Politechnika Śląska

Wydział Automatyki, Elektroniki i Informatyki

Mariusz Duka

Rozprawa doktorska

Wyznaczanie rankingu stron WWW algorytmem ISOWQ Rank

Promotor:

dr hab. Artur Strzelecki
Uniwersytet Ekonomiczny w Katowicach

Gliwice 2023

Pragnę podziękować:

Promotorowi,
Panu dr. hab. Arturowi Strzeleckiemu,
prof. Uniwersytetu Ekonomicznego w Katowicach
za opiekę merytoryczną, wyrozumiałość, za cenne uwagi i sugestie,
za zaangażowanie, dzięki któremu możliwe było napisanie tej pracy.

Panu dr. hab. Markowi Sikorze,
prof. Politechniki Śląskiej w Gliwicach
za współpracę i całą pomoc, jakiej udzielił mi w czasie,
gdy rodziła się tematyka mojej pracy doktorskiej.

Spis treści

Wykaz oznaczeń używanych w pracy

4G	–	czwarta generacja sieci telefonii komórkowej,
5G	–	piąta generacja sieci telefonii komórkowej,
Alexa Rank	–	system rankingowy oparty na ruchu wygenerowanym przez użytkowników na określonej stronie internetowej,
ARPANET	–	(ang. Advanced Research Projects Agency Network) pierwsza sieć rozległa oparta na rozproszonej architekturze i protokole TCP/IP,
ccTLD	–	(ang. country code top-level domain) dwuliterowa, krajowa domena najwyższego poziomu, zarezerwowana dla państwa lub terytorium zależnego,
CF	–	(ang. Citation Flow) wskaźnik jakości profilu hiperłączy dostępny w narzędziu Majestic,
CMS	–	(ang. Content Management System) wyposażony w panel administracyjny system do zarządzania treścią na stronie WWW,
CPC	–	(ang. Cost Per Click) współczynnik efektywności reklamy wyrażony jako stosunek kosztów emisji danej reklamy do liczby kliknięć w tę reklamę, występuje w modelu rozliczenia PPC,
CSS	–	(ang. Cascading Style Sheets) język służący do opisu formy prezentacji stron WWW,
CPM	–	(ang. Cost Per Mille) forma rozliczania reklamy internetowej polegająca na tym, że reklamodawca płaci za wyświetlenie reklamy przez tysiąc użytkowników,
CTR	–	(ang. Click Through Rate) współczynnik klikalności, stosunek między liczbą kliknięć a wyświetleniami reklamy mierzony w procentach,
DARPA	–	(ang. Defense Advanced Research Projects Agency) amerykańska agencja rządowa zajmująca się rozwojem techniki wojskowej, twórca protokołu TCP/IP,
DMOZ	–	wielojęzyczny katalog stron WWW, działający w latach 1998–2017,
DNSbl	–	usługa oparta na systemie DNS wykorzystywana do publikowania list adresów IP nadawców spamu,

EDGE – (ang. Enhanced Data Rates for GSM Evolution) technologia związana z pakietowym przesyłaniem danych w sieciach GSM, rozszerzenie technologii GPRS,

FLV – format multimedialny używany do dystrybucji plików wideo przez sieć internet,

GPRS – (ang. General Packet Radio Service) technologia związana z pakietowym przesyłaniem danych w sieciach GSM,

HITS – (ang. Hyperlink Induced Topic Search) algorytm rankingowy dla wyszukiwarek opracowany przez Jona Kleinberga w 1998 roku,

HSDPA – (ang. High Speed Downlink Packet Access) technologia związana z pakietowym przesyłaniem danych w sieciach GSM, nazywana też siecią trzeciej generacji (3G lub 3G+),

HTML – (ang. HyperText Markup Language) język pozwalający opisać strukturę informacji zawartych wewnątrz strony WWW,

HTTP – (ang. HyperText Transfer Protocol) protokół warstwy aplikacji odpowiedzialny za przesyłanie dokumentów hipertekstowych,

IoT – (ang. Internet of Things) internet rzeczy,

ISOWQ – (ang. International Studies of Website Quality) nazwa systemu rankingowego umożliwiającego wykonanie audytu strony WWW,

ISOWQ Rank – algorytm rankingowy wykorzystywany przez system ISOWQ,

LTE – (ang. Long Term Evolution) standard bezprzewodowej transmisji danych będący następcą systemów trzeciej generacji,

MOZ – nazwa firmy i opracowany przez nią algorytm rankingowy stosowany przez narzędzie MOZ Analytics (dostępne pod adresem https://moz.com), umożliwiające analizę i monitoring parametrów strony WWW,

MOZ DA – (ang. Domain Authority) wskaźnik przewidujący prawdopodobieństwo, w przedziale od 0 do 100, znalezienia się strony WWW na stronie z wynikami wyszukiwania (SERP),

MOZ PA – (ang. Page Authority) wskaźnik przewidujący prawdopodobieństwo, w przedziale od 0 do 100, znalezienia się konkretnego adresu URL na stronie z wynikami wyszukiwania (SERP),

MOZ EUID – (ang. External Equity Links) liczba hiperłączy przychodzących do strony WWW wykrytych przez narzędzie MOZ i uznanych za dobrej jakości,

PageRank – algorytm rankingowy dla wyszukiwarek opracowany w 1998 roku przez Larry'ego Page'a i Sergeya Brina,

PHP – skryptowy język programowania zaprojektowany do generowania stron WWW i budowania aplikacji webowych,

PPC – (ang. Pay Per Click) sposób rozliczania kampanii internetowych, w którym reklamodawca płaci za pojedyncze kliknięcie, a nie za wyświetlenie reklamy,

RSS – (ang. Really Simple Syndication) oparta na języku XML technika przesyłania nagłówków wiadomości publikowanych na blogach lub stronach WWW,

SEM – (ang. Search Engine Marketing) kompleksowe działania obejmujące marketing w wyszukiwarkach internetowych,

SEO – (ang. Search Engine Optimization) proces optymalizacji strony WWW w celu zwiększenia jej widoczności w organicznych wynikach wyszukiwania pod określonymi słowami kluczowymi,

SSL – (ang. Secure Socket Layer) protokół sieciowy używany do bezpiecznych połączeń internetowych, standard szyfrowania na stronach WWW,

SERP – (ang. Search Engine Results Page) strona wyników wyszukiwania wyświetlana przez wyszukiwarki internetowe w odpowiedzi na określone zapytania użytkowników,

TTFB – (ang. Time to First Byte) czas liczony od momentu wysłania przez użytkownika zapytania do serwera WWW do momentu odebrania pierwszego bajta danych,

TCP/IP – (ang. Transmission Control Protocol / Internet Protocol) zestaw protokołów określających wzajemną komunikację i wymianę danych w sieci internet,

TF – (ang. Trust Flow) wskaźnik wiarygodności profilu hiperłączy dostępny w narzędziu Majestic,

URL – (ang. Uniform Resource Locator) format adresowania zasobów w sieci internet i sieciach lokalnych,

W3C – (skrót od World Wide Web Consortium) organizacja zajmująca się ustalaniem standardów transferu danych w ramach protokołu HTTP,

WAP – (ang. Wireless Application Protocol) protokół umożliwiający dostęp do stron WWW poprzez urządzenia mobilne,

WCAG – (ang. Web Content Accessibility Guidelines) wytyczne dotyczące ułatwień w dostępie do treści publikowanych w sieci internet.

Wprowadzenie

Algorytmy PageRank i HITS to fundamenty dzisiejszych algorytmów rankingowych stosowanych w wyszukiwarkach, wykorzystujących do oceny jakości strony WWW strukturę hiperłączy, zgodnie z zasadą, że jakość ta jest mierzona liczbą odwołań z innych stron [1]. Regularnie wzrastająca liczba nowych serwisów internetowych wymusiła na twórcach wyszukiwarek zmianę algorytmów rankingowych w taki sposób, aby o jakości strony WWW nie decydowała tylko liczba odwołań, lecz również treść i struktura tekstu, wykorzystane technologie i optymalizacja kodu HTML.

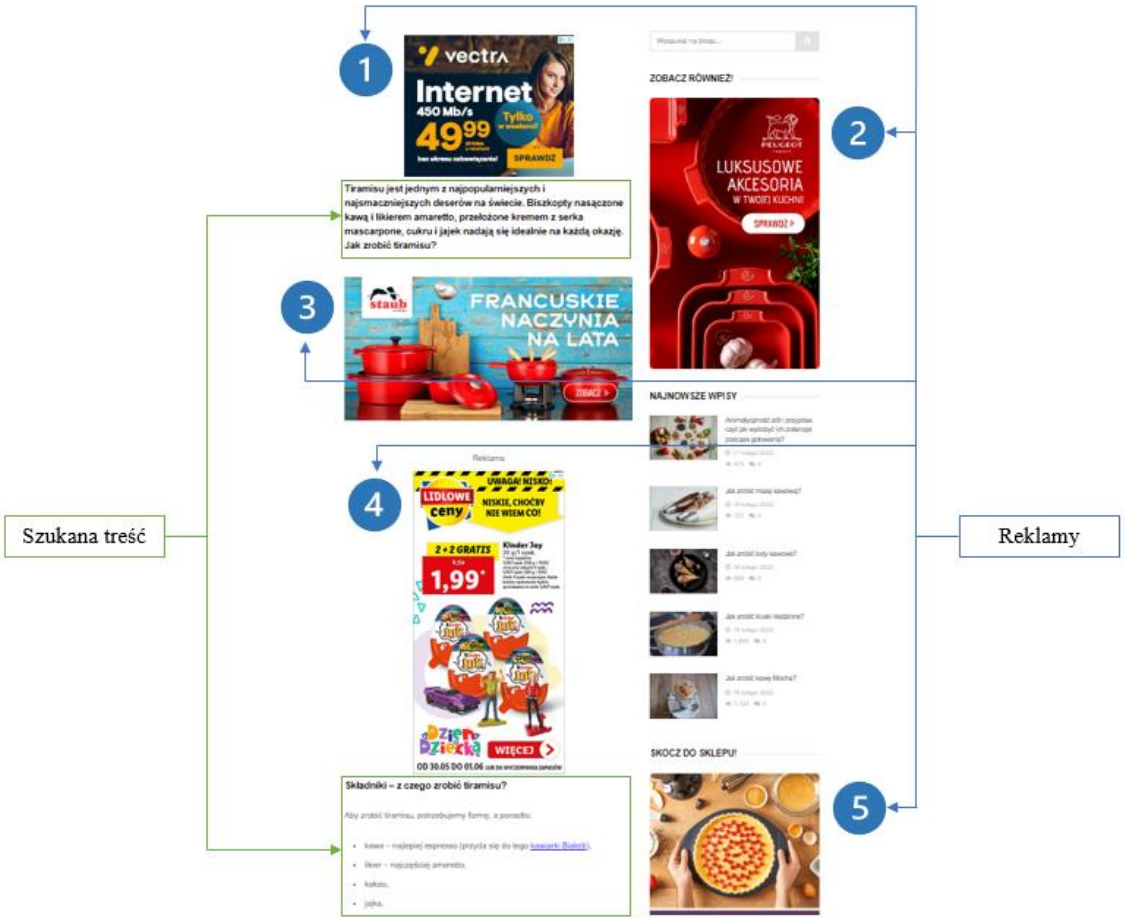
Badacze na całym świecie nieustannie podejmują próby odkrycia najważniejszych czynników wpływających na ranking w czołowych wyszukiwarkach, takich jak Google, Bing, Yahoo!, Yandex czy Baidu [2]. Do dziś nie opracowano skutecznej metody umożliwiającej odkrycie wszystkich takich czynników, co może wynikać z częstych zmian w algorytmach wyszukiwarek, jednak nie zniechęca to naukowców do dalszych badań. Zaowocowało to powstaniem nowych miar i metod, z rankingiem MOZ na czele [3], ułatwiających wyznaczanie wskaźników jakości strony WWW, na podstawie których można przewidzieć wzrost lub spadek pozycji rankingowej strony WWW w wyszukiwarkach.

