ZARZĄDZANIE MEDIAMI
Tom 8(3)2020, s. 243–257
doi:10.4467/23540214ZM.20.034.12052
www.ejournals.eu/ZM

Remigiusz Żulicki

Katedra Socjologii Kultury, Wydział Ekonomiczno-Socjologiczny Uniwersytetu Łódzkiego
remigiuszzulicki@gmail.com
ORCID: 0000-0003-2624-2422

Michał Żytomirski

Katedra Informatologii i Bibliologii, Wydział Filologiczny Uniwersytetu Łódzkiego
ORCID: 0000-0003-1400-9949
michal.zytomirski@gmail.com

PRÓBA WYPRACOWANIA METODOLOGII POMIARU BANIEK FILTRUJĄCYCH W WYSZUKIWARCE GOOGLE

Abstract

AN ATTEMPT TO DEVELOP A METHODOLOGY OF MEASURING FILTER BUBBLES IN GOOGLE SEARCH ENGINE

The authors proposed a method of quantitative measurement of the phenomenon of filter bubbles in Google search engine. Firstly, he topic was taken up because of the lack of such a method in the academic world and the proposal of a competitor to Google, DuckDuckGo. Secondly, because of the social consequences of the phenomenon of filter bubbles raised by activists and researchers. The aim was to test the used methodology of quantitative measurement of the phenomenon of information bubbles and to refine it. A pilot survey was conducted on a homogeneous group in an experimental scheme. It consisted in comparing sets of search results (SSR) in the normal mode of a web browser, logged into a Google account with SSR in the private (incognito) mode logged out of that account. One political search term was used – "Paweł Adamowicz". The SSR was treated as a sequence of characters, a string, and compared in two web browser's modes using the optimal string alignment distance method. The open sourced data and code allow readers to trace and reproduce the analyses made. The results do not indicate that the differences in the SSR are influenced neither by the world view in terms of liberalism/conservatism nor by the attitude to control the privacy of the Internet user. The influence was noted for the degree of personalization of the SSR in the normal web browser mode.

Key words: filter bubble, search engine, Google, quantitative research methods

Wstęp

Rozwój technologii cyfrowych wpływa na sposoby funkcjonowania całych społeczeństw i jednostek w nich funkcjonujących. Zmiany są widoczne zarówno na płaszczyznach wykorzystywania urządzeń cyfrowych (laptopy, smartfony) w celach zawodowych, edukacyjnych/naukowych, jak i prywatnych. Ważnym aspektem funkcjonowania urządzeń cyfrowych jest generowanie przez ich użytkowników cyfrowych śladów, czyli:

> (…) zmian w kodzie binarnym systemu teleinformatycznego, a także urządzenia cyfrowego zdolnego do przetwarzania, wysyłania, gromadzenia pakietów danych, będących wynikiem ingerencji zewnętrznej (fizycznej) bądź wewnętrznej (zdalnej) (Kasprzak, 2015).

Przykładowa typologia danych cyfrowych, opracowana na podstawie analizy narzędzi Google Analytics, określa, że dane cyfrowe mogą odnosić się m.in. do danych:

- statystycznych, np. częstotliwość wizyt użytkowników na danych stronach;
- geograficznych, odnoszących się do lokalizacji użytkowników wykorzystujących określone narzędzia/technologie cyfrowe;
- behawioralnych, sposobów wykorzystywania urządzeń/technologii cyfrowych przez użytkowników;
- technologicznych, określających, jakimi urządzeniami posługują się użytkownicy.

Wykorzystywanie urządzeń cyfrowych na globalną skalę oraz fakt indeksowania zachowań ich użytkowników tworzy dotychczas niespotykane w historii ludzkości możliwości. Twierdzi się, że dzięki poznaniu i przeanalizowaniu cyfrowych śladów użytkowników cyberprzestrzeni możliwe jest określenie ich cech psychologicznych, a co za tym idzie – dopasowywanie treści cyfrowych do ich umiejętności, kompetencji, potrzeb i oczekiwań (por. Kosinski, Stillwell, Graepel, 2013). Oznacza to, że treści są personalnie dopasowywane do użytkowników, co z kolei zamyka użytkowników Sieci w bańkach filtrujących.

Charakterystyka baniek filtrujących

Termin „bańki filtrujące" (*filter bubble*) w odniesieniu do środowiska cyfrowego został zaproponowany przez Eliego Parisera (2011), który określa to zjawisko jako oddźwięk pracy algorytmów dopasowujących treść do potrzeb użytkowników. Zwraca on szczególną uwagę na trzy cechy baniek filtrujących:

- użytkownicy tworzą swoje indywidualne bańki poprzez dokonywanie czynności w środowisku cyfrowym;
- nie mają oni wglądu w swoje bańki – nie są w stanie stwierdzić, jak są indeksowani, np. w środowisku Google (kryteria doboru są narzucone). Jed-

nocześnie użytkownicy nie mają zwyczaju porównywania swoich wyników wyszukiwania z innymi odbiorcami podobnych treści;
- użytkownicy nie mają wyboru, nie można pozbyć się swojej bańki.

