DBAJ O JAKOŚĆ DANYCH W GOOGLE ANALYTICS

Tekst: Maciej Ilczyszyn

PODEJŚCIE „DATA DRIVEN DECISION MAKING" (PODEJMOWANIE DECYZJI W OPARCIU O DANE) ZYSKUJE CORAZ WIĘKSZE GRONO ZWOLENNIKÓW. PODEJMOWANIE DECYZJI NA PODSTAWIE WIDZIMISIĘ PREZESA NA SZCZĘŚCIE ODCHODZI DO LAMUSA. W BIZNESIE INTERNETOWYM WDROŻENIE NOWEGO PODEJŚCIA JEST STOSUNKOWO ŁATWE, PONIEWAŻ NARZĘDZIA ANALITYCZNE SĄ DOSTĘPNE DLA KAŻDEGO. CZY JEDNAK ZAWSZE MOŻEMY IM BEZWZGLĘDNIE ZAUFAĆ?

Skoro analiza zgromadzonych danych ma kierować biznesem, to wiarygodność informacji, które zbieramy, jest absolutnie kluczowa. Niestety narzędzia do badania zachowań użytkowników na stronach internetowych nie są w pełni odporne na dane pochodzące z odsłon wykonywanych, nie przez prawdziwych użytkowników, a przez różnego rodzaju automatyczne skrypty, boty, pająki. Nie inaczej jest z najpopularniejszym Google Analytics.

Pochodzenie fałszywych danych w Google Analytics

Google Analytics na stronie WWW to, bez wchodzenia w szczegóły, uruchamiający się w odpowiednim momencie java script. Kiedyś wszystkie roboty internetowe poruszające się po witrynach w celu skanowania ich treści, linków, sprawdzania pozycji, takie jak np. Googlebot, nie uruchamiały skryptów znajdujących się na stronach.

Pierwszy robot, który wyłamał się z tej konwencji został stworzony przez Semalt i pojawiał się w raportach źródeł ruchu Analyticsa jako semalt.com/referral już w pierwszym kwartale 2014 r. Były to pojedyncze odsłony i nie dotykały tak wielu witryn.

Na przełomie lat 2014 i 2015 problem stał się palący. Analogicznie działających robotów pojawiło się znacznie więcej i dla witryn z niewielkim ruchem stanowiły już poważne zaburzenie statystyk.

Jak rozpoznać roboty w źródłach ruchu witryny

Najlepiej posłużyć się raportem *Źródło/medium* w dziale *Cały ruch* sekcji *Pozyskanie*. Wizyty, w teorii przekierowane z innych witryn (referral), które mają współczynnik odrzuceń oraz udział nowych sesji równe 0% lub 100%, to najczęściej roboty (rys. 1).

Istnieje niestety jeszcze inny rodzaj robotów, które potrafią lepiej ukryć się w raportach Google Analytics. Załóżmy, że obserwujesz wzrost odwiedzin bezpośrednich (direct/none).

Rysunek 1. **Raport źródeł ruchu zawierający mylne dane**

	Źródło / Medium	Pozyskiwanie			Zachowanie		
		Sesje ↓	% nowych sesji	Nowi użytkownicy	Współczynnik odrzuceń	Strony / sesja	Śr. czas trwania sesji
		1 353 % całości: 100,00% (1 353)	73,91% Śr. dla widoku danych: 73,91% (0,00%)	1 000 % całości: 100,00% (1 000)	71,40% Śr. dla widoku danych: 71,40% (0,00%)	1,77 Śr. dla widoku danych: 1,77 (0,00%)	00:01:05 Śr. dla widoku danych: 00:01:05 (0,00%)
1.	google / organic	674 (49,82%)	62,61%	422 (42,20%)	61,57%	2,15	00:01:29
2.	4webmasters.org / referral	293 (21,66%)	97,61%	286 (28,60%)	99,66%	1,01	00:00:01
3.	(direct) / (none)	187 (13,82%)	70,05%	131 (13,10%)	62,57%	1,75	00:01:17
4.	best-seo-offer.com / referral	37 (2,73%)	100,00%	37 (3,70%)	100,00%	1,00	00:00:00
5.	podroze-angelika.pl / referral	21 (1,55%)	14,29%	3 (0,30%)	76,19%	1,38	00:01:30
6.	twojbilet.eu / referral	15 (1,11%)	93,33%	14 (1,40%)	66,67%	2,20	00:01:22
7.	bing / organic	12 (0,89%)	58,33%	7 (0,70%)	50,00%	1,92	00:02:05
8.	barbara-travel.rzeszow.pl / referral	10 (0,74%)	100,00%	10 (1,00%)	90,00%	1,20	00:00:08
9.	buttons-for-your-website.com / referral	9 (0,67%)	100,00%	9 (0,90%)	100,00%	1,00	00:00:00