Wiedza na temat zasad działania algorytmów rankingowych to najpilniej strzeżona tajemnica twórców wyszukiwarek. Biorąc pod uwagę dzisiejszą rolę wyszukiwarek w biznesie, próby odkrycia tych zasad z pewnością nie ustaną i będą przedmiotem wielu badań i publikacji naukowych.

Motywacja badań i uzasadnienie wyboru tematu

Internet stał się dziś kluczowym kanałem komunikacji większości firm i organizacji, a także zwykłych użytkowników. Nowo powstające strony WWW wypełniają internet treściami niepewnej jakości [4]. Wyszukiwarki starają się dostarczać treści optymalnych dla użytkowników, wykorzystując w tym celu algorytmy analizujące strukturę hiperłączy i techniki wpływające na optymalizację w obrębie strony WWW. Użytkownicy zaś, wyświetlając w przeglądarce adres strony uzyskany z wyszukiwarki, oczekują szybkiej i wyczerpującej informacji na przesłane zapytanie. Z praktyki wiadomo, że w wynikach wyszukiwania znajdują się odnośniki do zasobów budzących wiele zastrzeżeń związanych z ich jakością, np. stron zawierających publikacje o niskiej wartości merytorycznej lub treść powieloną z innych serwisów. Dotyczy to również stron WWW, na których nie jest zachowana równowaga między rozmiarem treści a wielkością kodu, ze względu na

nadmiar elementów niepowiązanych merytorycznie, przeważnie modułów reklamowych. Przykład takiej strony przedstawiono na rysunku 1. Zanim użytkownik dotrze do miejsca, gdzie prezentowana jest właściwa treść, która jest odpowiedzią na zapytanie zadane w wyszukiwarce, jest zmuszony do przejrzenia co najmniej pięciu reklam graficznych, co znacznie zaniża użyteczność takiej strony.



Rysunek 1. Liczba reklam w stosunku do objętości tekstu na stronie WWW z przepisami kulinarnymi. Opracowanie własne na podstawie strony https://www.garneczki.pl/blog/jak-zrobic-tiramisu/, maj 2022 roku

W wynikach wyszukiwań znajdują się również odnośniki do stron nieistniejących, generujących w przeglądarkach komunikaty o braku szyfrowania lub błędach w składni kodu HTML, co w praktyce uniemożliwia ich poprawne użytkowanie. Choć algorytmy wyszukiwarek starają się szybko reagować na tego typu problemy przez aktualizację pozycji rankingowych, nie jest możliwe całkowite ich wyeliminowanie.

Dostępne narzędzia informatyczne, takie jak MOZ Analytics, umożliwiają analizę techniczną i ocenę jakości strony WWW za pomocą własnych algorytmów rankingowych. Narzędzia te są przeznaczone dla przedsiębiorstw w ramach płatnych

abonamentów, co może być przeszkodą dla małych firm i twórców stron prywatnych. Opracowanie konkurencyjnego systemu umożliwiającego bezpłatną analizę techniczną i ocenę jakości strony WWW w czasie rzeczywistym było motywacją do zaprojektowania algorytmu rankingowego ISOWQ Rank i systemu rankingowego ISOWQ.

Impulsem do zaprojektowania nowego algorytmu rankingowego była próba odkrycia czynników rankingowych wpływających na ranking MOZ z zastosowaniem badań porównawczych. Należy przy tym podkreślić, że zasada działania algorytmu MOZ, na którym oparty jest płatny system analityczny MOZ Analytics, nie jest publicznie znana. Założono, że czynniki wpływające na ranking są związane z analizą treści i struktury tekstu, optymalizacją kodu źródłowego, użytych technologii oraz z detekcją problemów technicznych występujących na stronie i serwerze WWW. Biorąc pod uwagę renomę, jaką cieszy się narzędzie MOZ wśród specjalistów SEO, stworzenie konkurencyjnego algorytmu rankingowego o zbliżonej skuteczności umożliwiłoby poznanie czynników wpływających nie tylko na ranking MOZ, ale również na ranking w wyszukiwarkach.

W ramach tej pracy dokonano przeglądu literatury, korzystając ze stron WWW skierowanych do naukowców, takich jak Scopus i IEEE Xplore. Kwerendę publikacji przeprowadzono w okresie od 5 listopada 2021 roku do 30 maja 2022 roku.

Cele i teza pracy

Celem niniejszej rozprawy doktorskiej były badania związane z opracowaniem i oceną skuteczności algorytmu rankingowego ISOWQ Rank, który nadaje stronom WWW określoną wartość, oznaczającą ich jakość. Badania miały potwierdzić, czy istnieje dodatnia korelacja pomiędzy algorytmami ISOWQ Rank i MOZ, a także wykazać, jaki wpływ na tę korelację mają poszczególne czynniki rankingowe, a w szczególności treść i struktura tekstu na stronie WWW. W trakcie badań przeanalizowano aktualną wiedzę z zakresu metod ustalania rankingu dla serwisów internetowych i technik optymalizacji w obrębie strony WWW i poza nią. Skuteczność algorytmu zmierzono w badaniach porównawczych, w których wykazano dodatnią korelację pomiędzy punktacją uzyskaną za pomocą algorytmu ISOWQ Rank a punktacją obliczoną przez algorytm MOZ.

Drugim celem pracy była implementacja algorytmu ISOWQ Rank i jego praktyczne zastosowanie. W ramach prac zaprojektowano i wdrożono system informatyczny składający się z dwóch niezależnych segmentów. Pierwszy segment objął podsystem odpowiedzialny za analizę danych, drugi zaś był przeznaczony do ich prezentowania i obsługi

użytkownika. W trakcie 11-letniej pracy systemu wykonano ponad 1,3 mln analiz stron WWW. Wszystkie zebrane dane udostępniono bezpłatnie na stronie internetowej projektu pod adresem www.isowq.org.

Tezę pracy można ująć w formie następującego stwierdzenia: algorytm ISOWQ Rank w sposób optymalny wyznacza ranking stron WWW przez nadanie im określonej wartości, oznaczającej ich jakość. Stosowanie się do wytycznych w opracowanej metodyce oceny jakości strony WWW za pomocą algorytmu ISOWQ Rank pozwala zwiększyć wartość rankingową wyznaczoną przez algorytm MOZ, co w konsekwencji może mieć pozytywny wpływ na pozycję rankingową w wyszukiwarkach internetowych.

Struktura pracy

Praca doktorska jest podzielona na cztery rozdziały.

Rozdział 1. zawiera podstawową wiedzę pochodzącą z literatury na temat wyszukiwarek stron WWW oraz rodzajów wyników w wyszukiwarkach z podziałem na płatne i organiczne. Ponadto przedstawiono metody optymalizacji pod wyszukiwarki stosowane zarówno bezpośrednio na stronie WWW, jak i poza nią.

W rozdziale 2. przedstawiono podstawową wiedzę pochodzącą z literatury na temat fundamentów dzisiejszych algorytmów rankingowych i badań naukowych związanych z próbami odkrycia czynników rankingowych wyszukiwarek. Omówiono zasadę działania algorytmów rankingowych PageRank i HITS na przykładowych strukturach połączeń pomiędzy stronami WWW. Na potrzeby prezentacji algorytmów PageRank i HITS w niniejszej pracy opracowano narzędzie programistyczne w języku Python umożliwiające obliczenie wartości rankingowych dla dowolnej struktury linkujących się wzajemnie stron WWW. Narzędzie jest dostępne bezpłatnie na portalu GitHub[1].

W rozdziale 3. szczegółowo opisano zasadę działania algorytmu ISOWQ Rank. Przedstawiono jego pseudokod i metodę implementacji. Ponadto omówiono architekturę systemu ISOWQ oraz jego budowę i strukturę baz danych, a także zaprezentowano przykładowy raport techniczny dla strony WWW.

W rozdziale 4. przedstawiono wyniki badań porównawczych algorytmów rankingowych ISOWQ Rank i MOZ. Ponadto zaprezentowano kod źródłowy w języku R, za pomocą którego wykonano obliczenia na potrzeby przeprowadzonego badania.

[1] PageRank-HITS, https://github.com/mariuszduka/PageRank-HITS, kwiecień 2022 r.

Wyniki badań i wnioski

Weryfikacja empiryczna opiera się na porównaniu wyników uzyskanych za pomocą algorytmów ISOWQ Rank i MOZ dla wybranej grupy stron WWW. Grupa badawcza obejmuje serwisy internetowe zróżnicowane pod względem wykorzystanych technologii, budowy kodu HTML, treści i struktury tekstu. Do oceny ich wspólnej zależności wykorzystano współczynnik korelacji τ-Kendalla przy zadeklarowanym poziomie istotności 0,05. Otrzymane wyniki wskazują jednoznacznie na dodatnią korelację pomiędzy punktacją uzyskaną z użyciem algorytmu ISOWQ Rank a punktacją MOZ, co oznacza, że wzrost jednej powinien spowodować wzrost drugiej.

Analizując wyniki badań, można wnioskować, że optymalizacja treści i struktury tekstu na stronie WWW ma istotne znaczenie przy ustalaniu rankingu w wyszukiwarkach. Świadczy o tym dodatnia korelacja pomiędzy czynnikiem rankingowym ISOWQ Rank związanym z oceną treści na stronie WWW a punktacją MOZ. Fragmenty badań opublikowano w 2020 roku w recenzowanym międzynarodowym kwartalniku „Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska", wydanym w języku angielskim przez Politechnikę Lubelską [5].

Efektem niniejszej pracy jest potwierdzenie skuteczności algorytmu rankingowego ISOWQ Rank w ocenie jakości stron WWW. Ocena ta opiera się na aktualnej wiedzy związanej z optymalizacją w obrębie strony WWW i poza nią, której celem jest wzrost pozycji rankingowej w wyszukiwarkach. Dodatnia korelacja z algorytmem MOZ, którego zasada działania nie jest publicznie znana, świadczy o właściwie dobranych parametrach i nadaniu im odpowiednich wag podczas oceny jakości stron WWW.

1. SEO jako metoda budowania ruchu z wyników organicznych

W niniejszym rozdziale zebrano podstawową wiedzę na temat pierwszych wyszukiwarek i katalogów internetowych oraz przedstawiono, jak zmieniał się rynek wyszukiwarek w ostatnich trzech dekadach. Zebrano aktualną wiedzę na temat dominujących rodzajów wyników wyszukiwania prezentowanych w wyszukiwarce Google i metod wpływania na pozycje rankingowe w wynikach organicznych. Omówiono popularne narzędzia informatyczne umożliwiające analizę najważniejszych obszarów związanych z technicznym audytem stron WWW.

