PRZEGLĄD BIBLIOTECZNY 2015 z. 2
PL ISSN 0033-202X

ADAM JACHIMCZYK
Instytut Dziennikarstwa i Informacji
Uniwersytet Jana Kochanowskiego w Kielcach
e-mail: Adam.Jachimczyk@ujk.edu.pl

MAGDALENA CHRAPEK
Instytut Matematyki
Uniwersytet Jana Kochanowskiego w Kielcach
e-mail: Magdalena.Chrapek@ujk.edu.pl

KATALOGI ARTYKUŁÓW (PRESELL PAGES) – ANALIZA STRON INTERNETOWYCH WYKORZYSTYWANYCH W SEO

Adam Jachimczyk jest adiunktem w Instytucie Dziennikarstwa i Informacji Uniwersytetu Jana Kochanowskiego w Kielcach. Jego zainteresowania naukowe obejmują zastosowanie technologii informatycznej w działalności informacyjnej. Opublikował m.in.: Polskie spisy bibliograficzne na World Wide Web. W: *Bibliografia: teoria, praktyka, dydaktyka*. Red. J. Woźniak-Kasperek i M. Ochmański. Warszawa, 2009; Format metadanych ONIX – ONline Information eXchange. W: *Szósta Ogólnokrajowa Narada Bibliografów Warszawa 23-24 października 2008*. Warszawa, 2010; Cenzura w wersji online. W: *Nie po myśli władzy: studia nad cenzurą i zakresem wolności słowa na ziemiach polskich od wieku XIX do czasów współczesnych.* Red. D. Degen, M. Żynda. Toruń, 2012; Obowiązki i kompetencje przedstawicieli zawodów informacyjnych. Perspektywa pracodawcy. *Przegląd Biblioteczny* 2013, z. 2; razem z: M. Olczak-Kardas: Recenzje użytkowników księgarni internetowej jako źródło informacji o książce i czytelnictwie. Komunikat z badań. *Rocznik Bibliologiczno-Prasoznawczy* 2012, t. 4, z. 1 (15). Od 2011 r. jest redaktorem naczelnym internetowego periodyku naukowego „iNFOTEZY".

Magdalena Chrapek jest starszym wykładowcą w Instytucie Matematyki Uniwersytetu Jana Kochanowskiego w Kielcach. Jej zainteresowania naukowe obejmują metody statystyczne i techniki data mining. Opublikowała m.in.: On the extremal behaviour of some stationa-

ry Markov sequence. *Demonstratio Mathematica* 2008, z. 31; jest współautorką m.in. następujących publikacji: W. Kryczka, K. Paluch, D. Zarębska-Michaluk, Progress of liver disease in chronic hepatitis C patients who failed antiviral therapy. *Medical Science Monitor* 2003, Aug, 9 Suppl 3; M. Stachura, B. Wodecka, Estymacja indeksu ekstremalnego w oparciu o k-te wartości rekordowe – sugestia poprawy jakości estymacji. W: *Prace Naukowe. Metody matematyczne, ekonometryczne i komputerowe w finansach i ubezpieczeniach*", Uniwersytet Ekonomiczny w Katowicach 2010, 2012.

SŁOWA KLUCZOWE: SEO. Pozycjonowanie stron internetowych. Katalogi artykułów. Wyszukiwarki internetowe.

ABSTRAKT: **Teza/Cel artykułu** – Katalogi artykułów są jednym z narzędzi wykorzystywanych w pozycjonowaniu stron internetowych. Statystyczna analiza ponad 300 katalogów zmierzała do zbadania czynników wpływających na ich jakość mierzoną wskaźnikiem PageRank (PR). **Metody badań** – Analizowano następujące dane: rok rejestracji w bazie *Spis katalogów SEO*, wartość PR, system zarządzania katalogiem, żądanie linku zwrotnego i opłaty za wpis do katalogu oraz moderowanie. Wartości PR zostały zebrane za pomocą skryptu napisanego w języku Python. Analizę statystyczną przeprowadzono wykorzystując programy Microsoft Excel oraz R. Obecność w indeksie wyszukiwarki sprawdzono narzędziem SEOquake. **Wyniki** – Badanie ujawniło, że stale zakładane są nowe katalogi mimo negatywnych opinii specjalistów o ich skuteczności w pozycjonowaniu stron WWW. Ich jakość, mierzona wartością PR, jest jednak bardzo niska. Ponadto, ich niewielką wartość potwierdza także nieobecność dużego odsetka katalogów w indeksie wyszukiwarki Google. **Wnioski** – Tylko niewielki odsetek katalogów osiąga poziom, który może wpływać na ranking promowanych w nich innych witryn WWW. Warunkiem jest staranna moderacja i stosunkowo długi okres funkcjonowania, który sprzyja pozyskiwaniu wartościowych linków.

WSTĘP

Pozycjonowanie (SEO – Search Engine Optimization) obejmuje różnego typu działania, etyczne i nieetyczne, mające na celu poprawę pozycji określonej strony WWW w rankingach internetowych mechanizmów wyszukiwawczych (Henzinger, 2002, p. 3; Jerkovic, 2011, s. 19; Ledford, 2009, s. 17). Coraz doskonalsze algorytmy wyszukiwarek internetowych eliminują wprawdzie niskiej jakości treści z wyników wyszukiwania, ale stale są rozwijane nowe metody nieuczciwego pozycjonowania witryn internetowych. Sprzyja im rozwój serwisów (blogi, fora dyskusyjne, media społecznościowe) opartych na treści generowanej przez ich użytkowników, które pozwalają na zbudowanie stosunkowo tanim kosztem strony internetowej nastawionej wyłącznie na wzrost rankingu promowanej witryny WWW. W tym kontekście badacze zidentyfikowali zjawi-

sko spamu społecznościowego (*social spam*) (Nikolov & Menczer, 2011, pp. 103-108) – publikowania dużej liczby tekstów promujących stosunkowo często wątpliwej jakości towary i usługi różnego typu.