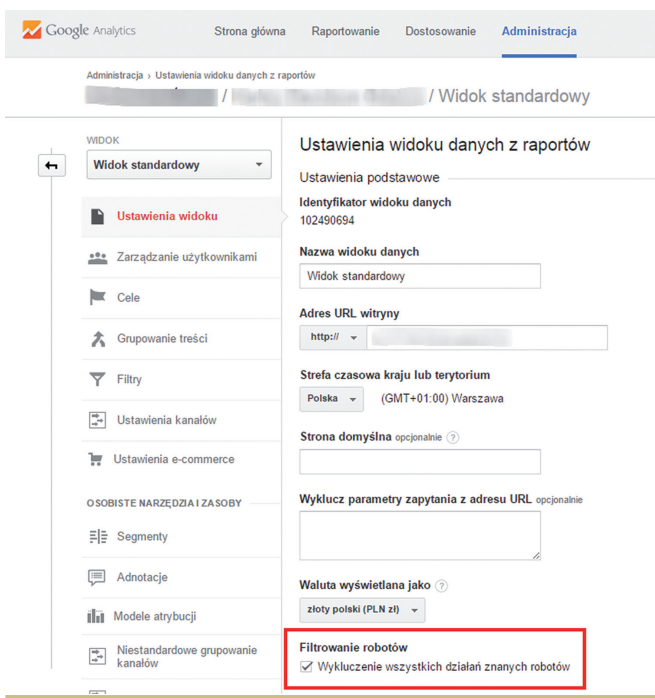
Rysunek 2. **Ruch niektórych robotów daje się zauważyć dopiero po głębszej analizie**

	Źródło / Medium	Nazwa hosta	Pozyskiwanie	
			Sesje ↓	% nowych sesji
			1 530 % całości: 100,00% (1 530)	73,46% Śr. dla widoku danych: 73,40% (0,09%)
1.	(direct) / (none)	[redacted]	504 (32,94%)	43,85%
2.	(direct) / (none)	(not set)	495 (32,35%)	100,00%
3.	google / organic	[redacted]	199 (13,01%)	70,85%
4.	site4.free-share-buttons.com / referral	(not set)	49 (3,20%)	100,00%
5.	best-seo-offer.com / referral	[redacted]	40 (2,61%)	100,00%
6.	guardlink.org / referral	(not set)	40 (2,61%)	100,00%
7.	poczta.wp.pl / referral	[redacted]	33 (2,16%)	39,39%

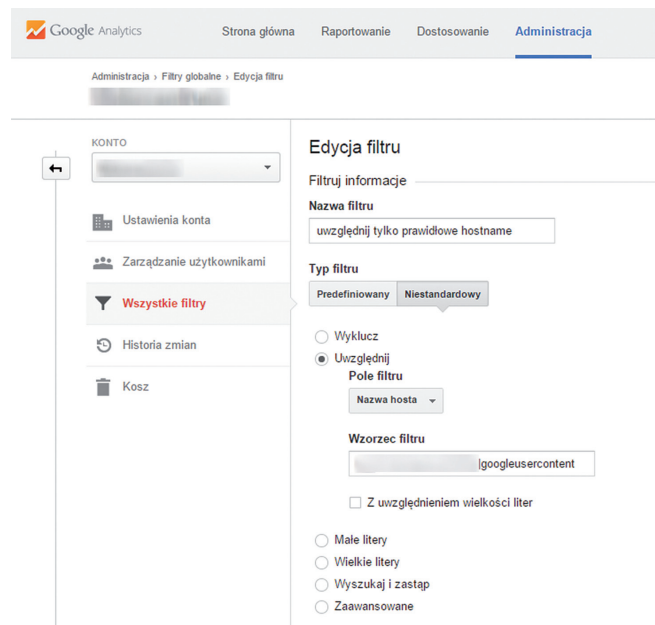
Tabela 1. **Wyniki testu funkcji *Wykluczania wszystkich działań znanych robotów***

	Widok zawierający wszystkie dane	Widok z włączoną funkcją *Wykluczania wszystkich działań znanych robotów*	Widok z odfiltrowanymi danymi wizyt robotów
Liczba sesji	4904	4898	4212
Odfiltrowanych sesji	-	6	692
% wizyt robotów, jaki został odfiltrowany	-	0,87%	100%