1.1. Pierwsze wyszukiwarki i katalogi internetowe

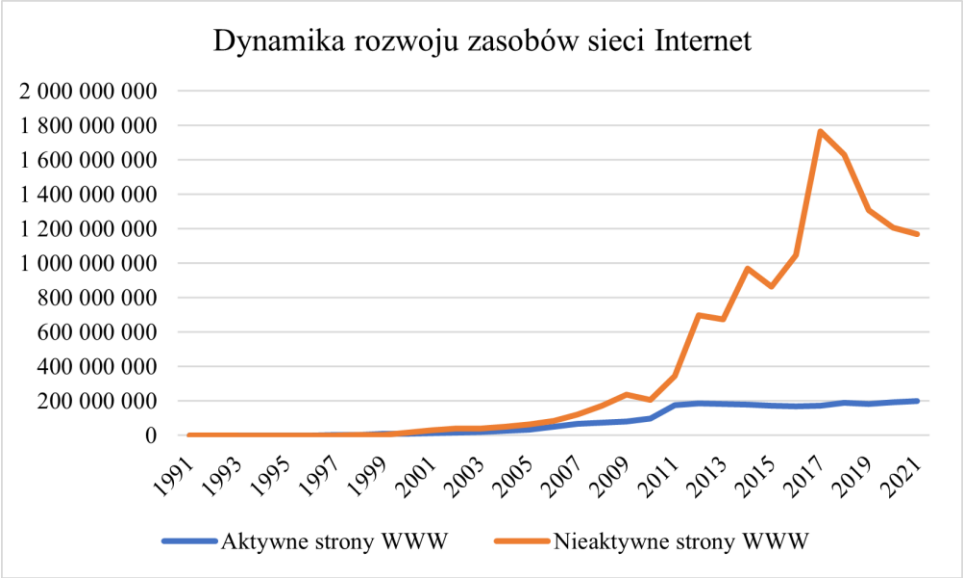
Wyszukiwarki odgrywają integralną rolę w życiu wielu ludzi, którzy często nie zdają sobie z tego sprawy. Stały się one potężnymi narzędziami do uzyskiwania przydatnych informacji rozproszonych w sieci [6]. Zwykły zakup na portalach Allegro czy eBay zazwyczaj rozpoczyna się od wyszukania produktu, podobnie jak rezerwacja wyjazdu na urlop najczęściej rozpoczyna się od wyszukania hotelu. Odpowiedź na każde zapytanie to kolejne wyszukanie w sieci.

Internet, jaki znamy, istniał przez wiele lat, zanim pojawiła się pierwsza strona WWW. Protokół internetowy (IP), jak również model warstwowej struktury protokołów komunikacyjnych (TCP/IP) to technologie wynalezione przez Boba Kahna i Vinta Cerfa już we wczesnych latach 70. XX wieku w Agencji Zaawansowanych Projektów Badawczych Departamentu Obrony Stanów Zjednoczonych (DARPA) [7]. Militarny charakter technologii sprawił, że nie była ona dostępna dla sektora cywilnego, poza lokalnymi sieciami akademickimi skupionymi we wspólnej sieci ARPANET [8]. Kiedy na początku lat 90. XX wieku protokół TCP/IP wszedł do komercyjnego użytku, nastąpił rozkwit branż związanych z usługami internetowymi. Fundamentalne znaczenie w owym czasie miało wynalezienie przez Tima Bernersa-Lee w 1989 roku technologii WWW. Berners-Lee wykorzystał istniejące protokoły HTTP i TCP/IP do budowy systemu składającego się z pierwszego serwera WWW i katalogu internetowego o nazwie WorldWideWeb [9].

System ten nadawał każdej stronie WWW ujednolicony format adresowania (URL), na podstawie którego była ona udostępniana publicznie w sieci. Berners-Lee stworzył również język HTML, oparty na znacznikach SGML-CERN, do formatowania treści tekstowych.

Od 1991 roku, kiedy to pojawiła się pierwsza strona WWW (info.cern.ch) [10] oraz pierwsze zręby języka HTML i serwerów WWW, przestrzeń internetową zapełniło ponad 1,4 mld stron WWW na ponad 233 mln unikatowych domen. Brytyjska firma Netcraft od 1995 roku monitoruje zasoby sieci internet, biorąc pod uwagę wykorzystane w witrynach internetowych technologie oraz oprogramowanie na serwerach dostawców usług hostingowych. Netcraft na podstawie analizy adresu IP oraz kodu źródłowego serwisu WWW oblicza liczbę aktywnych stron internetowych, tzn. takich, które prezentują konkretną treść, z pominięciem serwisów „w budowie” (ang. under construction), przekierowań lub domen wskazujących na identyczną treść, np. z włączoną u rejestratora usługą „parkowania”. Na podstawie cyklicznie przeprowadzanych analiz szacuje się, że w internecie

dostępnych jest ponad 200 mln aktywnych serwisów. Dynamikę rozwoju zasobów sieci internet przedstawia rysunek 2.



Rysunek 2. Dynamika rozwoju sieci internet od 1991 do 2021 roku, opracowanie własne na podstawie danych z serwisu netcraft.com

Rosnąca liczba nowych serwisów internetowych wymusiła powstanie wyszukiwarek, czyli systemów ułatwiających użytkownikom odszukanie w sieci konkretnych informacji. Dane gromadzone w takich systemach opierały się na analizie stron WWW pod kątem zawartych na nich treści oraz na analizie topologii sieci hiperłączy. Wiele z nich już nie istnieje, a w ich miejsce powstają nowe, starając się dogonić obecnych liderów. Z biegiem lat konkurencja wymusiła możliwość wyszukiwania również treści multimedialnych, co początkowo sprawiało algorytmom wiele problemów [11]; dziś funkcjonalność ta jest wbudowana w prawie każdej profesjonalnej wyszukiwarce. Przyrost nowych stron WWW, wynikający m.in. z chęci pojawienia się w wynikach wyszukiwania, początkowo utrudniał przedsiębiorstwom uruchomienie serwisu WWW pod dowolnie wybraną nazwą w domenie najwyższego poziomu – .com, co wymusiło utworzenie dodatkowych typów domen – narodowych i funkcjonalnych, takich jak utworzona w 2005 roku domena .eu.

Za pierwszą wyszukiwarkę uważa się Archie, uruchomioną w 1990 roku [12]. Później doszły m.in. W3Catalog, WebCrawler i Lycos. W połowie lat 90. XX wieku pojawiło się wiele nowych, takich jak Excite, AltaVista i Yahoo!. Był to okres pierwszych analiz ich zasobów [13] oraz pomiarów jakości zwracanych wyników [14].

Najpopularniejsza do dziś jest powstała w 1998 roku wyszukiwarka Google, która ze swoim algorytmem PageRank wytyczyła nowe standardy indeksowania i rankingowania zasobów internetowych. Po roku działalności liczba zindeksowanych adresów URL w wyszukiwarce wynosiła ponad 350 mln [15], a liczba indeksacji w kolejnych latach była już wyrażana w miliardach [16]. Szacuje się, że w 2000 roku liczba serwerów obsługujących wyszukiwarkę Google wynosiła 25 tys., a w 2010 roku liczba ta wzrosła do 900 tys. [17]. W tym samym roku co Google powstał katalog internetowy DMOZ[2], który w swoim najlepszym okresie miał bazę ponad 5 mln stron WWW. Obecność w katalogu DMOZ często była utożsamiana z wyższym rankingiem w wyszukiwarkach, a prestiżu dodawało to, że weryfikacją dodawanych stron zajmowali się wyselekcjonowani specjaliści. Moment powstania pierwszych wyszukiwarek i katalogów internetowych zbiega się z nadejściem portali społecznościowych, na czele z powstałym w 1997 roku serwisem „Six Degrees" [18].

Począwszy od pierwszych wyszukiwarek treści ze stron WWW pobierano [19] i indeksowano [20] za pomocą botów (ang. crawlers). Wówczas, przy niewielkiej liczbie stron i wyszukiwarek, nie generowały one tyle ruchu co dziś [21]. Przez lata pojawiały się nowe algorytmy, zwiększające wydajność botów [22], mierzono i porównywano ich zachowanie [23] i ruch generowany na stronach WWW [24], a także projektowano je do semantycznej analizy treści [25]. Ponieważ aktywność botów może budzić obawy dotyczące bezpieczeństwa i wydajności serwerów WWW [26], w wielu badaniach wyodrębniono cechy ruchu odróżniające boty od prawdziwych użytkowników [27] i opracowano metody ich automatycznej klasyfikacji [28]. Ponadto zwiększająca się liczba nowych stron internetowych wymagała opracowania metod klasyfikacji [29], kategoryzacji [30] i rekomendacji [31] botów, aby wyniki wyszukiwania w jak największym stopniu odpowiadały na konkretne zapytania [32].

Zagospodarowania globalnego rynku wyszukiwarek podjęła się również firma Microsoft, która równocześnie z Google wprowadziła swoją wyszukiwarkę MSN Search, licząc na to, że duża popularność systemu Windows przełoży się na wzrost popularności nowego narzędzia. Niestety, mimo wzmożonej aktywności marketingowej i kolejnych zmian nazwy wyszukiwarki – w 2006 roku na Microsoft Live, w 2007 roku na Live Search i ostatecznie w 2009 roku na Bing – nie udało się jej uzyskać więcej niż 10% rynku

[2] Wielojęzyczny katalog stron WWW, działający w latach 1998–2017.

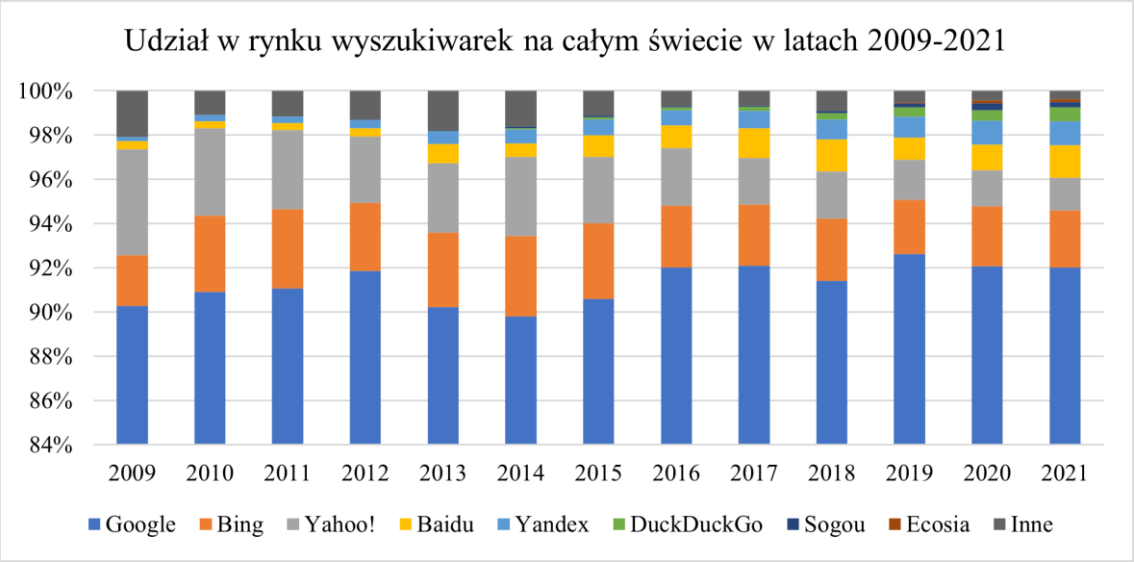
wyszukiwarek. Sukcesu nie zapewniło jej nawet to, że generowała bardziej precyzyjne niż wyszukiwarka Google wyniki uwzględniające lokalizację [33].

Choć wyszukiwarki Google, Bing czy Yahoo! są od lat światowymi liderami, istnieją lokalne wyszukiwarki, które kiedyś miały lub mają do dziś większy od nich udział w rynku w danym kraju [34]. Do grona takich wyszukiwarek należą rosyjski Yandex (1997) i chińskie Baidu (2000). Dominacja wyszukiwarki Yandex nad Google w Rosji wynikała z lepszego zrozumienia gramatyki języka rosyjskiego, co zaowocowało dokładniejszymi wynikami wyszukiwania rosyjskich stron WWW. Dzięki rozpowszechnieniu systemu Android i przeglądarki Chrome wyszukiwarka Google zdobywała coraz większy udział w wyszukiwaniu na rynku rosyjskim, w szczególności wśród młodych i bardziej zaawansowanych użytkowników, co ostatecznie doprowadziło Google na pozycję lidera. Jedynym krajem, w którym wyszukiwarka Google nie dominuje, są Chiny. Do najpopularniejszych wyszukiwarek internetowych w Państwie Środka należą Baidu z udziałem 59,85%[3] i Bing z udziałem 15,83%, a udział wyszukiwarki Google to zaledwie 3,70%. W Chinach wyszukiwarka Google miała ograniczone pole działania, głównie ze względów politycznych [35], co uniemożliwiało jej znaczący rozwój na tym rynku.