Z uwagi na rolę hiperłączy w obliczaniu ważności strony przez algorytm PageRank (PR), wykorzystywany przez najpopularniejszą na świecie wyszukiwarkę Google, jednym z narzędzi pozycjonowania są katalogi internetowe – uporządkowany tematycznie wykaz adresów URL stron WWW. Katalogi umożliwiają łatwe zbudowanie tzw. zaplecza, czyli dużej liczby stron odsyłających do pozycjonowanego serwisu WWW. Nie pełnią już funkcji źródła informacji o zasobach sieci, a stały się jednym ze sposobów budowania popularności strony internetowej mierzonej liczbą kierujących do niej hiperłączy (Jachimczyk, 2013, s. 229-230). Negatywne zjawiska, takie jak klonowanie, czyli powielanie jednego katalogu pod różnymi nazwami domenowymi (Gyöngyi & Garcia-Molina, 2004, pp. 7-8), wzajemne wiązanie hiperłączami katalogów zarządzanych przez jedną osobę (Davison, 2000), brak starannej moderacji, chaotyczna organizacja zawartości oraz najczęściej brak tematycznego związku z pozycjonowaną witryną spowodowały, że znaczenie katalogów w pozycjonowaniu witryn internetowych nieco spadło, ale nadal stanowią one dość popularne i stosunkowo łatwe w zarządzaniu narzędzie specjalisty SEO.

Wykrycie przez wyszukiwarkę Google nieuczciwych praktyk wykorzystania katalogów (Fishkin, 2007) zwróciło uwagę specjalistów SEO na inną metodę promowania stron WWW – katalogi artykułów (ang. *article directories,* nieco rzadziej używa się określeń *hosted content, hosted marketing content, hosted pages, hosted marketing pages* (Kowalczyk, 2012; Gryszko, 2007; Szymanski, 2007). Polscy specjaliści SEO stosują także nazwę „presell pages" (kolokwialnie nazywane preclami). Niekiedy łączy się je z omówionymi niżej tzw. farmami treści, ale między tego typu witrynami internetowymi występują pewne różnice (*Article directory*, 2014).

W algorytmach robotów indeksujących poza linkami istotne znaczenie ma również otaczająca je treść. Katalogi artykułów umożliwiają publikowanie artykułu wraz z hiperłączem do promowanej strony WWW. Takie powiązanie z punktu widzenia mechanizmu indeksującego wyszukiwarki jest traktowane jako naturalne i wpływa na wyższą ocenę strony, gdyż treść artykułu jest najczęściej tematycznie powiązana z promowaną witryną internetową (Enge et al., 2013, s. 329, 342-343). Z tego względu przez stosunkowo krótki okres katalogi artykułów były powszechnie wykorzystywane przez specjalistów SEO jako skuteczne narzędzie pozycjonowania stron WWW.

Katalogi artykułów, w zdecydowanej większości opierające się na popularnym oprogramowaniu Wordpress, przypominają wielotematyczne blogi (nieco rzadziej występują katalogi poświęcone jednej tematyce) (Olejnik, 2007). Mają zazwyczaj otwarty charakter, ale ich administrato-

rzy niekiedy pobierają opłaty za moderację artykułu lub za możliwość publikowania tekstów w dużej liczbie katalogów (Gatunekchroniony.pl, b.d.; 16ton.pl, b.d.; LGDPaluki, b.d.).

Zgłaszany do publikacji tekst musi spełniać pewne warunki. Powinien być oryginalny, wcześniej nie publikowany oraz powiązany tematycznie ze stroną, do której kieruje umieszczony w artykule link. Teksty są stosunkowo krótkie, ale muszą zawierać określoną przez administratora minimalną liczbę znaków – od 1200 do 2500 (*Katalog Budowlany*, b.d.; *Twoje informacje – artykuły tematyczne*, b.d.; *Artykuły i recenzje*, b.d.). Moderator katalogu określa też zasady dotyczące umieszczania hiperłączy w artykule. Zazwyczaj ich liczba w tekście nie może przekroczyć trzech i powinny one występować np. po pierwszych 300 lub 500 znakach (*Katalog Budowlany*, b.d.; *Twoje informacje – artykuły tematyczne*, b.d.; *Projektowanienazywo.pl*, b.d.; *Artykuły i recenzje*, b.d.). Każdy artykuł powinien być także opisany właściwymi słowami kluczowymi oraz zawierać ilustrację. Moderatorzy kładą również nacisk na poprawność gramatyczną oraz unikanie dodawania artykułów i linków do stron WWW, których treść może naruszać polskie prawo.

Przeszukiwanie zawartości katalogu odbywa się najczęściej na podstawie słowa kluczowego wpisywanego do formularza wbudowanej w system zarządzania treścią wyszukiwarki. Brakuje w niej jednak możliwości dodawania bardziej rozbudowanych kryteriów wyszukiwawczych. Do pewnego stopnia przeglądanie zawartości katalogu ułatwia niekiedy lista kategorii oraz dodawanych przez użytkowników słów kluczowych opisujących każdy artykuł. Wartość tego sposobu charakteryzowania zawartości, tak charakterystycznego dla tzw. folksonomii, czyli indeksowania treści przez użytkowników serwisów, obniża jednak poziom kontroli słownictwa. Efektem jest dowolność w doborze słownictwa, błędy pisowni, albo tworzenie metadanych nie mających związku z indeksowaną treścią (Babik, 2010, s. 183-185).

