, which results in a set of rules that allow for effective detection of bots that are responsible for the click frauds, assuming that their behavior differs in some aspects from human behavior. This analysis proves that involving device type as an extra variable improves the effectiveness of rules used for fraud detection and that the proposed algorithm provides a flexible and efficient solution for the given problem.

Keywords: click fraud, iForest algorithm, anomalies detection, Internet bot.