Obecnie niezwykle trudno byłoby wprowadzić nową wyszukiwarkę i konkurować na rynku zdominowanym przez Google; wymagałoby to dużej pomysłowości twórców [36]. Trudno też dziś sobie wyobrazić realnego globalnego konkurenta dla Google czy dla Baidu w Chinach. W tym kontekście interesującymi projektami są wyszukiwarki DuckDuckGo i Ecosia, które wprawdzie mają udziały liczone w ułamkach procenta, ale starają się odnaleźć na globalnym rynku wyszukiwarek. Fundamentem założonej w 2009 roku wyszukiwarki DuckDuckGo była koncentracja na prywatności i niegromadzeniu żadnych informacji, które umożliwiają identyfikację użytkownika. W odróżnieniu od innych wyszukiwarek po kliknięciu danego wyniku przekierowanie następuje w sposób, który zapobiega wysyłaniu wyszukiwanych słów kluczowych do strony WWW. Dzięki temu strony WWW nie wiedzą, jak dany użytkownik znalazł je w sieci. Podobnie powstała w 2009 roku wyszukiwarka Ecosia postawiła na prywatność i zrezygnowała z gromadzenia informacji o swoich użytkownikach. Warto dodać, że pomysłem na zaistnienie na rynku było przekazywanie części dochodów z reklam do organizacji społecznych zajmujących się sadzeniem drzew, m.in. w Brazylii, Etiopii, Burkina Faso, na

[3] Na podstawie serwisu StatCounter.com, październik 2022 r.

Madagaskarze, w Senegalu i Indonezji. W odróżnieniu od pozostałych wyszukiwarek, które samodzielnie gromadzą dane, Ecosia udostępnia wyniki dostarczone przez Bing i Yahoo!.



Rysunek 3. Udział w rynku wyszukiwarek na całym świecie w latach 2009–2021, opracowanie własne na podstawie danych z serwisu StatCounter.com

Globalny udział w rynku wyszukiwarek w latach 2009–2021 jest pokazany na rysunku 3. Dane dotyczące wyszukiwarki Yandex obejmują wyszukiwania globalne i lokalne w domenie yandex.ru. Wyszukiwarki AltaVista, Ask i WebCrawler pominięto ze względu na ich śladowy udział.

Przez lata zmieniały się nie tylko wyszukiwarki i ich algorytmy rankingowe, ale również sposób, w jaki korzystano z interfejsu i wyników wyszukiwania. Analiza badań okulograficznych wyszukiwarek internetowych przeprowadzonych w latach 2004–2017 i 2004–2019 wskazuje na postępującą zmianę sposobu postrzegania wyników w wyszukiwarkach internetowych [37][38]. Początkowo użytkownicy odbierali treści w formacie zbliżonym do litery F, co było fundamentem koncepcji złotego trójkąta, w którym znajdowały się hiperłącza do najczęściej odwiedzanych stron WWW. W obszar ten z czasem przeniesiono hiperłącza sponsorowane, które dotychczas znajdowały się po prawej stronie wyników wyszukiwania. Istotnym wnioskiem z analizy badań okulograficznych jest wskazanie, że istnieje tendencja do zmian interfejsu wyszukiwarek przez wprowadzanie nowej funkcjonalności, co jest wynikiem postępującego wzrostu udziału urządzeń mobilnych wykorzystywanych do wyszukiwania informacji.

Wyszukiwanie na urządzeniach mobilnych jest dziś czymś naturalnym, a zgodnie z raportem Google[4] w 2015 roku liczba takich wyszukiwań przekroczyła liczbę tych, które pochodziły z komputerów stacjonarnych. Trend ten utrzymuje się do dziś – telefony komórkowe i tablety stały się podstawowymi urządzeniami wyszukiwania. W światowym rankingu urządzenia mobilne mają 58,33%[5] udziału w liczbie wyszukiwań, a urządzenia stacjonarne i tablety, odpowiednio, 39,65% i 2,02%. Warto dodać, że możliwość wyszukiwania mobilnego jest dostępna od 1998 roku, kiedy to wprowadzono do użytku protokół aplikacji bezprzewodowych (WAP). Była to pierwsza mobilna funkcja umożliwiająca przeglądanie stron WWW, jaką udostępniono w telefonach komórkowych. Rozwój standardów bezprzewodowego przesyłania danych spowodował, że technologia WAP została wyparta przez GPRS, następnie EDGE, HSDPA, a wreszcie LTE, 4G i 5G. Rozpowszechnienie się urządzeń mobilnych sprawiło, że każdy mógł mieć przy sobie niewielki komputer, umożliwiający dostęp do nieograniczonego wyszukiwania i przeglądania informacji. Nie bez znaczenia jest tu silna pozycja systemu Android, który poprzez wbudowaną przeglądarkę Chrome umacnia dominującą pozycję wyszukiwarki Google.

Rozwój wyszukiwarek przyniósł również możliwość wyszukiwania głosowego, kiedy to użytkownik zadaje pytanie, zamiast wpisywać słowa kluczowe w przeglądarce. Choć możliwość ta wydaje się interesująca, często jest krytykowana, ponieważ pomimo tego, że generowany przez wyszukiwarkę głos brzmi realistycznie, niewiele osób chce słuchać, jak komputer czyta im zawartość stron WWW z listy 10 najlepszych wyników. Wyszukiwarki starają się poprawiać dokładność odpowiedzi za pomocą wyszukiwania semantycznego [39], polegającego na próbie zrozumienia języka naturalnego, co biorąc pod uwagę liczbę języków, jest zadaniem bardzo trudnym [40].

Przyszłością wyszukiwarek, oprócz zaimplementowania ich w systemach opartych na internecie rzeczy (IoT) [41], z pewnością będzie wykorzystanie sztucznej inteligencji, [42] m.in. do porównywania i kupowania usług takich jak bilety lotnicze i noclegi w hotelach [43]. Produkty, koszyk na zakupy i płatności mogłyby być prezentowane bezpośrednio na stronie z wynikami wyszukiwania. Wyszukiwarki otrzymywałyby prowizję od całkowitej transakcji, a nie tylko opłatę za kliknięcie.

Obecnie podział globalnego rynku wyszukiwarek jest następujący: na pierwszym miejscu, z udziałem 92,37%, jest rozwijana od 25 lat Google[6], dalsze zajmują Bing

[4] Raport dotyczył 10 krajów, w tym USA i Japonii.
[5] Na podstawie serwisu StatCounter.com, październik 2022 r.
[6] Ibidem.

(3,57%), Yahoo! (1,31%), Yandex (0,82%), DuckDuckGo (0,61%) i Baidu (0,58%), a pozostałym przypada 0,74%. Należy przy tym zwrócić uwagę na to, że przedstawione udziały oszacowano na podstawie analizy ruchu na kontrolnej grupie stron WWW. Biorąc pod uwagę specyfikę rynku chińskiego związaną z ograniczeniami nakładanymi na wyszukiwarki internetowe [44] oraz liczbę ludności Chin, z dużym prawdopodobieństwem można założyć, że udział wyszukiwarki Baidu jest niedoszacowany.

1.2. Rodzaje wyników w wyszukiwarkach

Strona z wynikami wyszukiwania (SERP) wyświetlana przez wyszukiwarki w odpowiedzi na określone zapytania użytkowników obejmuje bezpłatne wyniki organiczne, czyli naturalne wyniki wyszukiwania, o których kolejności decyduje algorytm wyszukiwarki, oraz płatne, w formie reklam tekstowych i produktowych. O pozycji rankingowej w wynikach organicznych decyduje algorytm wyszukiwarki, natomiast w przypadku reklam płatnych decydującym czynnikiem jest mechanizm aukcyjny platformy.

Przez lata SERP zmieniał swój wygląd i formę, głównie za sprawą Google, które, jako lider, wytycza trendy w formach prezentowania wyników wyszukiwań. Pierwotnie SERP przedstawiał listę hiperłączy do stron zawierających wyszukiwane słowo kluczowe i choć początkowo rezultaty wyszukiwań były zbieżne w większości popularnych wyszukiwarek [45], z biegiem lat wymagania użytkowników wymusiły zaprojektowanie nowych i bardziej funkcjonalnych form prezentacji wyników [46].

Algorytmy wyszukiwarek odpowiadają na zapytania już nie tylko przez analizę liczby hiperłączy polecających daną stronę, ale przede wszystkim starają się zrozumieć kontekst i intencję pytającego. Aby zrealizować to założenie, Google wdraża i rozwija różne rodzaje wyszukiwania oraz funkcje uatrakcyjniające i ułatwiające poszukiwanie informacji, miejsc czy produktów [47].

Wyszukiwarka Google zajmuje pierwsze miejsce także pod względem oferowanych rodzajów wyników wyszukiwania [48]. Żadna inna wyszukiwarka nie ma tak wielu różnych typów; poniżej opisano te najczęściej występujące.

1.2.1. Hiperłącza do podstron z wewnętrzną wyszukiwarką

Rozszerzeniem wyników organicznych, najczęściej przy zapytaniach o konkretną markę, są hiperłącza do podstron, pole wyszukiwania do podstron w obrębie wynikowej strony WWW, logo oraz informacje o firmie czy instytucji. Forma prezentacji hiperłączy ma

pomóc użytkownikowi w dotarciu do najbardziej wartościowych, według algorytmu, treści. Właściciel strony WWW nie ma wpływu na wyświetlanie tych elementów, decyduje o tym algorytm wyszukiwarki.

1.2.2. Karty informacyjne

Karty informacyjne (ang. rich snippets) wprowadziło w 2012 roku Google, jako odpowiedź na zmieniający się sposób kierowania zapytań [49]. Celem zmian było generowanie wartościowych stron wyników wyszukiwania — z ciekawą i wiarygodną treścią. Karty informacyjne wyświetlają specjalnie oznaczone dane strukturalne zawarte w kodzie strony WWW, takie jak cena i jej zakres, ocena, liczba opinii lub głosów, data publikacji, imię i nazwisko autora. Branża kulinarna ma własne rodzaje wyników wyszukiwania, rozszerzone o miniatury zdjęć przepisów, dzięki którym użytkownik może również dokonać wyboru wzrokowo.

1.2.3. Polecane fragmenty z odpowiedzią

Polecane fragmenty z odpowiedzią (ang. direct answer snippets), nazywane również fragmentem polecanym (ang. featured snipped), zostały wprowadzone w 2016 roku przez Google, jako zwięzła odpowiedź w formie akapitu, listy lub tabeli. Fragmenty polecane dają szybką odpowiedź na zapytanie użytkownika sformułowane w formie pytania, a także umożliwiają użytkownikom uzyskanie odpowiedzi bez konieczności odwiedzania źródłowej witryny internetowej [50].

Najczęściej spotykanymi typami odpowiedzi bezpośrednich są odpowiedzi w formie akapitu, ponieważ ten rodzaj odpowiedzi jest najbardziej czytelny, a jednocześnie najwygodniejszy do odczytania przez systemy wyszukiwania głosowego. Forma listy najczęściej pojawia się w przepisach kulinarnych, a tabela zazwyczaj występuje w przypadku zapytań dotyczących porównania cen lotów lub danych związanych z produktami finansowymi [50].

Uzyskanie pozycji zerowej w wynikach wyszukiwania, czyli w polecanym fragmencie z odpowiedzią, jest pożądane przez twórców stron WWW i podkreśla ekspercki charakter serwisu. Aby zwiększyć prawdopodobieństwo, że treść serwisu będzie wykorzystana przez Google w polecanym fragmencie, wymaga od autorów stron WWW odpowiedniego skomponowania treści, ze szczególnym rozróżnieniem słów kluczowych składających się z dwóch lub trzech wyrazów z krótkim opisem w formie akapitu bądź listy.

Ważną informacją dla twórców stron WWW jest również to, że strona musi znajdować się na pierwszej stronie SERP dla danego zapytania kluczowego [50].