Zjawisko baniek filtrujących, według dostawców treści takich jak Google, ma skutkować dostarczaniem spersonalizowanych treści do odbiorców. Zgodnie z tą koncepcją jest to traktowane jako korzyść, bowiem np. czas wyszukiwania/przeglądania Internetu w celu dotarcia do optymalnych dla danej osoby wyników będzie obniżony. Jednocześnie bańki filtrujące określane są jako „ograniczenia przestrzeni poznawczej", które skutkują obniżeniem kreatywności i umiejętności uczenia się użytkowników (Rogers, 2018). To ograniczenie jest immanentną cechą personalizacji treści, ta bowiem nastawiona jest na komfort poznawczy użytkownika – jednostka w wyniku personalizacji ma otrzymywać treści jak najbardziej zbliżone do swoich poglądów, preferencji i potrzeb (Szpunar, 2018, s. 193–195).

W powyższej argumentacji, jak i ogólnie w naszym projekcie koncentrujemy się na bańkach filtrujących jako na zjawisku technologicznym – działaniu systemów personalizujących treści, któremu przypisywane są (raczej negatywne) konsekwencje społeczne. Zjawisko to ujmowane bywa jednak szerzej. W przekonujący sposób opisano bańki filtrujące jako składające się z trzech komponentów: indywidualnego, społecznego i technologicznego (Geschke, Lorenz, Holtz, 2019, s. 133–134). Ze względu na cele niniejszego badania świadomie koncentrujemy się na aspekcie technologicznym zjawiska, pamiętając o rzekomych konsekwencjach społecznych.

Inspiracją do podjęcia przez nas tematu był tekst opublikowany przez firmę DuckDuckGo (2018). Firma ta oferuje wyszukiwarkę internetową niepobierającą danych użytkowników. W przedstawionym badaniu bańki filtrujące ujęto jako „manipulowanie wynikami wyszukiwania na podstawie danych personalnych". DuckDuckGo jako firma komercyjna, która konkuruje z Google w sposób bezpośredni, nie jest obiektywnym źródłem informacji o badanym obszarze.

Badanie DuckDuckGo zostało przeprowadzone w czerwcu 2018 roku na grupie złożonej z ochotników – 87 obywateli USA. Uczestnicy zostali poproszeni o wyszukanie w wyszukiwarce Google następujących fraz: „gun control", „immigration", „vaccinations". Respondenci mieli dokonać tych czynności dokładnie 14 czerwca 2018 roku o godzinie 21:00, następnie musieli przesłać informacje o wynikach. W badaniu wykorzystano wyniki wyszukiwania pochodzące od wspomnianych 87 uczestników (76 – wyszukiwanie na komputerze, 11 na urządzeniu mobilnym). Podczas analizy wyników wyszukiwania uwzględniono tylko domeny najwyższego poziomu witryn:

> Wyniki wyszukiwania Google zwykle zawierają dziesięć bezpłatnych linków. Chociaż kolejność tych linków jest naprawdę ważna (tzn. link nr 1 uzyskuje ~ 40% kliknięć, link nr 2 ~ 20%, link nr 3 ~ 10% itd.) (tamże).

Określono, że podczas wyszukiwania frazy „gun control" w przypadku 76 respondentów wyodrębniono 62 zestawy wyników wyszukiwania – autorzy badania upatrują w tym dowód na występowanie personalizacji/filtrów wyszukiwawczych

w badanym środowisku. Dodatkowo badacze stwierdzili, że wyszukiwanie w trybie prywatnym z uprzednim wylogowaniem się z konta Google nie zapewnia ochrony przed bańkami filtrującymi.

Różnorodne badania poświęcone zjawisku baniek filtrujących wskazują głównie na społeczne zagrożenia, które są wywoływane poprzez personalizujące rozwiązania cyfrowe. W Polsce zjawisko jest, w naszej ocenie, słabo rozpoznane empirycznie. Pojawiły się wartościowe teksty przeglądowe lub polemiczne (Furman, 2018; Malinowski, 2016; Szpunar, 2018; Szpyt-Wiktorowska, Wiktorowski, 2018), a także publikacje badawcze (Książek, 2019; Matuszewski, 2018; Popiołek, Sroka, 2019; Szpyt-Wiktorowska, 2018), jednak nie było dotychczas prób wypracowania metodologii ilościowego pomiaru zjawiska.

Realizowane są interesujące badania i analizy ilościowe baniek informacyjnych (np. Bakshy, Messing, Adamic, 2015; Möller, Trilling, Helberger, van Es, 2018; Nguyen, Hui, Harper, Terveen, Konstan, 2014). W pierwszym z wymienionych stwierdzono, że na Facebooku działania ludzkie bardziej niż działanie algorytmów ogranicza ekspozycja użytkowników na treści niezgodne z ich poglądami. W drugim w toku symulacji obliczono, że zarówno spersonalizowane, jak i niespersonalizowane algorytmy selekcji treści prowadzą do podobnej różnorodności w tematach wskazywanych użytkownikom treści. W trzecim – na podstawie rozłożonego w czasie badania użytkowników systemu rekomendującego filmy – wskazano, że osoby częściej wybierające filmy rekomendowane oglądały bardziej zróżnicowane treści niż osoby wybierające filmy spoza rekomendacji, przy czym z czasem zróżnicowanie dla wszystkich użytkowników zmniejszało się. Istnieje więcej tego rodzaju wysokiej jakości badań, z których każde daje inny wgląd w zjawisko baniek informacyjnych, niemniej nie odnajdujemy w literaturze propozycji wypracowania metody, także metody ograniczonej do jednego systemu selekcji (np. wyszukiwarki Google).