Katalogi artykułów stosunkowo szybko przestały być jednak efektywnym narzędziem pozycjonowania witryn internetowych. Było to spowodowane wypełnianiem ich bardzo niskiej jakości treścią, wielokrotnie kopiowaną z innych stron internetowych. Podobnie, jak w przypadku katalogów stron internetowych, specjaliści od pozycjonowania zaczęli tworzyć grupy katalogów zarabiając na możliwości dodawania wpisów do jak największej ich liczby (*Solidne precle SEO*, b.d.). W celu uniemożliwienia wykrycia zduplikowanej treści, była ona przeredagowywana (często przez automatyczne narzędzia), co w rezultacie prowadziło do powstania nonsensownego tekstu (Kowalczyk, 2012, Olejnik 2007)[1]. Takie

[1] Przykład takiego tekstu: „ […] Relaks w ostatnim roku bieżące dawna zaledwie chwila. Ogrom historii wytrwałych w planu spowodował, iż udało mi się wygospodarować ale dwa dni na pojęcie oddechu […]". (*Blog*, b.d).

artykuły nie miały najczęściej żadnego związku tematycznego z promowaną witryną i nie wpływały na ranking strony, a nawet mogły go obniżyć. Spamerski charakter katalogów artykułów wywołał reakcję firmy Google, która w kwietniu 2012 r. zaktualizowała algorytm wyszukiwarki w celu ukarania stron otrzymujących linki z tego typu stron (*Article directory*, 2014).

Bardzo trudno oszacować skalę działania presell pages. Baza danych *Spis katalogów* (*Spis Katalogów SEO*, b.d.) rejestruje ich ok. 600, ale dokładniejsza analiza wskazuje, że aktywnych jest mniej więcej połowa. Z kolei właściciel strony Emseo dysponuje płatną bazą ok. 3 tys. katalogów artykułów (*Emseo*, b.d.).

Niekiedy z katalogami artykułów wiąże się tzw. farmy treści (content farm) (*Article directory*, 2014), które również są formą spamowania wyszukiwarek internetowych, ale znacznie trudniejszą do wykrycia przez ich algorytmy. Farmy treści (najbardziej znane to m.in. eHow, Livestrong) zazwyczaj mają zamknięty charakter, gdyż ich właściciele opierają się na pewnej grupie wyselekcjonowanych płatnych współpracowników, którzy są w stanie, w stosunkowo krótkim czasie, napisać dużą liczbę niezbyt długich (do 500 wyrazów) tekstów. Tematyka tych poprawnie napisanych, niekiedy recenzowanych artykułów zazwyczaj dotyczy zagadnień, których najczęściej poszukują użytkownicy wyszukiwarek. W przeciwieństwie do katalogów artykułów teksty nie są anonimowe, niekiedy towarzyszy im również lista wykorzystanych źródeł (*What is content farm?* 2011; Spavlik, 2011, p. 18; Lalik, 2011; Kennedy, 2010, p. 19; Notess, 2011, p. 47).

Celem farm treści jest na ogół własna promocja w wyszukiwarkach, dlatego w artykułach brak hiperłączy do innych witryn. Farmy zarabiają na reklamach wyświetlanych przy publikowanych artykułach. W związku z tym ich tematyka koncentruje się na tematach, które będą aktualne przez stosunkowo długi okres i dzięki temu dłużej będą generowały przychód z wyświetlanych reklam (Notess, 2011, p. 47; Grensing-Pophal, 2010, p. 15).

Nacisk na opracowywanie jak największej liczby tekstów w jak najkrótszym czasie sprzyja powstawaniu tekstów lakonicznych, powierzchownych i odtwórczych (Nikolov & Menczer, 2011, p. 105; Notess, 2011, p. 46), które jednak lokują się stosunkowo wysoko w rankingach internetowych wyszukiwarek. Firma Google stara się wprawdzie aktualizować algorytm wyszukiwarki w celu obniżenia rankingu tego typu witryn WWW, ale te działania są stosunkowo nieskuteczne, gdyż farmy treści nie naruszają zasad stawianych webmasterom przez producenta najpopularniejszej wyszukiwarki. Obniżeniem rankingu zostały ukarane tylko te farmy treści, które kradną i duplikują treść na swoich witrynach (Spavlik, 2011, p. 18).

STAN BADAŃ

W literaturze naukowej brak prac poświęconych katalogom artykułów. Badacze koncentrują się na technikach spamerskich i metodach ich wykrywania. Niektóre z opisanych przez nich technik odnoszą się także do katalogów artykułów. Brian Davison opisał tzw. nepotyczne linki (nepotistic links), które łączą strony internetowe zarządzane przez jedną osobę lub organizację (Davison, 2000). Do cech stron spamerskich zalicza się także rejestrowanie wielu nazw domenowych odnoszących się do jednego adresu IP czy nadmierne duplikowanie treści (Fetterly & Manasse & Najork, 2004). Popularną spamerską praktyką jest także klonowanie katalogów internetowych (Gyöngyi & Garcia-Molina, 2004, pp. 7-8)

Analizy poświęcone katalogom internetowym znajdziemy również w pracach specjalistów SEO. Ich wnioski zachowują także aktualność w stosunku do katalogów artykułów. Kurtis Bohrnstedt z firmy SEOmoz po zbadaniu ponad 2500 katalogów doszedł do wniosku, że 20% z nich to katalogi spamerskie (Bohrnstedt, 2012). Rand Fishkin z SEOmoz wskazał, że pewne cechy katalogów, takie jak ogólna tematyka, brak moderacji i akceptacja wszystkich zgłoszeń, promocja katalogu jako strony z wysokim PR, powiązanie hiperłączami z innymi katalogami, czy żądanie linku zwrotnego, mogą wskazywać na ich potencjalnie manipulacyjny charakter (Fishkin, 2007).

ZAKRES PRACY

Analizie statystycznej poddano 317 katalogów artykułów wybranych z bazy dostępnej na stronie *Spis Katalogów SEO* (*Spis Katalogów SEO*, b.d.).

Główne cele analizy:

1. Ocena popularności katalogów artykułów jako narzędzia SEO. Popularność tę mierzyliśmy dynamiką ich powstawania w kolejnych latach ustaloną w oparciu o datę wpisania danego katalogu do *Spisu katalogów SEO* (*Spis Katalogów SEO*, b.d.). W tym aspekcie badanie pozwala zweryfikować twierdzenia specjalistów SEO o malejącym znaczeniu katalogów jako narzędzia pozycjonowania witryn internetowych. W tej części nasza analiza zmierzała także do zbadania zależności między czasem działania katalogu a jego wskaźnikiem PR.