1.2.4. Panel wiedzy

Panel wiedzy (ang. knowledge panel) to zbiór danych wyświetlany z prawej strony wyników wyszukiwania. Stanowi on wygodną formę dostarczania informacji na temat znanych osób, zespołów muzycznych, postaci historycznych, placówek naukowych i kulturalnych. Google, decydując o umieszczeniu informacji w panelu wiedzy, analizuje, podobnie jak w przypadku fragmentu polecanego, dane strukturalne na stronie WWW i jej pozycję rankingową [51].

1.2.5. Wizytówka firmy – Google Moja Firma

W związku z tym, że panel wiedzy jest generowany automatycznie przez algorytm, Google umożliwiło przedsiębiorstwom zarządzanie informacjami biznesowymi za pomocą firmowych wizytówek pojawiających się z prawej strony, w ramach darmowej platformy Google Moja Firma (ang. Google My Business). Aby można było dodać informację do firmowej wizytówki, firmę musi zarejestrować reprezentujący ją uprawniony organ [52]. Poprawnie przygotowana wizytówka, wzbogacona o zdjęcia i szczegółowy opis działalności, ma istotny wpływ na pozycje rankingowe w wyszukiwarce Google [53].

Głównymi elementami firmowej wizytówki są przyciski z hiperłączami do strony WWW, trasa dojazdu w serwisie Google Maps, dane teleadresowe, godziny pracy, a także uzupełniające informacje o lokalizacji. Wizytówka może dodatkowo zawierać fotografie, jak też sekcję z pytaniami i odpowiedziami oraz opiniami użytkowników dodanymi bezpośrednio w panelu Google Moja Firma lub recenzjami dodanymi na profilach społecznościowych. Dostępne są również przyciski umożliwiające dodanie nowej opinii lub fotografii.

1.2.6. Wyniki lokalne – Google Maps

SERP w wyszukiwarce Google może być uzupełniony o dodatkowe wyniki lokalne, najczęściej dotyczące usług w najbliższej okolicy (np. restauracje, hotele, mechanicy), w zależności od geograficznej lokalizacji użytkownika. Lista wyników jest generowana za pośrednictwem serwisu Google Maps, umożliwiającego wyszukiwanie obiektów, przeglądanie map, a także udostępniającego informacje o natężeniu ruchu ulicznego w czasie

rzeczywistym. Google Maps do ustalenia dokładnej lokalizacji osób i miejsc wykorzystuje system GPS, składający się z 27 satelitów orbitujących wokół Ziemi. Lista wyników lokalnych uwzględnia również takie informacje jak lokalizacja na mapie miasta, trasa dojazdu, odnośnik do strony firmowej, ocena użytkowników oraz numer telefonu [54].

1.2.7. Translator, pogoda, czas na świecie, kalkulator

SERP w wyszukiwarce Google cały czas się rozwija, dostarczając użytkownikowi coraz to nowej funkcjonalności. Celem twórców wyszukiwarek jest takie udostępnienie funkcji bezpośrednio w SERP, aby użytkownik nie musiał ich szukać na innych stronach WWW. To może umożliwić sprzedaż produktów lub usług bezpośrednio na stronie wyszukiwarki, co będzie miało przełożenie na wyniki finansowe ich właścicieli.

Wyszukiwarka Google umożliwia tłumaczenie tekstu za pomocą automatycznego tłumacza, który pozwala wybrać dowolną kombinację dwóch języków spośród kilkudziesięciu dostępnych. Dodatkowo tłumacz Google ma wbudowaną funkcję głosową umożliwiającą naukę wymowy danego słowa.

Wpisawszy w wyszukiwarkę „pogoda Bytom", otrzymamy informację o aktualnej temperaturze i warunkach atmosferycznych w konkretnym mieście, jak również prognozę na kolejne dni, prawdopodobieństwo opadów oraz prędkość i kierunek wiatru. Wyszukiwarka może poinformować o czasie lokalnym w dowolnym miejscu na świecie, a także dokonać za pomocą kalkulatora obliczeń matematycznych.

1.2.8. Płatne wyniki wyszukiwania – łącza sponsorowane

Łącza sponsorowane w wyszukiwarkach internetowych to popularna i efektywna forma reklamy internetowej. Zapewnia szybkie rezultaty i pozwala wygenerować duży ruch na stronie WWW [55]. Łącza sponsorowane są elementem działań SEM, czyli marketingu w wyszukiwarkach, w którego skład wchodzą ruch organiczny (SEO) oraz reklamy (PPC) [56]. Kampanie reklamowe polegają na wykupieniu powierzchni reklamowej w konkretnych wyszukiwarkach i portalach internetowych.

Reklamy w wyszukiwarkach dzielą się na reklamy tekstowe i graficzne. Główny cel reklam tekstowych to zwiększenie ruchu na stronie WWW, o czym świadczy to, że płaci się za nie na podstawie kosztu kliknięcia (CPC). Wyszukiwarki starają się eksponować takie reklamy, dopasowując je do jak najszerszej grupy wyszukiwanych słów kluczowych, aby zmaksymalizować swoje zyski [57]. Natomiast reklamy graficzne

z zastosowaniem elementów wideo mają na celu przede wszystkim budowanie wizerunku marki [58] i są rozliczane na podstawie kosztu tysiąca wyświetleń (CPM) [59].

Budowanie świadomości marki poprzez reklamę w wyszukiwarkach internetowych jest zaliczane do strategii „marketingu wychodzącego" (ang. outbound marketing), w odróżnieniu od „marketingu przychodzącego" (ang. inbound marketing), polegającego na podejmowaniu działań umożliwiających samodzielne odnalezienie nadawcy danego przekazu przez odbiorcę [60].

Czołową platformą reklamową jest założona w 2000 roku Google Ads[7], która pozwala na wyświetlanie łączy sponsorowanych w wynikach wyszukiwarki Google oraz na stronach współpracujących w ramach programu Google AdSense, sprzedawanych w najpopularniejszych modelach wyceny emisji reklam, CPC i CPM. Pozycja reklamy wśród innych reklam jest uzależniona od wysokości zadeklarowanej przez reklamodawcę ceny za kliknięcie oraz popularności reklamy, obliczanej za pomocą wskaźnika CTR, oznaczającego procent osób, które kliknęły w reklamę, w stosunku do liczby jej wyświetleń.

Reklamy w wyszukiwarkach oprócz tekstowego opisu i hiperłącza mogą zawierać zdjęcia, nazwę i cenę reklamowanego produktu. W wynikach wyszukiwania po nakierowaniu wskaźnika myszy na obiekt reklamy wyświetlane są również informacje o kosztach wysyłki.

1.3. Elementy SEO

SEO, czyli optymalizacja pod wyszukiwarki internetowe[8], to działania, które mają doprowadzić do osiągnięcia przez daną stronę WWW jak najwyższej pozycji rankingowej w wynikach wyszukiwania [61]. Zakres tych działań jest bardzo szeroki i związany z odpowiednim doborem słów kluczowych [62], optymalizacją treści strony WWW [63], jej strukturą, jak również optymalizacją kodu HTML [64], elementów graficznych [65] oraz powiązań z mediami społecznościowymi [66]. Najważniejszym celem działań SEO jest poprawa pozycji rankingowej strony WWW dla wybranych słów kluczowych w wynikach organicznych.

W praktyce działania SEO obejmują takie elementy jak:

- struktura treści – redakcja artykułów oraz opisów kategorii, ofert czy produktów [67],

[7] Nazwa Google Ads (wcześniej AdWords) obowiązuje od 24 lipca 2018 r.
[8] W Polsce przyjął się również termin „pozycjonowanie stron WWW".

- struktura strony WWW – modyfikacja elementów nawigacji, linkowanie wewnętrzne [68], dane strukturalne [69],
- warstwa techniczna – szybkość wyświetlania strony WWW w przeglądarce [70], szybkość pobierania treści przez roboty internetowe, szyfrowanie SSL [71], dostosowanie do urządzeń mobilnych,
- użyteczność i zaufanie – dostosowanie strony zgodnie ze standardem WCAG [72], przejrzystość i oryginalność treści, liczba i sposób rozmieszczenia reklam,
- linkowanie – liczba i jakość hiperłączy prowadzących do strony WWW z innych stron [73].

Wymienione wyżej działania możemy podzielić na te dokonywane bezpośrednio na stronie WWW, jak i te, które przeprowadza się poza nią [74]. Optymalizacja na stronie WWW i poza nią powinna wynikać ze spójnej i starannie zaplanowanej strategii, ponieważ tylko w ten sposób można uzyskać wymierny efekt w postaci lepszej pozycji rankingowej w SERP.

1.3.1. Optymalizacja w obrębie strony WWW

W praktyce działania w obrębie strony WWW mają fundamentalne znaczenie dla procesu jej optymalizacji dla wyszukiwarek i powinny być wykonane w pierwszej kolejności. Najkorzystniejszym wariantem jest wprowadzenie odpowiednich modyfikacji w kodzie strony WWW już na początkowym etapie jej budowy. Kluczowe w procesie optymalizacji jest to, aby wszystkie elementy, takie jak struktura, treść i odpowiednie występowanie w niej słów kluczowych, od początku wpływały na jej efektywność. W ramach optymalizacji przebudowuje się także istniejące strony WWW, głównie w przypadkach, kiedy firma nie zamierza finansować budowy witryny od początku. Działania optymalizacyjne należy realizować cyklicznie, wraz ze zmieniającymi się wytycznymi algorytmów wyszukiwarek [67].

Do najważniejszych elementów optymalizacji w obrębie strony WWW, znacząco wpływających na pozycję w wynikach wyszukiwania, zalicza się:

- znacznik TITLE – tytuł strony WWW ma kluczowe znaczenie, ponieważ pomaga użytkownikom szybko ocenić zawartość wyniku i jego trafność [75],
- znacznik META DESCRIPTION – opis lub streszczenie strony WWW stanowią część wyniku wyszukiwania w wyszukiwarkach [76],

- nagłówki H1–H6 – ułatwiają odbiorcy zorientowanie się w tematyce i hierarchii prezentowanych treści [77],
- słowa kluczowe w tekście, ich dobór i gęstość występowania [78],
- atrybut ALT, opisujący elementy graficzne – tych informacji używają roboty indeksujące i oprogramowanie czytnika ekranu, aby pomóc niewidomym użytkownikom zrozumieć zawartość obrazów [79],
- szczegółowy opis plików wideo, ułatwiający ich wyszukanie [80],
- linkowanie wewnętrzne – ułatwia nawigację po stronie WWW [81],
- kod HTML dostosowany do urządzeń mobilnych [82], zgodny ze standardami W3C [83],
- przyjazne adresy URL – najczęściej adresy krótkie, proste i czytelne, które zawierają słowa kluczowe [84],
- szybkość wyświetlania strony WWW w przeglądarkach internetowych [85].

1.3.2. Optymalizacja poza stroną WWW

Pozyskiwanie hiperłączy (ang. link building) z innych stron WWW, forów, blogów, optymalizacja wizytówki Google czy rekomendacje z mediów społecznościowych to działania związane z optymalizacją realizowane poza stroną WWW [86]. Pozyskiwane hiperłącza są traktowane przez wyszukiwarki jako polecenia, dlatego mają znaczący wpływ na pozycje rankingowe w SERP [87].

W procesie pozyskiwania hiperłączy dobrze sprawdzają się firmowe blogi, media społecznościowe, zaprzyjaźnione portale, strony publikujące artykuły sponsorowane i fora tematyczne [73]. Istotne jest to, aby publikowane w takich miejscach treści były wartościowe i zawierały hiperłącza powiązane z optymalizowaną stroną WWW [88]. Proces ten jest czasochłonny, zwłaszcza wtedy, kiedy koszt pozyskania wartościowych hiperłączy jest minimalny. Zbyt nachalne umieszczanie hiperłączy w różnych miejscach może skutkować efektem odwrotnym do zamierzonego.