Metody co do zasady zbliżone do rozwiązania proponowanego przez nas, a wywodzące się z pomysłu Eliego Parisera (2011) – czyli porównania wyników wyszukiwania w wyszukiwarce Google dla tego samego hasła przez różne osoby – były stosowane przez Tomasza Książka (2019) oraz Joannę Szpyt-Wiktorowską (2018). Pierwsza z wymienionych pozycji (Książek, 2019) zawiera powierzchowne analizy (brak testów statystycznych, a jedynie podsumowania zebranych w ankiecie odpowiedzi) i niedoskonałości merytoryczne (np. prezentacje rozkładów na mało czytelnych, trójwymiarowych wykresach kołowych, opisywanie różnic o X punktów procentowych jako różnic o X procent), zatem w naszej ocenie nie należy traktować tej publikacji jako wzoru w kontekście metod badań czy analiz baniek informacyjnych. Druga z propozycji (Szpyt-Wiktorowska, 2018) opiera się na cennej poznawczo idei porównania źródeł internetowych, których użytkowanie deklarują respondenci, z ich wynikami wyszukiwania w Google. Autorka zawarła także ważne wnioski (m.in. podkreślając rolę działań SEO dla pozycjonowania danego źródła na liście wyników wyszukiwania, a w konsekwencji dostępu do tego źródła bez względu na to, jak personalizowane są wyniki). W przypadku metodologii tego tekstu brakuje jednak opisu sposobu doboru respondentów.

Środowisko Google przez swój globalny charakter tworzy jedno z największych źródeł informacji o działaniach ludzi w sieci, co w naszej ocenie stawia je jako najciekawsze pole badawcze w odniesieniu do baniek filtrujących. Zakres danych o użytkownikach, które podlegają analizom, to (Google, 2019):

1. Pole:
   a. Aplikacje
   b. Przeglądarki
   c. Urządzenia
2. Zakres:
   a. Aktywność
      i. Wyszukiwane hasła
      ii. Oglądane filmy
      iii. Wyświetlenia reklam i treści oraz interakcja z nimi
      iv. Informacje związane z głosem i dźwiękiem, gdy używasz funkcji audio
      v. Zakupy
      vi. Osoby, z którymi się kontaktujesz lub którym udostępniasz materiały
      vii. Aktywność w witrynach i aplikacjach innych firm, które korzystają z naszych usług
      viii. Historia przeglądania w Chrome synchronizowana z kontem Google
   b. Informacje o lokalizacji
      i. GPS
      ii. IP
      iii. Dane z czujników urządzeń
      iv. Informacje o rzeczach w pobliżu urządzenia, na przykład punktów dostępu do sieci Wi-Fi, stacji bazowych sieci komórkowych i urządzeń z włączoną obsługą Bluetooth
   c. Dane te zbierane są w celach:
      i. Świadczenia usług
      ii. Utrzymywania i ulepszania usług
      iii. Opracowywania nowych usług
      iv. Zapewniania spersonalizowanych usług, w tym treści i reklam
      v. Pomiarowych
      vi. Kontaktowych

Metodologia badań własnych

Zrealizowaliśmy badanie pilotażowe na homogenicznej grupie badawczej 105 studentek i studentów Wydziału Zarządzania Uniwersytetu Łódzkiego. Celem było sprawdzenie wykorzystywanej metodologii ilościowego pomiaru zjawiska baniek informacyjnych i dopracowanie jej na potrzeby badania właściwego, które odbędzie się w roku akademickim 2019/2020.

Nasza metodologia, inspirowana wspomnianymi badaniami firmy DuckDuckGo, zakłada – po pierwsze – badanie homogenicznej, celowo dobranej grupy. W ten sposób chcemy kontrolować jak najwięcej nierejestrowanych źródeł zmienności wyników wyszukiwania. Po drugie – eksperymentalny charakter badania, gdzie każdy/a z uczestników/uczestniczek wyszukuje w wyszukiwarce Google dwukrotnie to samo hasło – w trybie normalnym z zalogowaniem na konto Google[1] i w trybie prywatnym, bez zalogowania. Po trzecie – poszukiwanie źródeł zmienności wyników wyszukiwania w wymiarach: światopoglądu (konserwatyzm/liberalizm) oraz nastawienia do kontroli prywatności użytkownika Internetu. Tym samym nasza metodologia ma pozwalać na:

- opis rozkładu wyników wyszukiwania w grupie (skoro użytkownicy nie mają w zwyczaju porównywania swoich wyników wyszukiwania z innymi odbiorcami podobnych treści, to możemy porównać je w ten sposób);
- na sprawdzenie, czy korzystanie z trybu prywatnego przeglądarki wraz z wylogowaniem z konta Google może być strategią wyjścia poza bańki informacyjne, a także na próbę określenia źródeł zmienności wyników wyszukiwania i źródeł zmienności różnic w wynikach wyszukiwania dla tej samej osoby.