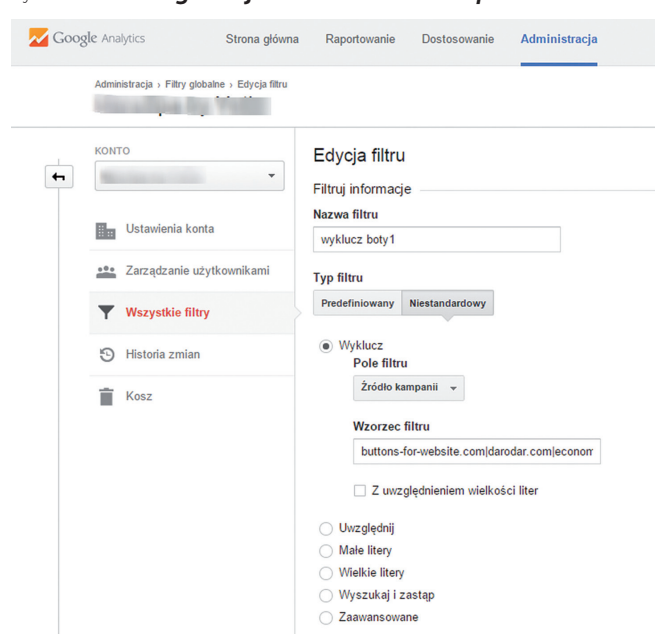
Rysunek 3. **Ustawienia widoku w Google Analytics**



Rysunek 4. **Konfiguracja filtru nazwy hosta**



Rysunek 5. **Konfiguracja filtru *źródła kampanii***



Zadowolony raportujesz, że prowadzona kampania radiowa przynosi efekty. Po dodaniu do raportu wymiaru dodatkowego *Nazwa hosta*, zauważasz prawdziwe pochodzenie tego wzrostu (rys. 2). Połowa odwiedzin bezpośrednich, to odwiedziny robotów.

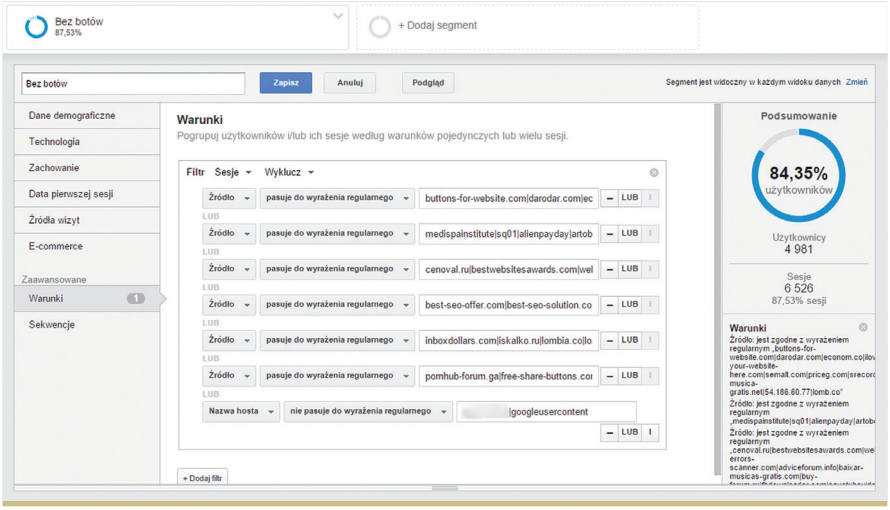
Jak duży to problem?

Na tyle duży, że Google w lipcu 2014 r. dodało do Analyticsa funkcję *Wykluczenia wszystkich działań znanych robotów*, włączaną w ustawieniach widoku. Teoretycznie funkcja ta powinna wystarczyć, jednak lista „znanych robotów", którą posługuje się Google, jest wyraźnie rzadko aktualizowana.

Wyniki eksperymentu przeprowadzonego na przełomie czerwca i lipca 2015 r. wykazały słabą skuteczność wspomnianej funkcji Google Analytics (tab. 1).

Twórcy narzędzi do monitoringu pozycji, obecności brandu w sieci czy do badania linkowania witryn nie poprzestają w wysiłkach i wciąż tworzą nowe roboty ukrywające się pod nowymi nazwami. Do tej pory Google pozostaje w ogonie peletonu, więc musimy sobie sami

Rysunek 6. **Konfiguracja segmentu niezawierającego danych z odwiedzin robotów**



poradzić z problemem. Na szczęście jest to możliwe.