Wdrożenie wizytówki firmowej Google i wypełnienie jej informacjami pozwala osiągnąć lepszą widoczność, przede wszystkim na urządzeniach mobilnych. Wyszukiwarka, wykorzystując lokalizację geograficzną, dopasowuje najlepsze wyniki do potrzeb użytkownika. Ważnym czynnikiem wpływającym na pozycje rankingowe w SERP jest również nazwa i wiek domeny [89].

W związku z tym, że działania poza stroną WWW wymagają doświadczenia, częstym błędem jest pozyskiwanie hiperłączy z niepewnych źródeł [90], oznaczonych jako niebezpieczne dla użytkownika [91], lub wykorzystywanie narzędzi do sztucznego generowania treści [92], które wpisują się w działania sprzeczne z polityką wyszukiwarek [93], co w konsekwencji może doprowadzić do usunięcia strony WWW z wyników wyszukiwania [94].

1.4. Narzędzia i audyt SEO

Osoba odpowiedzialna za optymalizację strony WWW, wprowadzająca zmiany na stronie i na bieżąco monitorująca postępy swoich działań, jest nazywana specjalistą SEO. Specjaliści SEO w procesie diagnostyki SEO i optymalizacji dla wyszukiwarek wykorzystują wiele narzędzi. W praktyce nie ma idealnego zestawu takich narzędzi, który gwarantowałby sukces. To dlatego, że każdy specjalista SEO tworzy taki zestaw programów, jaki będzie spełniał jego oczekiwania w długofalowym procesie optymalizacji [95]. Dodatkowym czynnikiem wpływającym na dobór narzędzi są środki finansowe, które można przeznaczyć na działania związane z optymalizacją poza stroną WWW.

Narzędzia SEO umożliwiają analizę kluczowych obszarów, które są decydujące dla efektywności optymalizacji pod wyszukiwarki, umożliwiającą znalezienie błędów w kodzie źródłowym i brakujących elementów w zawartości strony WWW. Analiza, będąca audytem SEO, skupia się na zagadnieniach związanych zarówno z optymalizacją, jak i użytecznością strony WWW i jej bezpieczeństwem.

Najważniejsze obszary audytu SEO to:

- analiza komunikacji strony WWW z robotami internetowymi na poziomie plików robots.txt [96] i sitemap.xml [97],
- analiza tytułu strony i znaczników META odpowiedzialnych za prezentację strony WWW w wynikach wyszukiwania [98],
- analiza struktury strony WWW i nawigacji – hiperłączy wewnętrznych, zewnętrznych oraz konstrukcji adresów URL [81],
- analiza treści na stronie WWW, w tym jej rozmiaru oraz występowania słów kluczowych i duplikatów [99],
- analiza statusów odpowiedzi serwera WWW [100],
- analiza poprawności kodu HTML [101],
- analiza wydajności – szybkości ładowania strony WWW [102],

- analiza występowania danych strukturalnych [103],
- analiza działania serwisu WWW na urządzeniach mobilnych [104],
- weryfikacja logów serwera WWW [105],
- analiza profilu hiperłączy [106],
- analiza kluczowych konkurentów w wynikach organicznych [107].

Narzędzia umożliwiające przeprowadzenie audytu SEO możemy podzielić na te oferowane przez wyszukiwarki – Google Search Console[9] w przypadku Google i Bing Webmaster Tools w przypadku Bing, oraz te, które analizują profil hiperłączy, strukturę i wydajność strony WWW, z których najpopularniejsze to Ahrefs, Majestic i Semrush. Ponadto wykorzystuje się narzędzia do analizy pozycji rankingowej w SERP i złożoności strony WWW [108] oraz wtyczki dla systemu WordPress – Yoast SEO i All in One SEO.

1.4.1. Google Search Console

Do opracowania audytu SEO trzeba mieć, biorąc pod uwagę dominację wyszukiwarki Google, dostęp do bezpłatnego narzędzia Google Search Console, czyli podstawowego narzędzia do analizy technicznej stron WWW. Narzędzie to pozwala na szybsze zindeksowanie podstron, pomaga też zweryfikować ich stan techniczny i uzyskać informację o napotkanych problemach z dostępem do danego zasobu, a także zweryfikuje stan mapy strony WWW [109].

Statystyki, jakie udostępnia Google Search Console, obejmują:

- liczbę wyświetleń i kliknięć w hiperłącza z SERP,
- odsetek wyświetleń (wskaźnik CTR) powodujących kliknięcia w hiperłącza z SERP,
- średnią pozycję rankingową w SERP,
- listę najczęściej występujących zapytań do wyszukiwarki powodujących wyświetlenie hiperłączy w SERP,
- listę podstron najczęściej wyświetlanych w SERP,
- listę krajów, z których pochodzi największa liczba zapytań,
- listę urządzeń, których używano podczas wyszukiwania.

[9] Nazwa Google Search Console (wcześniej Google Webmaster Tools) obowiązuje od 20 maja 2015 r.

Google Search Console udostępnia informacje na temat błędów napotkanych przez roboty indeksujące i liczbę poprawnie zindeksowanych podstron wokół całej domeny. Umożliwia ręczne wskazanie adresu URL do mapy strony WWW zapisanej w pliku tekstowym w formacie XML, zawierającej zbiór hiperłączy do wszystkich podstron i artykułów na blogu w domenie głównej. Narzędzie dostarcza również wiedzy na temat podstawowych wskaźników internetowych, czyli informacji o skuteczności stron WWW, na podstawie danych zebranych podczas korzystania z nich – wskazuje podstrony o słabej jakości i wymagające poprawy oraz adresy URL dobrej jakości.

W momencie wystąpienia problemów z brakiem dostępu do zasobów, kiedy serwer WWW zwraca kod błędu 404, można tymczasowo lub całkowicie usunąć adres URL z wyników organicznych. Usunięcie adresu URL jest możliwe tylko w obrębie analizowanego serwisu WWW, a usunięcie zasobów z innych domen jest możliwe wyłącznie we wskazanych przez Google przypadkach [110].

1.4.2. Ahrefs

Ahrefs to popularne narzędzie stosowane na co dzień przez specjalistów SEO do wykonywania audytów SEO i optymalizacji pod wyszukiwarki, analizy konkurencji i monitorowania efektów. Narzędzie jest dostępne w płatnej subskrypcji w cenie od 99 do 999 dolarów miesięcznie[10], w zależności od pakietu.

Narzędzie udostępnia informacje na temat profilu hiperłączy prowadzących do strony WWW, podzielonych na grupy i kategorie, wraz z listą wszystkich wykrytych domen i adresów IP odsyłających do strony WWW, a także pełną listą tekstów zakotwiczenia występujących w hiperłączach tekstowych. Ahrefs dostarcza informacji na temat słów kluczowych, które generują ruch z wyników organicznych, wraz z analizą konkurencji. W przypadku kampanii łączy sponsorowanych na platformie Google Ads narzędzie Ahrefs wykryje je na podstawie słów kluczowych powiązanych ze stroną WWW i wyświetli dane o tych, które generują największą liczbę wyświetleń i kliknięć.

Ahrefs analizuje również elementy graficzne, skrypty JavaScript i arkusze stylów kaskadowych (CSS) pod kątem szybkości wczytywania w przeglądarkach internetowych. Informuje, które zasoby są udostępniane za pomocą nieszyfrowanego połączenia, w przypadku zaś obrazów – które z nich nie są opisane atrybutem ALT w kodzie strony WWW.

[10] Na podstawie serwisu ahrefs.com, luty 2022 r.

Twórcy narzędzia Ahrefs opracowali system do ustalania najpopularniejszych stron WWW w sieci internet. Ustalając pozycję rankingową, Ahrefs uwzględnia:

- liczbę odsyłających hiperłączy i domen,
- tekst zakotwiczenia w hiperłączach,
- słowa kluczowe w wynikach organicznych i płatnych wynikach wyszukiwania,
- wskaźniki Ahrefs domain rating, Ahrefs URL rating, Ahrefs rank,
- ruch z wyników organicznych.

Narzędzie Ahrefs oferuje również wtyczkę do popularnego systemu zarządzania treścią WordPress, umożliwiającą analizę treści oraz monitorowanie hiperłączy zwrotnych, rekomendacji dla wpisów i podstron, a także sprawdza ogólny stan techniczny strony WWW.

1.4.3. Majestic

Majestic to popularne narzędzie do analizy profilu hiperłączy zwrotnych dowolnie wybranej strony WWW. Jest dostępne w płatnej subskrypcji w cenie od 49 do 399 dolarów miesięcznie[11], w zależności od pakietu.

Narzędzie analizuje zakotwiczenia (ang. anchors) co do łącznej liczby odwołujących się za pomocą nich domen i hiperłączy, a także informuje o liczbie usuniętych hiperłączy zwrotnych. Majestic udostępnia wskaźniki wiarygodności – TF (Trust Flow), i jakości profilu linków – CF (Citation Flow), które przyjmują wartości od 0 do 100 [111] i są istotne dla specjalistów SEO w trakcie optymalizacji poza stroną WWW.

TF, wskaźnik wiarygodności strony WWW, jest ustalany na podstawie wartości TF stron WWW, z których pochodzą hiperłącza zwrotne. Wskaźnik ten ułatwia wybór strony będącej optymalnym źródłem hiperłączy. CF to zaś wskaźnik szacujący jakość profilu hiperłączy na podstawie liczby hiperłączy zwrotnych. W przeciwieństwie do TF opiera się na ich liczbie, a nie na jakości.

Wartości wskaźników TF i CF powinny być zbliżone, co oznacza, że strona WWW ma zdywersyfikowane źródła hiperłączy przychodzących. Sytuacja, w której wartość wskaźnika TF jest znacznie wyższa niż CF, co w praktyce oznacza dużą liczbę hiperłączy zwrotnych o wysokim rankingu, może się spotkać z negatywną reakcją algorytmów wyszukiwarek, jeśli uznają taki profil hiperłączy za nienaturalny. Natomiast zbyt wysoka

[11] Na podstawie serwisu majestic.com, luty 2022 r.

wartość wskaźnika CF względem TF może oznaczać wykorzystanie praktyk sztucznego pozyskiwania hiperłączy o niskiej jakości.

Istotne znaczenie dla algorytmów wyszukiwarek mają hiperłącza pochodzące z górnej części strony WWW, z menu głównego i z treści, natomiast mniejsze mają hiperłącza z dolnej części, dlatego ważne jest przemyślane pozyskiwanie hiperłączy zwrotnych. Majestic dostarcza szczegółowych informacji na temat kontekstu hiperłączy, czyli otoczenia, w którym się znajdują. Hiperłącza mogą pochodzić z menu, zlokalizowanego w górnej części strony WWW lub jej stopce, z treści artykułu lub podstrony, a także z grafiki umieszczonej w treści lub z banneru reklamowego.

Majestic udostępnia dodatkowe funkcje, takie jak generator słów kluczowych, narzędzie do śledzenia kampanii reklamowej oraz możliwość pobierania listy hiperłączy zwrotnych i generowania raportu standardowego i zaawansowanego z możliwością zapisu w pliku w formacie PDF. Ponadto wszystkie informacje analityczne są udostępniane w przeglądarce internetowej za pośrednictwem wtyczki Backlink Analyzer.

1.4.4. Semrush

Semrush to narzędzie do przeprowadzania audytów SEO, pozwalające na analizę efektów optymalizacji serwisów WWW pod wyszukiwarki. Jego możliwości rozszerzono o funkcje dla specjalistów od marketingu, którzy zajmują się szeroko pojętym marketingiem w wyszukiwarkach [112]. Narzędzie jest dostępne w płatnej subskrypcji w cenie od 119,95 do 449,95 dolara miesięcznie[12], w zależności od pakietu.