2. Oszacowanie odsetka katalogów spamerskich, czyli takich, które stosują nieuczciwe techniki pozycjonowania.

Dokonaliśmy tego poprzez:

a) porównanie adresu IP z nazwą domenową. Dzięki temu wykryliśmy stopień występowania zjawiska klonowania katalogów, które jest popularną metodą wykorzystywaną w nieuczciwym pozycjonowaniu.

b) analizę indeksowania katalogów przez wyszukiwarkę Google i skorelowanie indeksowania z wartością PR. Brak strony w indeksie wyszukiwarki również najczęściej oznacza nieuczciwy charakter strony.

3. Ustalenie liczby moderowanych i niemoderowanych katalogów, płatnych i bezpłatnych oraz wymagających w zamian za publikację artykułu umieszczenia tzw. linku zwrotnego do katalogu, rodzaju wykorzystywanego oprogramowania zarządzającego katalogiem. W tej części zbadaliśmy także, jak cechy katalogów, takie jak moderacja oraz odpłatność, wpływają na ich jakość mierzoną wskaźnikiem PR. Nie badaliśmy wpływu żądania umieszczenia linku zwrotnego na promowanej stronie z uwagi na małą liczbę (tylko 6) katalogów stawiających taki warunek. Staranne zarządzanie katalogiem pozwala uniknąć wpisów o spamerskim charakterze (np. skopiowanych z innych witryn WWW), które obniżają jego jakość i pośrednio wpływają także na ranking witryny, do której kieruje hiperłącze. Z moderacją może być także powiązana kwestia odpłatności za publikację artykułu, ale w tym przypadku nie ma prostej reguły, która mówi, że płatne katalogi charakteryzują się lepszą jakością. Mogą one bowiem uzależniać publikację tekstu od wniesienia opłaty nie zwracając uwagi na wartość zgłaszanego tekstu.

OGRANICZENIE BADANIA

Wyznacznikiem jakości strony uczyniliśmy wartość PR. Algorytm wyszukiwarki Google oblicza ją na podstawie liczby hiperłączy kierujących do strony z innych witryn WWW. Uwzględnia także jakość witryn, na których znajduje się hiperłącze (Liu, 2007, pp. 245-246). Duża liczba linków nie musi się więc przekładać na odpowiednio wysoką wartość PR analizowanej witryny. W wielu przypadkach algorytm nie będzie jej przypisywał żadnej wartości z uwagi na niską jakość stron WWW zawierających hiperłącze kierujące do pozycjonowanej witryny. W ten sposób można np. manipulować rankingiem konkurencyjnych stron internetowych. Dokładny sposób wyliczania wartości PR nie jest jednak znany, gdyż stanowi tajemnicę handlową firmy Google. Upublicznia ona wprawdzie informację o wskaźniku PR dla indeksowanych przez nią stron WWW (wyświetlają ją niektóre narzędzia SEO, m.in. *Mozbar* lub *SEOquake*), ale, jak zaznaczają specjaliści, wartość ta jest uaktualniana stosunkowo rzadko i nie odzwierciedla rzeczywistego wskaźnika PR obliczanego przez firmę Google (Joyce, 2013; Joyce, 2012; Chant, 2011). Ostatnia publicznie znana aktualizacja PR miała miejsce między 5 a 6 grudnia 2013 r. i ją wzięliśmy pod uwagę w analizie (Anderson, 2014).

Brak PR może świadczyć o spamerskim charakterze katalogu. Trzeba także uwzględnić fakt, że nowo powstałe strony również mogą jeszcze nie

posiadać ustalonego wskaźnika (Dover & Dafforn, 2011 s. 94-95). Nie do końca jest jasne, jak traktować strony z PR=0. Może to oznaczać karę nałożoną na stronę przez wyszukiwarkę, ale według niektórych specjalistów SEO zerowa wartość oznacza, że do strony nie kierują żadne wartościowe linki (*PR0 – Google's PageRank 0 Penalty,* b.d.). Dlatego dla potrzeb naszej analizy, badając wpływ wieku katalogu na wskaźnik PR i korelację między PR a obecnością w indeksie Google, przyjęliśmy podział katalogów na trzy grupy: katalogi z PR>0, z PR=0 oraz katalogi bez ustalonej wartości PR.

Inne ograniczenie badania wynika też z faktu, że PR jest wyliczany osobno dla każdej strony tworzącej serwis WWW. Możliwa jest więc sytuacja, że strona główna, dla której sprawdzaliśmy PR, ma niższą wartość od wielu innych stron tworzących katalog (Bailyn & Bailyn, 2012, s. 18-19).

Brak strony w indeksie wyszukiwarki Google może być spowodowany nie tylko spamerskim charakterem, ale także jej pewnymi parametrami, które uniemożliwiają wyszukiwarce indeksowanie. Na przykład, znacznik meta w kodzie HTML <META NAME="ROBOTS" CONTENT="NOINDEX"> zabrania robotom wyszukiwarek internetowych indeksowania określonej strony (Jones, 2012; *Metatagi*, b.d.). W przypadku katalogu artykułów raczej mało prawdopodobne jest jednak zakazywanie wyszukiwarce indeksowania zawartości. Natomiast, jak zwraca uwagę specjalista SEO Bob Jones, możliwa jest sytuacja, że nazwa domenowa katalogu była już wcześniej wykorzystywana przez inną spamerską witrynę i wyszukiwarka nadal odmawia jej indeksowania (Jones, 2012).

Ponadto trzeba mieć na uwadze, że data rejestracji w wykazie nie musi być tożsama z datą powstania katalogu, ale tylko na tej podstawie mogliśmy w przybliżeniu ustalić jego datę rozpoczęcia działalności.

MATERIAŁ I METODY BADAWCZE

Materiału do analizy dostarczyła nam moderowana baza danych polskich katalogów internetowych *Spis Katalogów SEO* (*Spis Katalogów SEO,* b.d.). Plik, który stał się podstawą analizy (pobrany w dniu 12 września 2014 r.) obejmował 6581 katalogów, w tym 626 (niecałe 10%) zakwalifikowanych jako presell pages.

