

Metoda estymacji rzeczywistej liczby użytkowników (Real Users) witryn internetowych

1. Wstęp

Jednym z podstawowych wskaźników oglądalności witryn internetowych jest liczba użytkowników (internautów), którzy odwiedzili badaną witrynę w określonym przedziale czasu. Estymacja wartości tego wskaźnika (w odróżnieniu np. od liczby odsłon wykonanych na badanej witrynie) wymaga uwzględnienia wielu skomplikowanych zjawisk, których źródeł należy szukać zarówno w technologii używanej w komunikacji internetowej, jak i różnych sposobach korzystania z internetu przez internautów.

Niniejszy dokument opisuje unikalną i autorską metodę estymacji rzeczywistej liczby użytkowników (Real Users) badanych witryn internetowych w zadanym okresie czasu opracowaną przez firmę badawczą Gemius SA. Metoda ta wykorzystuje wyniki badania site-centric (gemiusTraffic) oraz informacje o liczebności całej populacji internautów.

2. Liczba cookies a liczba internautów

Wszystkie systemy site-centric oraz systemy adserwerowe prezentują liczbę cookies zarejestrowanych na badanej witrynie jako wskaźnik estymujący liczbę użytkowników odwiedzających badaną witrynę. Niestety, wskaźnik ten z przyczyn obiektywnych nie uwzględnia zjawiska „kasowalności cookies” polegającego na tym, że pewna liczba użytkowników co pewien czas (świadomie, bądź przypadkowo) kasuje na swoich komputerach zarejestrowane cookies i w ten sposób widziana jest przez systemy site-centric wiele razy w badanym okresie. Zjawisko „kasowalności cookies” ma decydujący wpływ na to, że liczba rejestrowanych cookies (zwłaszcza dla dłuższych okresów czasu) jest dużo większa, niż rzeczywista liczba użytkowników badanej witryny. Przykładowo: internauta, który codziennie kasuje na swoim komputerze cookie, rozpoznawany jest przez systemy site-centric oraz systemy adserwerowe w okresie jednego miesiąca jako trzydziestu różnych użytkowników!

Estymując rzeczywistą liczbę użytkowników badanej witryny trzeba wziąć pod uwagę również dodatkowe czynniki, które wpływają na fakt, że liczba zarejestrowanych cookies nie odpowiada rzeczywistej liczbie użytkowników. Wpływ występowania zjawisk współkorzystania wielu osób z tego samego komputera i profilu użytkownika (tego samego cookie) oraz korzystania z internetu przez te same osoby na wielu komputerach jest aktualnie przedmiotem badań, choć już dziś można stwierdzić, że wykorzystanie informacji o liczebności i strukturze społeczno-demograficznej całej populacji internautów pozwala poprawnie estymować rzeczywistą liczbę internautów odwiedzających badane witryny internetowe.

3. Algorytm Real Users

W celu estymacji rzeczywistej liczby użytkowników badanej witryny w zadanym okresie czasu (oznaczymy ją przez U_W) w pierwszym etapie estymujemy taką liczbę cookies, która byłaby zarejestrowana na badanej witrynie, gdyby nie występowało zjawisko „kasowalności cookies”. Następnie obliczamy zasięg badanej witryny i ostatecznie liczbę użytkowników badanej witryny. W tym celu przeprowadzane są następujące obliczenia:

- ◆ Obliczamy liczbę odsłon wygenerowaną przez wszystkich użytkowników na badanej witrynie - oznaczmy tę liczbę przez O_W .
- ◆ Następnie obliczamy liczbę tych cookies, co do których mamy pewność, że istniały przez cały badany okres (cookie, które na pewno istniały przez cały badany okres to takie cookie, które na pewno istniały przez badany okres i po badany okres) - oznaczmy - oznaczmy tę liczbę przez C_D .
- ◆ Obliczamy liczbę odsłon wygenerowanych przez wybrane w punkcie b) cookies - oznaczmy ją przez O_D .
- ◆ Obliczamy liczbę cookies, jaka zarejestrowana byłaby na badanej witrynie, gdyby nie występowało zjawisko „kasowalności cookies” zgodnie ze wzorem:
$$C_W = (O_W/O_D) * C_D$$
- ◆ W analogiczny sposób (przyjmując w miejsce badanej witryny zbiór wszystkich badanych witryn) obliczamy liczbę cookies, która byłaby zarejestrowana na wszystkich badanych witrynach, gdyby nie występowało zjawisko „kasowalności cookies” - oznaczmy ją przez C_P .