Semrush udostępnia następujące funkcje:

- analiza ruchu organicznego i hiperłączy zwrotnych,
- dostęp do informacji o najlepszym słowie kluczowym dla działań SEO,
- porównanie serwisu WWW z maksymalnie pięcioma konkurencyjnymi w celu znalezienia brakującego słowa kluczowego w treści, a także porównania profili hiperłączy zwrotnych, aby znaleźć źródła, w których należy je utworzyć,
- analiza treści stron WWW konkurencji, w tym analiza występowania nagłówków i łącznej długości słów kluczowych z długim ogonem, czyli składających się zwykle z więcej niż trzech słów,
- analiza wyników organicznych dla danego słowa kluczowego,

[12] Na podstawie serwisu semrush.com, luty 2022 r.

- edytor treści ułatwiający pracę specjalistom SEO, który pozwala na pisanie tekstów czytelnych zarówno dla ludzi, jak i dla robotów indeksujących.

Narzędzie Semrush umożliwia określenie liczby wyszukiwań każdego słowa kluczowego, kosztów kliknięcia i trudności słów kluczowych oraz pozwala sprawdzić, które strony zajmują najlepsze pozycje w rankingach. Ponadto ułatwia identyfikację tych słów kluczowych, które znajdują się w treści strony WWW i są wykorzystywane przez konkurencyjne serwisy. Po połączeniu z usługami Google Search Console i Google Analytics wszystkie dane są dostępne w jednym, przejrzystym panelu.

1.5. Podsumowanie

W tym rozdziale dokonano przeglądu wyszukiwarek stron WWW, które począwszy od udostępnionej w 1990 roku Archie zyskały popularność na całym świecie. Zaprezentowano rodzaje wyników w wyszukiwarkach z podziałem na płatne i organiczne, a także przedstawiono metody optymalizacji pod wyszukiwarki, zarówno te, które stosuje się bezpośrednio na stronie WWW, jak i te, które wykorzystuje się poza stroną. W następnym rozdziale omówiono najważniejsze algorytmy rankingowe, które są fundamentem działania dzisiejszych wyszukiwarek.

2. Badania związane z analizą wyników w wyszukiwarkach

W tym rozdziale zebrano podstawową wiedzę na temat algorytmów rankingowych wykorzystywanych przez wyszukiwarki stron WWW, a szczegółowo skupiono się na omówieniu algorytmów PageRank oraz HITS. Przedstawiono w nim najważniejsze informacje na temat metod nadawania rankingu stronom WWW oraz wyniki badań naukowych związanych z próbami odkrycia najważniejszych czynników wpływających na ranking w wyszukiwarkach. Omówiono też algorytmy rankingowe wprowadzone przez twórców narzędzi do analityki marketingowej – MOZ Rank i Ahrefs Rank.

2.1. Fundamenty dzisiejszych algorytmów rankingowych

Wyszukiwarki internetowe są dziś dla większości z nas podstawowym źródłem wiedzy na każdy temat. Światowe koncerny technologiczne, takie jak Google, Microsoft, Meta czy Yahoo!, w specjalnie zaprojektowanych wyszukiwarkach udostępniają gromadzone latami informacje. To od twórców wyszukiwarek zależy, jak skonstruowane są strony wyników wyszukiwania, które pozycje na liście wyników będą wyróżnione, a które zasoby będą pomijane w wynikach. Za każdą wyszukiwarką stoi jej najbardziej strzeżona tajemnica w postaci algorytmu i metody nadawania rankingu konkretnemu zasobowi, których celem jest to, aby wynik wyszukania był dla odbiorcy jak najbardziej użyteczny [113].

Już w przypadku pierwszych wyszukiwarek, w latach 90. XX wieku, pojawił się problem z ustaleniem optymalnej listy wyników [114]. Pierwsze metody opierały się na analizie słów kluczowych i znaczników HTML, jednak nie dawały zadowalających rezultatów, głównie ze względu na niską wiarygodność i częste nadużycia ze strony projektantów stron WWW. Problemy z ustaleniem, które zasoby w sieci można uznać za istotne, doprowadziły do opracowania specjalnych algorytmów: HITS, stworzonego w 1998 roku przez Jona Kleinberga [115], oraz PageRank, autorstwa Sergeya Brina i Larry'ego Page'a [116], które stały się fundamentem większości dzisiejszych algorytmów rankingowych.

Kleinberg przyjął, że zasoby są połączone ze sobą, tworząc graf skierowany, w którym wierzchołkami są strony WWW, a krawędziami hiperłącza [117]. Struktura grafu jest zorganizowana w taki sposób, że krawędź jest skierowana ze strony linkującej na linkowaną. Kleinberg oparł swój algorytm na dwóch założeniach, określając je idealnymi, a mianowicie że istnieją strony WWW odgrywające rolę autorytetu (ang.

authority), czyli takie, na które wskazuje wiele hiperłączy, oraz strony pełniące funkcję koncentratora (ang. hub), czyli takie, które wskazują na strony autorytatywne [118].

Podstawą algorytmu było założenie, że idealna strona WWW powinna zawierać hiperłącza do innych wartościowych stron, a także być linkowana przez inne, równie ważne strony. Kleinberg opracował model oparty na hiperłączach do nadawania autorytetu i metodę, która identyfikowała zarówno relewantne, jak i autorytatywne strony dla zapytań o szerokiej tematyce [119].

Sergey Brin i Larry Page podobnie jak Kleinberg założyli, że sieć połączonych ze sobą hiperłączami stron WWW przypomina graf [120]. Przyjęli również założenie, że o wadze publikacji świadczy liczba odwołań z innych publikacji, czyli waga strony WWW może być mierzona liczbą hiperłączy wskazujących tę stronę z innych stron [121].

Algorytm PageRank, w odróżnieniu od zwykłego zliczania hiperłączy, stosuje ważenie ich wartości, co powoduje, że strona WWW może uzyskać wysoką pozycję rankingową, jeśli jest linkowana ze stron o wysokim rankingu. Takie podejście, swego czasu nowatorskie, powodowało, że jedno hiperłącze przychodzące ze strony o wysokim rankingu miało większą wartość niż wiele hiperłączy ze stron o niskim rankingu. W związku z tym, że algorytm opierał się tylko na informacji o hiperłączach, wystąpił problem zapętlenia, kiedy dwie strony wzajemnie do siebie odsyłają, ale poza tym – nigdzie indziej. Do obejścia tej pułapki zastosowano model losowego surfera (ang. random surfer model), którego działanie polega na tym, że w przypadku wykrycia problemu zapętlenia algorytm przechodzi do losowego hiperłącza, tak jakby wykonał to prawdziwy użytkownik [122].

Algorytm PageRank od samego początku istnienia wyszukiwarki Google miał istotny wpływ na pozycję w wynikach wyszukiwania. Jawność formuły algorytmu doprowadziła do sytuacji, w której wartość PageRank można było osiągnąć sztucznie [123], z wykorzystaniem specjalnie zaprojektowanych systemów do wzajemnego linkowania (ang. link farm) [124]. Przez takie praktyki, które w 1996 roku Eric Convey jako pierwszy nazwał spamem internetowym (ang. web spam) [125], twórcy Google musieli często zmieniać algorytm rankingowy [126].

Przełom w świadomości twórców stron WWW co do obecności w wyszukiwarkach nastąpił po nagłośnieniu słów amerykańskiego poety Kennetha Goldsmitha, który w 2005 roku na konferencji Elective Affinities Conference na Uniwersytecie Pensylwanii stwierdził, że jeśli czegoś nie ma w internecie, to nie istnieje[13] [127]. Choć Goldsmith miał na

[13] W oryginale: „If it doesn't exist on the internet, it doesn't exist".

myśli ideę powszechnego i bezpłatnego dostępu do zasobów naukowych, jego słowa stały się katalizatorem zmian w postrzeganiu znaczenia wyszukiwarek.

2.2. Próby odkrycia czynników rankingowych

Od momentu, kiedy Google stało się dominującą wyszukiwarką, a algorytm PageRank istotnym elementem w wyznaczaniu pozycji rankingowych, próbowano odkryć pozostałe czynniki, które wpływają na ranking [128]. Choć dziś wiemy, że dotąd nie udało się badaczom odkryć wszystkich reguł, na podstawie których Google ustala ranking, warto wspomnieć o próbach osiągnięcia tego celu.

W 2003 roku Ali Khaki-Sedigh i Mehdi Roudaki opublikowali pracę [129], w której opisali wykorzystanie metody najmniejszych kwadratów do modelowania dynamiki zmian zachodzących w wynikach wyszukiwarki Google. Wykonując cykliczne zapytania do wyszukiwarki, obserwowano pierwsze 100 pozycji w rankingu. Analizując wyniki, przygotowano zbiór uczący. Ustalono, że najniżej analizowaną pozycją rankingową będzie 87., aby uniknąć utraty ciągłości gromadzonych danych w wyniku częstych zmian w rankingu.

Z powodu braku wystarczających informacji na temat czynników rankingowych w wyszukiwarce Google parametry do obliczeń wybrano na podstawie wiedzy opartej na badaniach empirycznych. Jako parametry wejściowe wybrano wartość PageRank i skupiono się na analizie występowania słów kluczowych w kodzie strony WWW oraz w atrybutach ALT opisujących pliki graficzne.

Weryfikacja metody polegała na wyliczeniu pozycji rankingowych dla zbioru testowego, w którego skład wchodziło pięć stron WWW. Różnice między rankingiem wyszukiwarki Google a pozycjami wyliczonymi przez model wynosiły od 5 do 35 pozycji, co badacze uznali za dopuszczalny błąd. Dokładne wyliczenie pozycji rankingowej tą metodą nie było możliwe, głównie ze względu na brak dostępu do informacji na temat konstrukcji algorytmu rankingowego, choć jak wskazują Sedigh i Roudaki, można ją wykorzystać do modelowania dynamiki rankingu.

W 2005 roku Albert Bifet, Carlos Castillo, Paul-Alexandru Chirita oraz Ingmar Weber opublikowali pracę [130], w której zaproponowali, by do wyliczenia rankingu w wyszukiwarkach wykorzystać regresję logistyczną, maszynę wektorów nośnych i binarne drzewo klasyfikacyjne. Przyjęto założenie, że wyszukiwarki nie stosują tych samych

kryteriów oceny dla wszystkich stron WWW, natomiast dzielą je na grupy tematyczne i w ramach tych grup następuje wyliczenie rankingu.

Opracowano zestawy homogenicznych zapytań jedno-, dwu- i wielowyrazowych, a następnie przypisano je do grup niepowiązanych tematycznie. Zapytania rozdzielono w trzech zbiorach – treningowym, weryfikacyjnym i testowym.

Proces zbierania danych polegał na analizie wyników wyszukiwania w wyszukiwarce Google. Weryfikowano również obecność strony WWW w katalogu DMOZ oraz analizowano tekst zakotwiczenia stron linkujących. Listę hiperłączy do strony WWW uzyskiwano zapytaniem „link:" w wyszukiwarce.

Parametry wejściowe do obliczeń wytypowano na podstawie analizy tekstu na stronie WWW i występowania w nim konkretnych słów kluczowych. Zbadano miejsce występowania słów kluczowych, liczbę niepowtarzalnych zwrotów i częstotliwość ich występowania, porównano wielkość kodu strony WWW do wielkości tekstu, a także sprawdzono występowanie zwrotów podobnych. Przeanalizowano również wykorzystanie formatowania, obecność słów kluczowych w atrybutach ALT i TITLE oraz adresie URL, a także wartość PageRank i liczbę hiperłączy przychodzących z innych stron WWW i wychodzących do innych stron.

W badaniu uzyskano dokładność obliczenia rankingu w przedziale od 57% do 70%, w zależności od grupy tematycznej, co można uznać za wynik zadowalający. Wskazano, podobnie jak w przypadku badań Sedigha i Roudakiego, że wynikły błąd może być związany z brakiem dostępu do potwierdzonych informacji na temat czynników rankingowych wyszukiwarki Google, jak i z tym, że informacje o wartości PageRank i liczbie hiperłączy uzyskanej zapytaniem „link:" mogą być niedokładne.