Baza danych, na podstawie której analizowaliśmy katalogi artykułów, rejestruje je od 2005 r. Wykaz jest nieregularnie aktualizowany. Kryteria rejestracji w bazie obejmują m.in. następujące zasady (*Dodaj Katalog !!!*, b.d.):

– Akceptowane są tylko katalogi, które służą celom pozycjonowania stron internetowych.

– Nie są rejestrowane katalogi, które umożliwiają ustawienie atrybutu nofollow – uniemożliwiającego robotom wyszukiwarek podążanie za linkami umieszczonymi w artykule.

Przed analizą statystyczną zweryfikowaliśmy dane o katalogach artykułów. W badaniu uwzględniliśmy tylko działające katalogi. Wykluczyliśmy niedostępne strony internetowe, te, które nie były katalogami artykułów, lub takie, które przekierowywały na zupełnie inną witrynę. Ostatecznie analizie poddano 317 katalogów artykułów.

Analiza dotyczyła następujących informacji: rok wpisania do wykazu, wartość PR, rodzaj oprogramowania do zarządzania katalogiem, żądanie linku zwrotnego i opłaty za wpis do katalogu oraz jego moderowanie. W dniu 26 września 2014 r. zweryfikowano deklarowane w bazie danych wartości PR dla strony głównej tych katalogów. Weryfikacja okazała się konieczna, gdyż w wykazie stronom bez ustalonej wartości PR przypisano wskaźnik 0. Ponadto zbadano ile nazw domenowych katalogów jest przypisanych do jednego adresu IP.

Wartości PR zostały zebrane automatycznie za pomocą skryptu napisanego w języku Python. Analizę statystyczną przeprowadzono wykorzystując programy Microsoft Excel oraz R (*R Core Team*, 2013). Do oceny istotności różnic między średnimi stosowano test t-Studenta, zaś istotność różnic dla danych kategorialnych oceniano testem χ^2. Różnice uznawano za znamienne statystycznie, gdy p-value < 0,05. Obecność w indeksie wyszukiwarki została sprawdzona w dniu 21 października 2014 r. narzędziem SEOquake (wersja 2.8.15) – wtyczki do przeglądarki Mozilla Firefox (*SEOquake,* b.d.).

Biorąc pod uwagę fakt, że aktualizacja wskaźnika PR miała miejsce w grudniu 2013 r. w analizie uwzględniliśmy tylko katalogi zarejestrowane do grudnia 2013 r., a pominęliśmy strony rejestrowane w 2014 r., gdyż jako nowo powstałe nie objęła ich jeszcze aktualizacja PR. Do obliczenia okresu działania katalogu przyjęliśmy datę jego rejestracji w bazie oraz datę aktualizacji wartości PR, która miała miejsce między 5 a 6 grudnia 2013 r.

WYNIKI ANALIZY

Ogólną charakterystykę katalogów artykułów przedstawia tabela 1.

Analiza dat rejestracji katalogów artykułów wskazuje, że stale są zakładane nowe katalogi, mimo negatywnych opinii specjalistów o ich skuteczności w pozycjonowaniu stron internetowych. W porównaniu z katalogami stron internetowych stanowią one jednak znaczącą mniejszość, gdyż zarządzanie katalogiem artykułów wymaga znacznie staranniejszej administracji w celu uniknięcia dodawania niskiej jakości artykułów.

Stosunkowo duża różnica między liczbą katalogów zarejestrowanych w latach 2006-2011 a 2012-2014 wynika z faktu, że analizowaliśmy tylko katalogi działające, nie braliśmy pod uwagę tych, które nie są już do-

Tabela 1

Charakterystyka katalogów artykułów (n=317)

Dane	Liczba	Procent
Data rejestracji		
2006	1	0,3
2007	4	1,3
2008	12	3,8
2009	16	5,0
2010	25	7,9
2011	24	7,6
2012	65	20,5
2013	109	34,4
2014	61	19,2
PageRank		
Brak	172	54,3
0	87	27,4
1	31	9,8
2	21	6,6
3	4	1,3
4	1	0,3
5	1	0,3
Obecność w indeksie Google		
Jest	46	14,5
Brak	271	85,5
Moderowany		
Nie	58	18,3
Tak	259	81,7
Płatny		
Nie	264	83,3
Tak	53	16,7
Link zwrotny		
Nie	314	99,05
Tak	3	0,95
Wykorzystywany skrypt		
Wordpress	304	95,9
Autorski	5	1,6
MocneLinki	4	1,3
Inny	4	1,3

Źródło: opracowanie własne

stępne w sieci. Porównując te dwa okresy możemy postawić hipotezę, że przynajmniej połowa z katalogów zarejestrowanych w latach 2013-2014 przestanie być wkrótce dostępna, gdyż ich właściciele zrezygnują z możliwości płatnego przedłużenia ważności domeny internetowej, pod którą działa katalog.

Różnica między liczbą katalogów zarejestrowanych w 2013 i 2014 r. jest spowodowana tym, że baza obejmuje dane tylko do września 2014 r. Po-

równując dane z 2013 r. można więc zakładać, że w 2014 r. zostanie założonych również ponad 100 nowych katalogów.

Brak wskaźnika PR dla części katalogów, zwłaszcza tych istniejących w sieci ponad rok, wskazuje na ich potencjalnie spamerski charakter, co skutkuje karą w postaci nieprzyznania wartości PR. Brak PR charakteryzuje również strony WWW działające stosunkowo krótko, najwyżej kilka miesięcy, gdyż nie zdążyły jeszcze zebrać puli wartościowych linków z innych witryn internetowych.