Zdecydowaliśmy się na wyszukiwanie hasła „Paweł Adamowicz”[2], zainspirowani stwierdzeniem „bańka informacyjna jest szczególnie zgubna w przypadku poszukiwania treści politycznych” (DuckDuckGo, 2018). W tym samym kontekście to, iż w demokracji media powinny umożliwiać obywatelom styczność z różnorodnymi źródłami informacji oraz różnorodnymi opiniami, podnosili zarówno badacze, jak i np. Rada Europy (Möller, Trilling, Helberger, van Es, 2018, s. 959).

Wyniki badań własnych

Rozkład wyników wyszukiwania

Osoby uczestniczące w naszym badaniu zostały poproszone o podanie dokładnego adresu url dla pięciu pierwszych wyników wyszukiwania hasła „Paweł Adamowicz” w wyszukiwarce Google. Zebraliśmy po pięć pierwszych wyników dla trybu normalnego (przy zalogowaniu na konto Google) i prywatnego przeglądarki internetowej.

[1] Odpowiedzi zbieraliśmy poprzez formularz Google zamieszczony na naszej stronie http://www.googlebubble.uni.lodz.pl/, umożliwiając respondentom uczestnictwo tylko po zalogowaniu się na konto Google. Na potrzeby badania założyliśmy konto myfilterbubble@gmail.com, korzystając z uczelnianego PC i świeżo zainstalowanej przeglądarki internetowej Opera. Miało to na celu uniemożliwienie łączenia danych z naszych prywatnych kont Google z niniejszym badaniem

[2] Paweł Adamowicz był prezydentem Gdańska w latach 1998–2019, zmarł 14.02.2019 r. po zamachu na jego osobę podczas finału Wielkiej Orkiestry Świątecznej Pomocy. Był politykiem znanym, kontrowersyjnym, a jego działalność w ostatnich latach oraz tragiczną śmierć szeroko komentowano w polskich mediach tradycyjnych i Internecie (Szczęśniak, 2019).

Uzyskaliśmy odpowiedzi od 105 osób, jednak po wyczyszczeniu tekstów, wprowadzonych jako adresy url, pozostawiliśmy odpowiedzi 73 osób[3]. Zatem łącznie przeanalizowaliśmy 730 adresów url – po 10 dla jednego uczestnika. Każdy adres url skróciliśmy do domeny głównej. Każdej domenie przypisaliśmy znak (literę albo cyfrę), na potrzebę obliczenia różnicy między zestawami wyników wyszukiwania. Łącznie w obu trybach wyszukiwania uczestnicy naszego badania uzyskali 28 różnych domen (tabela 1).

Tabela 1. Liczebność domen głównych

Domena główna	Znak przypisany	Liczebność [łączna liczba wystąpień domeny, n = 730]
pl.wikipedia.org	I	138
www.facebook.com	V	138
www.wprost.pl	3	87
adamowicz.pl	A	77
twitter.com	M	51
wiadomosci.gazeta.pl	O	41
bialystok.onet.pl	B	37
trojmiasto.wyborcza.pl	L	31
www.youtube.com	4	24
wiadomosci.wp.pl	R	21
www.trojmiasto.pl	1	20
www.tvn24.pl	2	14
dziennikbaltycki.pl	C	6
fakty.interia.pl	D	6
www.fakt.pl	W	6
poranny.pl	J	5
natemat.pl	G	4
wiadomosci.radiozet.pl	Q	4
gdansk.naszemiasto.pl	E	3
wpolityce.pl	T	3
www.bstok.pl	U	3
www.tokfm.pl	Z	3
wmeritum.pl	S	2

[3] Podstawowym powodem wykluczenia uczestników badania było umieszczenie tego samego adresu url na więcej niż jednej z pięciu pozycji wyników wyszukiwania. Dokładną, powtarzalną procedurę przygotowania danych do analizy można odtworzyć, uruchamiając nasze skrypty w języku R. Każda czytająca osoba może nie tylko odtworzyć nasze wyniki, ale także prześledzić każdy kolejny krok operacji na danych, wszystko w środowisku *open source*. Dane i kod są dostępne na https://github.com/zremek/google_filter_bubble. Jest to zatem pierwsze w Polsce ilościowe badanie baniek informacyjnych, przygotowane według najwyższych standardów powtarzalności badań naukowych (*reproducible research*).

Domena główna	Znak przypisany	Liczebność [łączna liczba wystąpień domeny, n = 730]
www.gdansk.pl	X	2
trojmiasto.onet.pl	K	1
tysol.pl	N	1
wiadomosci.onet.pl	P	1
www.rp.pl	Y	1

Źródło: badania własne.