Wykluczamy fałszywe dane

Istnieją rozwiązania serwerowe, jednak poniżej zaprezentowany został sposób na pozbycie się niechcianych danych w samym Google Analytics.

1. Pierwszym krokiem powinno być utworzenie nowego widoku danych, tak aby mieć do dyspozycji co najmniej 2. Jeden z zupełnie niefiltrowanymi danymi, a drugi z danymi, które poddasz wszystkim możliwym zabiegom oczyszczającym.

2. Przejdź do widoku, który ma raportować tylko prawdziwe dane i włącz opisywaną funkcję *Wykluczania wszystkich działań znanych robotów* (rys. 3). Cóż, może kiedyś zacznie działać jak należy…

3. Stwórz filtr weryfikujący nazwę hosta.

Musi to być filtr, który spowoduje gromadzenie danych tylko z odsłon z właściwą nazwą hosta dla Twojej witryny (rys. 4).

Zawartość pola *Wzorzec filtru* to wyrażenie regularne, które oprócz domeny witryny powinno obejmować *googleusercontent* (np. strony tłumaczone Translatorem Google) i w zależności od potrzeb domeny takich usług jak bramki płatności, narzędzia do testowania landing page'y czy subdomeny, jeżeli stosujesz śledzenie w wielu domenach. Jeśli na tym etapie zapomnisz o jakiejś pozycji, to odzyskanie danych nie będzie możliwe. Dlatego zawsze warto mieć widok niefiltrowany. Przejrzyj dokładnie raport źródeł ruchu rozszerzony o nazwę hosta za ostatni rok i wypisz z niego wszystkie prawidłowe nazwy hosta. Zastosuj filtr do odpowiednich widoków danych.

4. Filtruj nieprawidłowe dane na postawie nazwy źródła.

Musisz utworzyć filtr, który będzie szukał w *źródle kampanii* określonych ciągów znaków (nazw domen robotów) i je wykluczał (rys. 5).

Robotów jest tak dużo, że wykonanie wszystkich potrzebnych filtrów, nawet przy zastosowaniu wyrażeń regularnych, byłoby strasznie czasochłonne. Na szczęście istnieją listy robotów, z których możemy kopiować całe wyrażenia regularne (np. bit.ly/1GtRCk5).

Jeszcze szybszym rozwiązaniem jest użycie **skryptu** stworzonego przez Simo Ahava'ę (bit.ly/1LcVsfX), **który utworzy odpowiednie filtry** na Twoim koncie.

Co z raportami historycznymi?

Po zastosowaniu 3 opisanych ustawień, gromadzisz tylko dane, którym możesz zaufać. Tak, ale co jeśli chcesz analizować dane historyczne, które nie są czyste? Jest dla Ciebie ratunek. Musisz utworzyć segment, w którym zastosujesz te same reguły, co w filtrach powyżej (rys. 6).

Po zastosowaniu takiego segmentu Google Analytics włącza próbkowanie, więc mimo wszystko warto tworzyć filtry i działać na raportach niepróbkowanych.

Podsumowanie

W świecie niezmąconych danych nie pozostaniesz jednak na zawsze. Regularnie pojawiają się bowiem nowe roboty, które będziesz musiał dopisywać do filtrów w swoim koncie. Być może dane zakłócą jeszcze inne rodzaje niepożądanych działań (np. fałszywe zdarzenia w raportach Google Analytics). Poza tematem artykułu pozostaje oczywiście filtrowanie własnych wizyt i prawidłowa implementacja kodów śledzących.

Analizując ruch w witrynach, które nie mają ilości wizyt rzędu kilkudziesięciu czy kilkuset tysięcy miesięcznie, można z powodu opisanych działań robotów zupełnie opacznie zinterpretować raporty w Google Analytics. Dbaj zatem o jakość gromadzonych danych, analizuj je i podejmuj na ich podstawie decyzje, bo to wciąż najlepsza metoda na kierowanie biznesem.

Maciej Ilczyszyn – analityk internetowy, specjalista ds. konwersji w Yetiz Interactive

Wdraża narzędzia analityczne. Dane z nich płynące przekuwa w lepsze wyniki finansowe właścicieli witryn internetowych. W pogoni za wyższą konwersją prowadzi badania użyteczności, testy A/B i optymalizuje kanały ruchu. Posiada certyfikaty Google AdWords i Google Analytics Individual Qualification. Optymalizuje konwersję stron z branży ubezpieczeniowej, turystycznej, motoryzacyjnej.

NAPISZ DO AUTORA:
maciej.ilczyszyn@yetiz.pl

yetiz