Ogółem prawie 55% katalogów nie posiada wskaźnika PR, a blisko 28% legitymuje się zerowym wskaźnikiem. Wśród katalogów zarejestrowanych w bazie do 2013 r. wskaźnika PR nie posiada 46,5%, a prawie 33% ma zerową wartość. Związek między PR i wiekiem katalogu okazał się wysoce statystycznie istotny (p-value=0,0002). Mediana wieku katalogów wynosiła 15,4 miesiąca i dlatego przyjęto 15 miesięcy jako graniczną wartość przy dychotomizacji wieku w poniższych analizach. Częstość występowania katalogów bez ustalonej wartości PR jest ponad 1,5 razy większa wśród tych, które działają najwyżej 15 miesięcy, niż wśród tych, które działają dłużej (zob. Tab. 2). Tu, do pewnego stopnia, na wyniki wpływa również fakt, że w grupie katalogów działających nie dłużej niż 15 miesięcy, będą się znajdować także strony istniejące zaledwie kilka miesięcy, którym wyszukiwarka nie przyznała jeszcze rankingu.

Z kolei katalogi z PR równym co najmniej 1 występują dwa razy częściej wśród katalogów działających ponad 15 miesięcy, niż wśród katalogów działających krócej. Zwraca uwagę stosunkowo duży, sięgający ponad 34%, odsetek katalogów bez ustalonego wskaźnika PR istniejących ponad 15 miesięcy. Katalogi bez PR działają istotnie ($p<0,002$) krócej, niż te, dla których PR jest ustalony (średnio 17,7 miesiąca vs. 25,6 miesiąca).

Trzeba podkreślić stosunkowo dużą liczbę katalogów z zerową wartością PR. Większy odsetek takich stron występuje zwłaszcza w grupie katalogów działających ponad 15 miesięcy. Bardzo trudno jednak, z powodu braku informacji o poprzednich wartościach PR, odpowiedzieć na pytanie, jaki odsetek analizowanych stron zwiększył swój wskaźnik PR, a u ilu te wartości się zmniejszyły.

Tabela 2

Rozkład PageRank w zależności od wieku katalogu – dane dla katalogów zarejestrowanych najpóźniej w 2013 r. (n=256)

PageRank	Do 15 miesięcy	Ponad 15 miesięcy	Do 15 miesięcy (%)	Ponad 15 miesięcy (%)
Bez wartości	74	45	59,2	34,4
0	34	50	27,2	38,2
1-5	17	36	13,6	27,5
Razem	125	131	100	100

Źródło: opracowanie własne.

O spamerskim charakterze katalogów artykułów świadczy korelacja między nazwą domenową a adresem IP. Ogółem stwierdzono 246 różnych adresów IP, z czego 34 były wykorzystywane przez więcej niż jeden katalog. Większości adresów IP wykorzystywanych przez więcej niż jeden katalog, odpowiadały dwa katalogi (24 przypadki adresów IP). Spośród adresów wykorzystywanych przez więcej niż dwa katalogi, najczęściej wykorzystywany był adres 5.9.82.27 (13 katalogów), a następnie adres 91.228.196.15 (9 katalogów) i adres 144.76.99.221 (8 katalogów). Te strony najczęściej charakteryzuje identyczny wygląd oraz tematyka artykułów.

Inną metodą sprawdzenia jakości katalogu jest analiza jego występowania w indeksie Google. Aż 85% katalogów zarejestrowanych w latach 2006-2014 nie jest indeksowanych przez wyszukiwarkę. Jeśli z analizy wykluczymy katalogi zarejestrowane w 2014 r. odsetek nieindeksowanych stron jest równie wysoki i sięga blisko 84%.

Tabela 3

Rozkład PageRank w zależności od występowania w indeksie Google – dane dla katalogów zarejestrowanych najpóźniej w 2013 r. (n=256)

PageRank	Katalogi nieindeksowane przez Google	Katalogi indeksowane przez Google	Katalogi nieindeksowane przez Google (%)	Katalogi indeksowane przez Google (%)
Bez wartości	113	6	52,8	14,3
0	81	3	37,9	7,1
1-5	20	33	9,3	78,6
Razem	214	42	100	100

Źródło: opracowanie własne.

Katalogi indeksowane przez Google działały statystycznie istotnie (p-value=0,002) dłużej niż nieindeksowane (średnio 32,9 miesięcy vs. 19,8 miesięcy).

Rozkład PR statystycznie istotnie (p-value<0,00001) różnił się wśród katalogów indeksowanych i nieindeksowanych przez Google (zob. Tab 3). Wartość PR równa co najmniej 1 około 8-krotnie częściej była obserwowana dla katalogów indeksowanych przez Google niż dla nieindeksowanych, zaś brak wartości PR notowano trzykrotnie rzadziej dla katalogów indeksowanych niż nieindeksowanych.

10% katalogów nieindeksowanych przez Google ma PR>0, co świadczy o tym, że taki odsetek stron został zakwalifikowany przez wyszukiwarkę jako spam i usunięty z jej indeksu. Z kolei wśród katalogów indeksowanych przez Google ponad 20% ma PR=0 lub nie ma określonego PR, co oznacza pewną poprawę ich jakości.

Tylko niecałe 20% katalogów artykułów zadeklarowało się w bazie danych jako niemoderowane. Niemniej jednak duży odsetek katalogów, które nie posiadają ustalonego wskaźnika PR, może wskazywać, że moderacja w tych katalogach jest bardzo pobieżna i niestaranna. Katalogi moderowane i niemoderowane zarejestrowane w latach 2006-2013 w sposób statystycznie istotny różniły się rozkładem PR (p-value=0,002): katalogi z wyższym PR częściej występowały wśród katalogów moderowanych niż niemoderowanych – katalogi z PR>0 stanowiły 23,4% katalogów moderowanych i tylko 9,8% niemoderowanych.

Stosunkowo niewielki odsetek, ok. 17% katalogów, deklarowało się jako płatne. Widać tutaj zasadniczą różnicę w stosunku do katalogów stron internetowych, gdzie ten odsetek przekracza 40% (Jachimczyk & Chrapek & Chrapek, b.d.). Przypuszczamy, że właściciele katalogów starają się zarabiać nie na moderowaniu zgłaszanych artykułów, a raczej na usłudze dodawania wpisów do jak największej liczby zarządzanych przez siebie katalogów, jak w przypadku strony *Solidne Precle SEO* (*Solidne Precle SEO*, b.d.). Nie obserwowano statystycznie istotnej różnicy między rozkładami PR w zależności od tego, czy katalog jest płatny, czy bezpłatny (p-value=0,69).