Domeny te ułożone były w różnej kolejności, od pierwszej do piątej pozycji na liście wyników wyszukiwania. W podziale na tryb przeglądarki różnice w kolejności były niewielkie dla najczęściej pojawiających się domen (pierwsza pozycja – pl.wikipedia.org; druga – www.facebook.com; trzecia – www.wprost.pl). Niektóre domeny (www.gdansk.pl; trojmiasto.onet.pl; wiadomosci.onet.pl) pojawiły się tylko w trybie normalnym przeglądarki, inne (tysol.pl; www.rp.pl) tylko w trybie prywatnym (rysunek 1).

Jako zestaw wyników wyszukiwania (dalej: ZWW) traktujemy pięć domen, zapisanych w kolejności wyświetlania w wynikach wyszukiwarki Google dla jednego uczestnika badania. Przykładowo, najczęściej pojawiający się ZWW (8 na 73 osoby w trybie prywatnym; 5 w trybie normalnym) IVB31 oznacza, że osoba w odpowiedzi na zapytanie „Paweł Adamowicz" w wyszukiwarce Google uzyskała następujące domeny w podanej kolejności (tabela 2).

Tabela 2. Przykładowy ZWW – IVB31

Znak przypisany	Domena główna
I	pl.wikipedia.org
V	www.facebook.com
B	bialystok.onet.pl
3	www.wprost.pl
1	www.trojmiasto.pl

Źródło: badania własne.

Łącznie w dwóch trybach przeglądarki dla 73 osób uzyskano 73 różne ZWW – 73 różne sposoby, na które uczestnicy zobaczyli w wynikach 28 różnych domen. Nie oznacza to jednak, że każdy uczestnik zobaczył inny zestaw, bowiem w trybie normalnym było to 50 ZWW, a w prywatnym 51. Niemniej ZWW powtarzają się rzadko. W trybie normalnym niepowtarzalny ZWW uzyskało 37 z 73 osób. W trybie prywatnym było to 38 osób. Zatem większość uczestników badania widziała ZWW, którego nie widziała żadna inna osoba. To właśnie określane jest jako bańki informacyjne. Żaden ZWW nie pojawił się więcej, niż osiem razy – wymienione IVB31 w trybie prywatnym. Pojedyncze ZWW występowały tylko w jednym trybie (rysunek 2).

Rysunek 1. Rozkład domen głównych

Rysunek 2. Rozkład ZWW

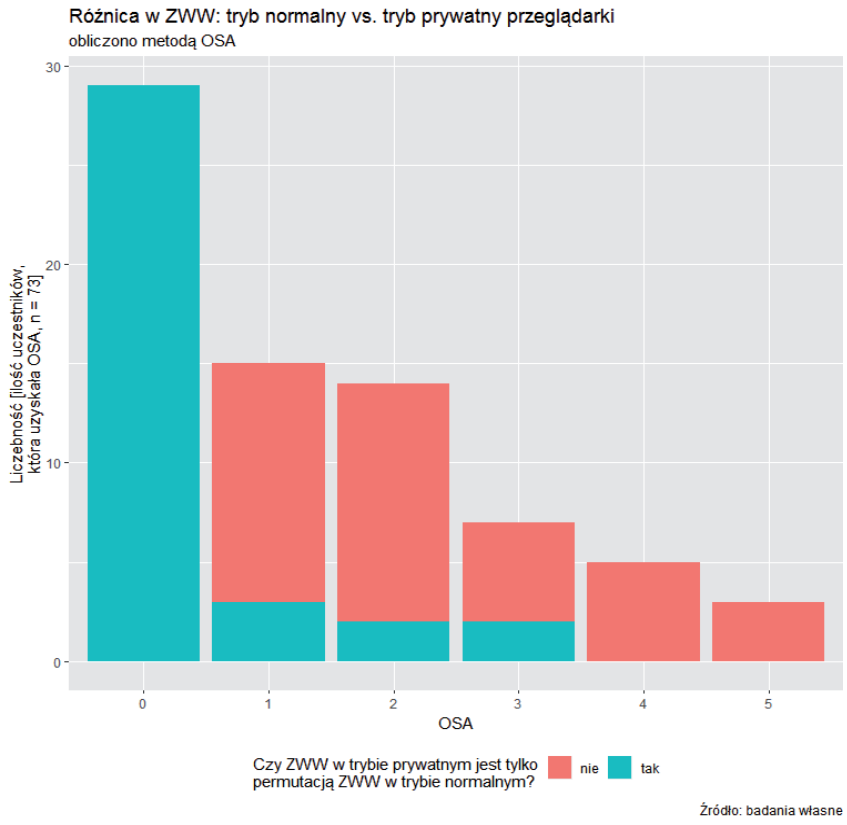
Źródło: badania własne

Różnica w zestawach wyników wyszukiwania: tryb normalny vs. tryb prywatny przeglądarki

Różnica wyrażona jest w liczbach całkowitych od 0 do 5. Zero oznacza, że wyniki były identyczne. Wynik wyższy oznacza, że wystąpiły różnice.

Różnicę obliczyliśmy metodą *optimal string alignment distance* (dalej: OSA) (van der Loo, 2014). Metoda ta daje wyniki identyczne z metodą zastosowaną w badaniu DuckDuckGo (2018).