- ◆ Obliczamy względny zasięg badanej witryny zgodnie ze wzorem: $Z_W = C_W / C_P$.
- ◆ Jeżeli przez P oznaczymy wielkość populacji internautów w badanym okresie, to liczbę użytkowników badanej witryny w zadanym okresie czasu obliczymy zgodnie ze wzorem: $U_W = Z_W * P$.

Do obliczeń bierzemy pod uwagę tylko te odsłony (i w konsekwencji również cookies), które zostały wygenerowane z komputerów z numerami IP pochodzącymi z obszaru Polski.

Podstawą powyższych obliczeń są następujące założenia:

- ◆ Istnieje możliwość wybrania grupy cookies, o której mowa w punkcie b) powyżej, tzn. istnieje możliwość monitorowania aktywności cookies nie tylko na badanej witrynie ale na jak największej liczbie witryn (lub najlepiej w całym internecie), jak też nie tylko w badanym okresie, ale również przed i po badanym okresie (dla okresu badania jakim jest jeden miesiąc najlepiej jest badać aktywność co najmniej miesiąc przed i miesiąc po). Jedną z konsekwencji takiego założenia jest naturalne opóźnienie w publikacji wyników badania – konieczny jest okres, w którym weryfikowany jest fakt przynależności cookies do wybranej w punkcie b) grupy cookies. Założenie to sprowadza się do technicznej możliwości monitorowania aktywności cookies i jest spełnione dzięki istnieniu badania gemiusTraffic.
- ◆ Wybrana w punkcie b) grupa cookies stanowi reprezentatywną grupę dla wszystkich cookies, zwłaszcza jeśli chodzi o średnią liczbę generowanych odsłon. Dla wszystkich dotychczasowych przypadków, założenie to zostało pozytywnie zweryfikowane analizami porównawczymi wszystkich możliwych charakterystyk behawioralnych wyłonionej w ten sposób grupy cookies w stosunku do grupy pozostałych cookies (analizy parametrów generowanych odsłon i wizyt, analizy częstotliwości i częstości odsłon i wizyt w dłuższych okresach, analizy geolokalizacyjne, analizy parametrów systemowych komputerów i przeglądarek itp.).
- ◆ Proporcja generowanej na wszystkich badanych witrynach przez wybraną w punkcie b) grupę cookies liczby odsłon oraz liczby wszystkich odsłon na wszystkich badanych witrynach jest identyczna jak proporcja generowanej na wszystkich witrynach (w tym nie badanych) przez tę samą grupę cookies liczby odsłon oraz liczby wszystkich odsłon na wszystkich witrynach. Hipoteza ta została pozytywnie zweryfikowana na podstawie analiz wyników badań panelowych, w których

brano pod uwagę proporcje odsłon generowanych na witrynach monitorowanych i niemonitorowanych przez system site-centric zarówno przez osoby kasujące jak i niekasujące pliki cookies, które przynajmniej raz odwiedziły co najmniej jedną z badanych witryn.

- ◆ Zasięg względny Z_w badanych witryn obliczony w zadanym okresie czasu na podstawie wybranej grupy cookies dobrze estymuje zasięg wśród użytkowników badanych witryn. Hipoteza ta jest obecnie przedmiotem dogłębnych badań, choć wszystkie dotychczasowe wyniki analiz pozwalają uznać tę hipotezę za prawdziwą.
- ◆ Wielkość populacji internautów P dobrze estymuje liczbę wszystkich internautów odwiedzających w badanym okresie wszystkie badane witryny.

4. Podsumowanie

Opisany powyżej algorytm real users pozwala oszacować rzeczywistą liczbę użytkowników badanych witryn internetowych. Uwzględnia on znane czynniki mogące potencjalnie zaburzać otrzymywane wyniki, w tym: zjawisko współkorzystania z jednego komputera przez wiele osób, korzystania przez jedną osobę z wielu komputerów/profilu oraz - co najważniejsze - zjawisko „kasowalności cookies”.