Znikomy jest udział tych katalogów artykułów, które w zamian za wpis żądają umieszczenia zwrotnego hiperłącza. W analizowanej grupie tylko 3 katalogi stawiały taki warunek. Takie żądanie jest jedną z cech strony spamerskiej i prawdopodobnie obawa przed karą ze strony wyszukiwarki skłania administratorów katalogów do rezygnacji z żądania umieszczenia takiego odnośnika.

Katalogi w zdecydowanej większości opierają się na popularnym systemie zarządzania treścią Wordpress. Jest on niezwykle łatwy w instalacji i prosty w obsłudze, gdyż został zaprojektowany pod kątem zarządzania internetowymi blogami. Wykorzystuje go blisko 96% katalogów artykułów.

PODSUMOWANIE

Omawiane w artykule strony internetowe lokują się bardzo nisko na liście z rezultatami wyszukiwania w wyszukiwarce Google, ale ich zadaniem nie jest przekazywanie informacji, tylko zakulisowe promowanie innych witryn internetowych. Mogą być także wykorzystane jako metoda szkodliwego pozycjonowania, nieuczciwej konkurencji polegającej na umieszczaniu na szkodliwej witrynie hiperłącza do strony internetowej konkurenta.

Analiza pokazuje, że reprezentują stosunkowo nieliczną grupę witryn internetowych niezwykle niskiej jakości. Tylko niecałe 19% stron głównych katalogów osiąga wartość PR większą od zera. Niemniej jednak stale powstają nowe katalogi artykułów z myślą o promowaniu kolejnych serwi-

sów internetowych. Około połowa katalogów przestaje działać po upływie roku, co jednoznacznie wskazuje, że jedynym celem ich powstania było spamowanie wyszukiwarek internetowych.

Tylko niewielki odsetek katalogów osiąga poziom, który może wpływać na ranking promowanych w nich innych witryn internetowych. Warunkiem jest, co udowadnia analiza, staranna moderacja i odpowiednio długi okres funkcjonowania, który sprzyja pozyskiwaniu wartościowych linków.

BIBLIOGRAFIA

16ton.pl (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.16ton.pl/jak--dodac-artykul/>.

Anderson, Shaun (2014). *Google Toolbar Pagerank Update History* [online], [dostęp: 04.11.2014]. Dostępny w WWW: <http://www.hobo-web.co.uk/google-pr-update/>.

Article directory (2014) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://en.wikipedia.org/wiki/Article_directory>.

Artykuły i recenzje (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://recenzje.info.pl/jak-dodac-artykul/>.

Babik, Wiesław (2010). *Słowa kluczowe.* Kraków: Wydaw. Uniwersytetu Jagiellońskiego.

Bailyn, Evan; Bailyn, Bradley (2012). *Przechytrzyć Google. Odkryj skuteczną strategię SEO i zdobądź szczyty wyszukiwarek.* Gliwice: Helion.

Blog (b.d) [online], [dostęp: 20.11.2014]. Dostępny w WWW: <http://ala12.orgfree.com/?p=2771>.

Bohrnstedt, Kurtis (2012). *Web Directory Submission Danger: Analysis of 2,678 Directories Shows 20% Penalized/Banned by Google* [online], [dostęp: 14.01.2015]. Dostępny w WWW: <http://moz.com/blog/web-directory-submission-danger>.

Chant, Rob (2011). *Introduction to Google PageRank: Myths & Facts* [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://searchenginewatch.com/article/2064605/Introduction-to-Google-PageRank-Myths-Facts>.

Davison, Brian D. (2000). *Recognizing Nepotistic Links on the Web*. [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.aaai.org/Papers/Workshops/2000/WS-00-01/WS00-01-005.pdf>.

Dodaj Katalog !!! (b.d.) [online], [dostęp: 20.10.2014] Dostępny w WWW: <http://www.katalogiseo.info/add.php>.

Dover, Danny; Dafforn, Erik (2011). *Sekrety SEO*. Gliwice: Helion.

Emseo (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://sklep-emseo.pl/prywatne-precle/8-prywatne-precle-1000-szt-.html>.

Enge, Eric; Spencer, Stephan; Stricchiola, Jessie; Fishkin, Rand (2013). *Sztuka SEO; optymalizacja witryn internetowych*. Gliwice: Helion.

Fetterly, Dennis; Manasse, Mark; Najork, Marc (2004). *Spam, Damn Spam, and Statistics: Using statistical analysis to locate spam web pages*. [online], [dostęp: 14.01.2015]. Dostępny w WWW: <http:// re-search.microsoft.com/pubs/59848/webdb2004.pdf>.

Fishkin, Rand (2007). *What Makes a Good Web Directory, and Why Google Penalized Dozens of Bad Ones*. [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://moz.com/blog/what-makes-a-good-web-directory-and-why-google-penalized-dozens-of-bad-ones>.

Gatunekchroniony.pl (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.gatunekchroniony.pl/dodaj-wpis/>.

Grensing-Pophal, Lin (2010). Content's latest creation model. Factory farmed or organically grown. *Econtent*, September, pp. 14-18.

Gryszko, Marta (2007). *Czym są Presell Pages?* [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.lexy.com.pl/blog/presell-pages.html>.

Gyöngyi, Zoltan; Garcia-Molina, Hector (2004). *Web Spam Taxonomy*. *Technical Report*. [online], [dostęp: 20.10.2014], pp. 1-11. Dostępny w WWW: <http://ilpubs.stanford.edu:8090/646/1/2004-25.pdf>.

Henzinger, Monika R.; Motwani, Rajeev; Silverstein, Craig (2002), *Challenges in web search engines. SIGIR Forum Fall* [online], vol. 36, no. 2, pp. 1-12 [dostęp: 20.10.2014]. Dostępny w WWW: <http://sigir.org/files/forum/F2002/henzinger.pdf>.