Rysunek 3. Różnica w ZWW obliczona metodą OSA



Do obliczenia OSA każdy ZWW jest traktowany jako ciąg pięciu znaków (por. tabela 1, rysunek 2). Brane są pod uwagę cztery scenariusze różnic między ZWW jako ciągami znaków (van der Loo, 2014, s. 114):

- substytucja (*substitution of a character*), np. 'foo'→'boo';
- usunięcie (*deletion of a character*), np. 'foo'→'oo';
- umieszczenie (*insertion of a character*), np. 'foo'→'floo';
- transpozycja (*transposition of two adjacent characters*), np. 'foo'→'ofo',

z których każdy ma wagę 1.

Naszym zdaniem w kontekście wyszukiwania informacji w Internecie scenariusz, w którym pomiędzy ZWW występuje tylko transpozycja, różni się jakościowo od pozostałych. Choć kolejność wyników wyszukiwania ma znaczenie dla ich klikalności, to gdy wynik nie pojawi się w ogóle, użytkownik na pewno na niego nie kliknie. Zatem za pomocą przeszukania możliwych permutacji sprawdziliśmy, czy ZWW uzyskany w trybie prywatnym składa się z tych samych znaków (czyli domen), co w trybie normalnym. Uczestnicy badania najczęściej (29 z 73) zobaczyli ten sam ZWW (OSA = 0) w obu trybach przeglądarki (rysunek 3). Mediana OSA = 1, Q1 = 0, Q3 = 2, średnia = 1,356. Jeszcze częściej ZWW w trybie prywatnym był tylko permutacją ZWW w trybie normalnym (36 z 73 osób). Nie więcej niż jedną różnicę w ZWW zobaczyły 44 osoby.

Wyniki nasze są bardzo zbliżone wyników do uzyskanych w badaniu firmy DuckDuckGo (2018). Firma uzyskała średnie OSA bliskie naszej = 1,356 dla dwóch z trzech wyszukiwań (*gun control* = 1,03; *immigration* = 1,38; *vaccinations* = 2,23). Jej zdaniem takie wyniki wskazują na utrzymanie bańki informacyjnej w trybie prywatnym przeglądarki (tamże).

Testowanie hipotez

Na etapie projektowania badania postawiliśmy robocze hipotezy, że wynik OSA mogą różnicować zmienne:

- rok urodzenia osoby uczestniczącej w badaniu
- płeć
- rodzaj studiów
- umiejętność korzystania z trybu prywatnego przeglądarki
- światopogląd w wymiarze liberalizm/konserwatyzm (DuckDuckGo, 2018)
- nastawienie do kontroli prywatności użytkownika Internetu

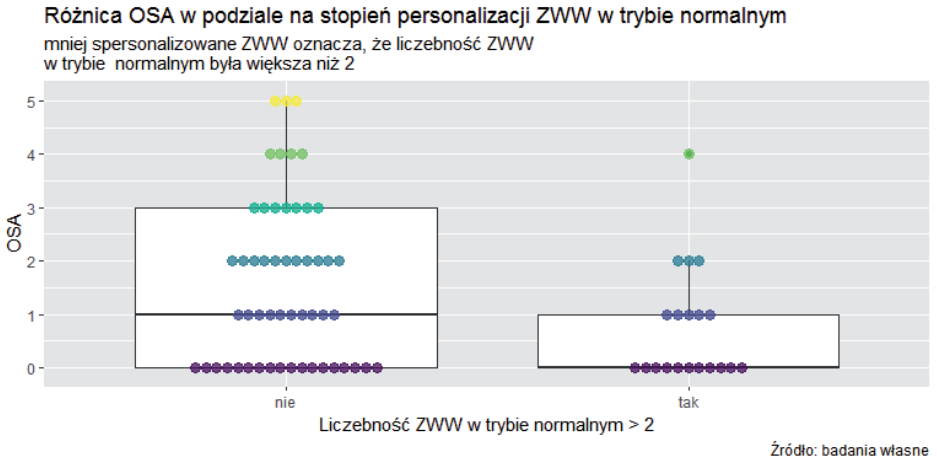
Ze względu na specyfikę realizacji badania sprawdziliśmy także, czy wynik OSA różnicował czas odpowiedzi:

- miesiąc i dzień
- miesiąc

We wszystkich wyżej wymienionych przypadkach różnice nie były istotne statystycznie.

W toku analizy postawiliśmy zatem kolejną hipotezę, że wyniki OSA może różnicować stopień personalizacji ZWW w trybie normalnym. Jako wskaźnik tego stopnia personalizacji przyjęliśmy liczbę wystąpień ZWW (od 0 do 5, por. rysunek 2), za punkt odcięcia przyjmując dwa wystąpienia. Dla mniej spersonalizowanych ZWW (czyli takich, które pojawiły się więcej niż dwojgu respondentom, Mediana OSA = 0) różnica OSA była istotnie niższa niż dla bardziej spersonalizowanych ZWW (które pojawiły się tylko pojedynczym respondentom lub dwojgu, Mediana OSA = 1), $W = 697{,}5$; $p < 0{,}05$; $r = -0{,}252$. Wartość r wskazuje na wielkość efektu (*effect size*) między małą a średnią (Field, Miles, Field, 2012, s. 666). Różnicę tę przedstawiamy na rysunku 4.

Rysunek 4. Różnica OSA a stopień personalizacji ZWW



Źródło: badania własne

Dyskusja

Zaproponowaną metodologię ilościowego pomiaru baniek informacyjnych uznajemy za właściwą dla badania rozkładu wyników wyszukiwania. Za właściwe uznajemy badanie na podstawie jednego hasła wyszukiwania, na homogenicznej grupie użytkowników, przy kontroli miejsca i czasu. Naszą metodologię uznajemy także za właściwą w sensie mierzenia różnic w wynikach wyszukiwania metodą OSA, ale jedynie w zrealizowanym tu schemacie eksperymentalnym, gdzie porównujemy parę wyników tej samej użytkowniczki/użytkownika. Będziemy dążyć do wypracowania metody pomiaru różnic nie par wyników, ale wyników w grupie użytkowników, tak by móc za pomocą jednoliczbowej statystyki ocenić zróżnicowanie np. dla wyników w trybie normalnym czy prywatnym. Nie tyle za niewłaściwe, ile nieudane uznajemy próby określenia źródeł zmienności wyników w wymiarach światopoglądu i nastawienia do kontroli prywatności użytkownika Internetu. Nie da się ocenić, czy niewłaściwie dobraliśmy źródła, czy sposób pomiaru zmiennych.

W odniesieniu do realizacji badania krytycznie oceniamy zbieranie adresów url przez formularz. Kopiowanie i wklejanie było dla uczestniczących w badaniu czasochłonne, niewygodne i prowadziło do obniżenia jakości zebranych danych. Dążymy do wypracowania szybkiej i wygodnej metody ekstrakcji adresów url ze źródła strony z wynikami wyszukiwania. Pozwoli ona na zapisanie informacji z całej pierwszej strony wyników wyszukiwania, dzięki czemu ZWW nie będzie ograniczony do pierwszych pięciu pozycji.

W odniesieniu do dalszych badań nad bańkami informacyjnymi z zastosowaniem naszego ujęcia sądzimy, że należałoby zająć się problematyką treści wyników wyszukiwania, w niniejszej pracy celowo pominiętą. Być może obok stosowanych

już rozwiązań typu *text mining* (por. Möller, Trilling, Helberger, van Es, 2018) warto byłoby sięgnąć po metody jakościowej analizy treści.

Nasze wyniki wskazują, że korzystanie z wyszukiwarki Google w trybie prywatnym przeglądarki internetowej i bez zalogowania na konto Google nie może być traktowane jako skuteczny sposób na wydostanie się z własnej bańki filtrującej. W kolejnych badaniach rozważamy wykonanie porównań wyników z różnych wyszukiwarek internetowych, np. poza dominującą Google wyszukiwarek DuckDuckGo, Qwant, Bing czy Yandex.

Nie chcemy przy tym stawiać się w roli aktywistów, uznających bańki informacyjne za jednoznacznie „złe", np. w sensie zagrożenia dla demokracji. Ze względu na złożony, co najmniej trojaki (indywidualno-społeczno-technologiczny) charakter badanego zjawiska zależy nam na wypracowaniu możliwie powtarzalnej, systematycznej metody badania przynajmniej jednego aspektu zjawiska baniek informacyjnych.

Podziękowania

Dziękujemy wszystkim uczestniczącym w naszym badaniu studentkom i studentom Wydziału Zarządzania UŁ.

Składamy podziękowanie za wsparcie w realizacji badania kadrze Wydziału Zarządzania UŁ. Badania realizowaliśmy początkowo w maju 2019 roku dzięki zgodzie prof. Tomasza Czapli i wsparciu organizacyjnym dr Olgi Dryni, a następnie w czerwcu 2019 roku dzięki pomocy prof. Anny Szychty, dra Grzegorza Skalskiego, dr Dominiki Kaczorowskiej-Spychalskiej, prof. Krystyny Iwińskiej-Knop.

Dziękujemy także za konsultacje metodologiczne dr Katarzynie Grzeszkiewicz-Radulskiej (Instytut Socjologii, Wydział Ekonomiczno-Socjologiczny UŁ).

Kierujemy również słowa wdzięczności do osób recenzujących artykuł za konstruktywną krytykę i cenne uwagi.

Bibliografia

Amrhein V., Greenland S., McShane B., *Scientists Rise Up against Statistical Significance*, „Nature" 2019, vol. 567 (7748), s. 305–307, https://doi.org/10.1038/d41586-019-00857-9.

Bakshy E., Messing S., Adamic L.A., *Exposure to Ideologically Diverse News and Opinion on Facebook*, „Science" 2015, vol. 348 (6239), s. 1130–1132, https://doi.org/10.1126/science.aaa1160.

DuckDuckGo, *Measuring the Filter Bubble: How Google Is Influencing What You Click*, 2018, https://spreadprivacy.com/google-filter-bubble-study/ (dostęp: 4.12.2018).