Jachimczyk, Adam (2013). Katalog internetowy. Nieoczekiwana zmiana miejsc. W: *Bibliografia: źródła, standardy, opracowania*. Praca zbior. pod red. Jerzego Franke. Warszawa: Wydaw. SBP, s. 217-230.

Jachimczyk, Adam; Chrapek, Magdalena; Chrapek, Zbigniew (b.d.). *Web directories. Selected features and their impact on the quality of directories.* (Materiał niepublikowany).

Jerkovic, John I., *Wojownik SEO: sztuka osiągania najwyższych pozycji w wynikach wyszukiwania* (2011). Gliwice: Helion.

Jones, Bob (2012). *8 Reasons Why Your Site Might Not Get Indexed.* [online], [dostęp: 13.11.2014]. Dostępny w WWW: <http://moz.com/ugc/8-reasons-why-your-site-might-not-get-indexed>.

Joyce, Julie (2012). *Why Links Matter*. [online], [dostęp: 20. 10. 2014]. Dostępny w WWW: <http://searchenginewatch.com/article/2166568/Why-Links-Matter>.

Joyce, Julie (2013). *Link Building A-Z Guide – Definitions & Terms* [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://searchenginewatch.com/article/2172916/Link-Building-A-Z-Guide-Definitions-Terms>.

Katalog Budowlany (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://budownictwo.hl1.pl/jak-dodc-wpis/>.

Kennedy, Shirley Duglin (2010). 'Content farms', information literacy and You. *Information Today*, November, pp. 17-19.

Kowalczyk, Tomasz, *Co to jest Presell Page i Precel?* (2012) [online], [dostęp: 20.10.2014] Dostępny w WWW: <http://seomoher.pl/podstawy-seo/co-to-jest-presell-page-i-precel.html>.

Lalik, Ewa (2011). *Coraz gorsza jakość wyszukiwania w Google? Przez content farms* [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.spidersweb.pl/2011/02/coraz-gorsza-jakosc-wyszukiwania-w-google-przez-content-farms.html>

Ledford, Jerri L., *SEO: biblia: wiedza obiecana* (2009), Gliwice: Helion.

LGDPaluki (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.lgdpaluki.pl/jak-dodac-artykul>.

Liu, Bing, *Web data mining: exploring hyperlinks, contents, and usage data* (2007). Berlin: Springer Verlag.

Metatagi, które Google potrafi zinterpretować (b.d.) [online], [dostęp: 13.11.2014]. Dostępny w WWW: <https://support.google.com/webmasters/answer/79812?hl=pl>.

Nikolov, Dimitar; Menczer, Filippo (2011). Social spam. In: *The death of the Internet*. Ed. by M. Jakobsson. Hoboken, NJ: John Wiley & Sons, pp. 103-117.

Notess, Greg R. (2011). Content farming, quick creation, and declining information quality. *Online*, May-June, pp. 46-48.

Olejnik, Łukasz (2007). *Strony typu „presell"* [online], [dostęp: 20. 10. 2014]. Dostępny w WWW: <http://googlepolska.blogspot.com/2007/08/strony-typu-presell.html>.

PR0 – Google's PageRank 0 Penalty (b.d.) [online], [dostęp: 28.10.2014]. Dostępny w WWW: <http://pr.efactory.de/e-pr0.shtml>

Projektowanienazywo.pl (b.d.) [online], [dostęp: 20. 10. 2014]. Dostępny w WWW: <http://www.projektowanienazywo.pl/jak-dodac-wpis>.

R wersja 3.0.2 R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/).

SEOquake (b.d.) [online], [dostęp: 04.11.2014]. Dostępny na WWW: <http://www.seoquake.com/>.

Solidne Precle SEO (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://precle.eu/>.

Spavlik, Janet (2011). Dispatches from the content farm trenches. *Econtent*, November, pp. 16-20.

Spis Katalogów SEO (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://www.katalogiseo.info/>

Szymanski, Kaspar (2007). *Site content and use of web catalogues* [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://googlewebmastercentral.blogspot.com/2007/03/site-content-and-use-of-web-catalogues.html>.

Twoje informacje – artykuły tematyczne (b.d.) [online], [dostęp: 20.10.2014]. Dostępny w WWW: <http://twoje.info.pl/dodaj-artykul/>.

What is content farm? (2011) [online], [dostęp: 20. 10. 2014]. Dostępny w WWW:
<http://www.econtentmag.com/Articles/Resources/Defining-EContent/What-is-a-Content-Farm-78370.htm>.

Artykuł w wersji poprawionej wpłynął do Redakcji 14 stycznia 2015 r.

ADAM JACHIMCZYK
Institute of Journalism and Information
Jan Kochanowski University in Kielce
e-mail: Adam.Jachimczyk@ujk.edu.pl

MAGDALENA CHRAPEK
Institute of Mathematics
Jan Kochanowski University in Kielce
e-mail: Magdalena.Chrapek@ujk.edu.pl

DIRECTORIES OF ARTICLES (PRESELL PAGES) – THE ANALYSIS OF WEBPAGES USED IN SEO

KEYWORDS: SEO (Search Engine Optimization). Webpage positioning. Directories of articles. Web search engines.

ABSTRACT: **Thesis/Objective** – Directories of articles are one of the tools used in SEO. The statistical examination of more than 300 directories sought to identify factors influencing their quality measured with PageRank (PR) indicator. **Research methods** – The following data were analyzed: year of registration in *SEO directories list*, PR value of the directory, directory management system, the request for the backlink, the fee for the article added to the directory, the moderation of the directory. The analysis was performed with Microsoft Excel and R software, the PR values were collected with software written in Python and the presence in Google search engine index was checked with SEOquake tool. **Results** – The research revealed that new directories were built despite the negative opinions of the professionals on their usefulness in webpage positioning. Their value measured with PR indicator is extremely low, confirmed with absence of considerable percentage of the directories in Google index. **Conclusions** – Very few directories achieve the level when they begin to have impact on the positioning of websites they list provided that they offer careful moderation and operate for a long period of time which helps to gather valuable links.