Field A., Miles J., Field Z., *Discovering Statistics Using R*, Sage Publications Ltd., London 2012.

Furman W., *Od pozornej wiedzy do komory pogłosowej i nadmiaru informacji. Krótki przegląd strachów medialnych*, „Zeszyty Prasoznawcze" 2018, t. 61, nr 2 (234), s. 201–208, https://doi.org/10.4467/22996362pz.18.014.9109.

Geschke D., Lorenz J., Holtz P., *The Triple-Filter Bubble: Using Agent-Based Modelling to Test a Meta-Theoretical Framework for the Emergence of Filter Bubbles and Echo Chambers*, „British Journal of Social Psychology" 2019, vol. 58, nr 1, s. 129–149, https://doi.org/10.1111/bjso.12286.

Google, *Polityka prywatności – prywatność i warunki*, 2019, https://policies.google.com/privacy?hl=pl (dostęp: 28.09.2019).

Kasprzak W., *Ślady cyfrowe: Studium prawno-kryminalistyczne*, Difin, Warszawa 2015.

Kosinski M., Stillwell D., Graepel T., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, „Proceedings of the National Academy of Sciences" 2013, vol. 110, nr 15, s. 5802–5805, https://doi.org/10.1073/pnas.1218772110.

Książek T., *Bańka filtrująca i błąd konfirmacji w świadomości użytkowników Internetu*, Stowarzyszenie Bibliotekarzy Polskich, Warszawa 2019.

Malhotra N.K., Kim S.S., Agarwal J., *Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model*, „Information Systems Research" 2004, vol. 15, nr 4, s. 336–355, https://doi.org/10.1287/isre.1040.0032.

Malinowski B., *How Does Facebook Traps Us in a Bubble: The Facebook's Content Filter Algorithm vs Filter Bubble Effect*, „Zarządzanie Mediami" 2016, t. 4, nr 1, s. 15–22, https://doi.org/10.4467/23540214ZM.15.002.5212.

Matuszewski P., *Wykorzystanie mediów informacyjnych w dyskusjach politycznych na Facebooku*, „Studia Medioznawcze" 2018, t. 1 (72), s. 27–42, http://studiamedioznawcze.pl/Numery/2018_1_72/matuszewski.pdf (dostęp: 21.09.2019).

Möller J., Trilling D., Helberger N., van Es B., *Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and their Impact on Content Diversity*, „Information, Communication & Society" 2018, vol. 21, nr 7, s. 959–977, https://doi.org/10.1080/1369118X.2018.1444076.

Nguyen T.T., Hui P.-M., Harper F.M., Terveen L., Konstan J.A., *Exploring the Filter Bubble*, [w:] *Proceedings of the 23rd International Conference on World Wide Web – WWW '14*, ACM Press, New York 2014, s. 677–686, https://doi.org/10.1145/2566486.2568012.

Popiołek M., Sroka K., *Bańka filtrująca i świadomość mechanizmów jej funkcjonowania wśród młodzieży – wyniki badania przeprowadzonego wśród gimnazjalistów*, „Zarządzanie Mediami" 2019, t. 7, nr 3, https://doi.org/10.4467/23540214ZM.19.011.11122.

Rogers R., *Aestheticizing Google Critique: A 20-Year Retrospective*, „Big Data & Society" 2018, vol. 5, nr 1, s. 1–13, https://doi.org/10.1177/2053951718768626.

Roguska B., *Charakterystyka poglądów potencjalnych elektoratów partyjnych*, Komunikat z badań Centrum Badania Opinii Społecznej (85), 2015, https://cbos.pl/SPISKOM.POL/2015/K_085_15.PDF (dostęp: 11.03.2019).

Szczęśniak A., *Materiałów na temat Adamowicza było w TVP prawie 1800 w 2018*, 2019, https://oko.press/materialow-oczerniajacych-adamowicza-bylo-w-tvp-ponad-100-pis-to-klamstwo-naprawde-telewizja-zajmowala-sie-adamowiczem-prawie-1800-razy/ (dostęp: 26.09.2019).

Szpunar M., *Koncepcja bańki filtrującej a hipernarcyzm nowych mediów*, „Zeszyty Prasoznawcze" 2018, t. 61, nr 2 (234), s. 191–200, https://doi.org/10.4467/22996362pz.18.013.9108.

Szpyt-Wiktorowska J., *Strategie mediów wobec baniek informacyjnych*, „Zarządzanie Mediami" 2018, t. 6, nr 2, s. 41–50, https://doi.org/10.4467/23540214zm.18.004.9026.

Szpyt-Wiktorowska J., Wiktorowski M., *Sfera publiczna i praktyka zarządzania mediami na przykładzie portalu internetowego*, „Zeszyty Prasoznawcze" 2018, t. 61, nr 1 (233), s. 81–95, https://doi.org/10.4467/22996362pz.18.006.8716.

van der Loo M.P.J., *The Stringdist Package for Approximate String Matching*, „The R Journal" 2014, vol. 6, nr 1, s. 111–122, https://doi.org/10.32614/rj-2014-011.