Należy zwrócić uwagę na to, że wspomniani wyżej badacze dokonali szczegółowej analizy tekstu i kodu strony WWW, wykorzystali wartość PageRank[14], a pomimo to nie udało im się odkryć czynników, które znacząco decydują o pozycjach rankingowych. Zadanie to jest coraz trudniejsze, głównie za sprawą twórców wyszukiwarek, którzy regularnie aktualizują swoje algorytmy rankingowe [131]. Utrudnieniem są również aktualizacje algorytmów wyszukiwarek, choćby – w przypadku wyszukiwarki Google – aktualizacja o nazwie Panda w lutym 2011 roku [132], Penguin w kwietniu 2012 roku [133] czy Hummingbird w sierpniu 2013 roku [134].

[14] Informacje o wartości PageRank od 2016 r. nie są dostępne publicznie.

2.3. Metody nadawania rankingu

Dostarczanie odpowiednich wyników w SERP jest prawdopodobnie najważniejszym czynnikiem, który sprawia, że wyszukiwarka internetowa jest użyteczna. Ciągłe dążenie do poprawy jakości wyników inspiruje naukowców do projektowania wydajnych algorytmów rankingowych, umożliwiających wyszukiwarkom umieszczanie najtrafniejszych stron WWW na szczycie listy wyników dla określonego zapytania [135]. Najpopularniejsze algorytmy, ustalając pozycję rankingową, zazwyczaj analizują treść na stronie WWW i strukturę hiperłączy [136].

Struktura powiązań pomiędzy stronami WWW to prawdopodobnie najczęściej używana funkcja w rankingach opartych na popularności. Klasycznymi reprezentantami takich algorytmów są PageRank i HITS, a także oparte na nich Weighted PageRank [137] i SALSA [138]. Algorytmy te nie opierają się tylko na liczbie hiperłączy, lecz wprowadzają do obliczenia rankingu pojęcie ich jakości. Zastosowanie tej metody powoduje wystąpienie problemu dominacji popularnych stron WWW nad nowo powstałymi o potencjalnie wysokiej jakości [139]. Algorytmy tego typu są podatne na sztuczne podnoszenie pozycji rankingowej za pomocą zaprojektowanych specjalnie w tym celu systemów wymiany hiperłączy [140].

PageRank to najpopularniejszy algorytm rankingowy, w którego otoczeniu cały czas pojawiają się konkurenci, będący jego uzupełnieniem [141] lub oferujący odmienne pomysły [142] i założenia [143]. Należą do nich są algorytmy WLRank [144], T-Fresh [145] i Level-Based Link Analysis [146], wyznaczające pozycję rankingową na podstawie długości tekstu zakotwiczenia hiperłączy, czy algorytmy Recency-sensitive Query-based [147], Actual PageRank [148] i Time-weighted PageRank [149], mierzące częstotliwość zmian hiperłączy na stronie WWW.

Ponadto algorytm Frank [150] oblicza ranking na podstawie liczby odwiedzin, T--Rank [151] – częstotliwości aktualizacji treści, Wavelet Rank [152] – struktury hiperłączy, DistanceRank [153] – różnicy w liczbie hiperłączy pomiędzy dwoma stronami WWW, a WordRank [154], FocusedRank [155] i A3Crank [156] – podobieństwa stron WWW i korelacji pomiędzy hiperłączami. Algorytmy te, choć zazwyczaj powstają w ramach badań naukowych [157], często są stosowane w projektach komercyjnych [158].

Algorytmy rankingowe są również wykorzystywane w narzędziach informatycznych, które do oceny jakości stron WWW stosują indywidualne miary. Za pomocą tych miar projektanci witryn oceniają prawdopodobieństwo uzyskania przez daną stronę wysokiej

pozycji rankingowej w wyszukiwarkach. Najpopularniejsze narzędzia i miary stosowane przez specjalistów SEO to MOZ Rank, Ahrefs Rank, Majectic TrustFlow i Citation Flow.

Na potrzeby prezentacji algorytmów PageRank i HITS w kolejnych rozdziałach niniejszej pracy opracowano narzędzie programistyczne w języku Python umożliwiające obliczenie wartości rankingowych dla dowolnej struktury linkujących się wzajemnie stron WWW. Narzędzie jest dostępne bezpłatnie na portalu GitHub[15].

2.3.1. Google PageRank

PageRank to algorytm obliczający ranking strony WWW, oparty na założeniu, że o wadze publikacji świadczy liczba odwołań z innych publikacji, czyli waga strony WWW może być mierzona liczbą hiperłączy wskazujących tę stronę z innych stron. Algorytm PageRank stosuje ważenie wartości hiperłączy, co powoduje, że strona WWW może uzyskać wysoką wartość rankingową, jeśli linkowana jest ze stron o wysokim rankingu.

Algorytm PageRank jest przedstawiony równaniem (1):

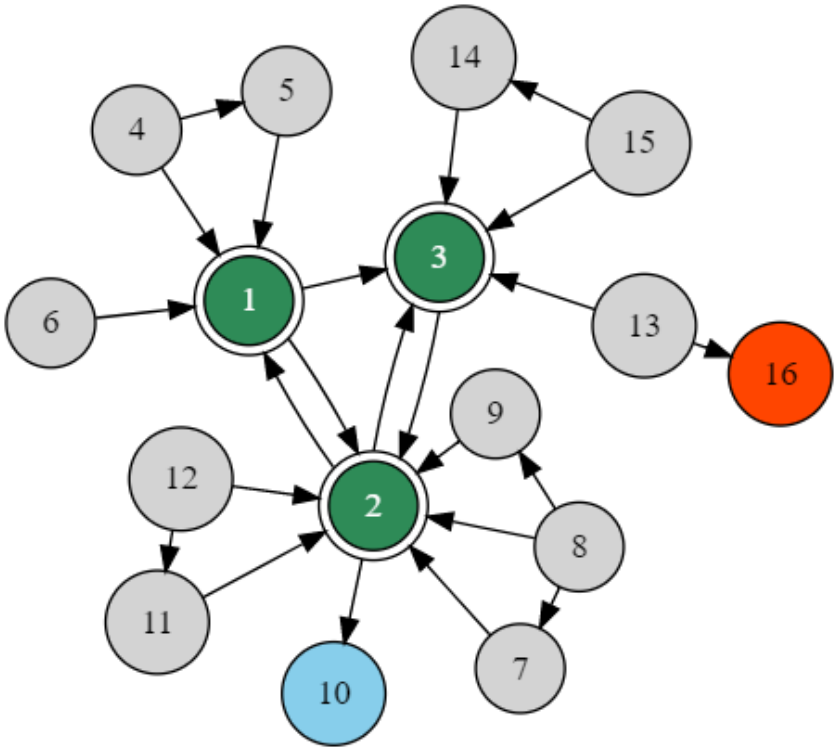
$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

gdzie *PR(A)* to wartość PageRank strony *A*, $PR(T_1)$ to wartość PageRank strony T_1, $C(T_1)$ to liczba łączy wskazujących stronę T_1, N to liczba stron internetowych, d zaś to współczynnik tłumienia, mieszczący się w zakresie $0 < d < 1$, dla którego zwykle przyjmuje się wartość 0,85 [159].

PageRank jest obliczany iteracyjnie jako suma wartości PageRank wszystkich stron WWW wskazujących na wybraną stronę podzielona przez liczbę łączy na każdej z tych stron [160]. Wskaźnik PageRank jest mierzony w skali logarytmicznej od 1 do 10, co oznacza, że jego zwiększenie z poziomu 0 na 1 jest zdecydowanie łatwiejsze niż z poziomu 4 na 5.

Przykładowa struktura stron WWW jest przedstawiona na rysunku 4. Strony o numerach 1, 2 i 3 mają znacznie więcej hiperłączy przychodzących od pozostałych stron, zatem zgodnie z zasadą algorytmu uzyskają wyższe wartości PageRank. Strony o numerach 10 i 16 pełnią funkcję kontrolną, a ich wartości są ustalane na podstawie tylko jednego hiperłącza przychodzącego.

[15] PageRank-HITS, https://github.com/mariuszduka/PageRank-HITS, kwiecień 2022 r.



Rysunek 4. Przykładowa struktura połączeń pomiędzy stronami WWW, opracowanie własne

Tabela 1. przedstawia wyliczone przez algorytm PageRank wartości dla każdej strony WWW w pierwszych 10 iteracjach. Współczynnik tłumienia *d* ustalono na 0,15, co oznacza, że 85% wartości PageRank jest przekazywane stronom linkowanym. Stabilizacja wyników następuje w piątym cyklu, w którym strony o numerach 1, 2 i 3 uzyskały najwyższe wartości PageRank, odpowiednio: 0,164, 0,352 i 0,21. Wyniki dla stron kontrolnych, numer 10 i 16, które wynoszą, odpowiednio, 0,111 i 0,016, jednoznacznie wskazują, że jakość hiperłączy zwrotnych jest kluczowa dla uzyskania dobrego wyniku.

Tabela 1. Wartości PageRank w kolejnych iteracjach algorytmu, opracowanie własne

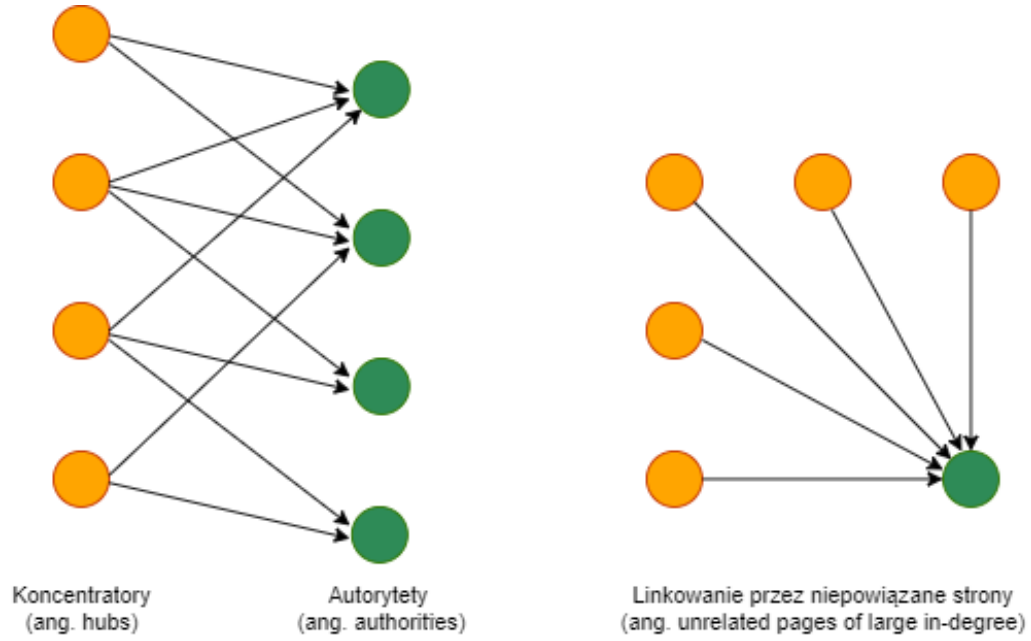
Iter.	PR(1)	PR(2)	PR(3)	PR(5)	PR(7)	PR(9)	PR(10)	PR(11)	PR(14)	PR(16)	PR(4,6,8,12,13,15)
0	1	1	1	1	1	1	1	1	1	1	1
1	0,167	0,355	0,29	0,001	0,02	0,001	0,101	0,03	0,03	0,001	0,001
2	0,128	0,398	0,208	0,015	0,011	0,014	0,124	0,011	0,011	0,015	0,011
3	0,178	0,344	0,207	0,016	0,015	0,014	0,109	0,017	0,017	0,016	0,011
4	0,164	0,351	0,209	0,016	0,015	0,015	0,111	0,017	0,017	0,016	0,011
5	**0,164**	**0,352**	**0,21**	**0,016**	**0,015**	**0,014**	**0,111**	**0,017**	**0,017**	**0,016**	**0,011**
6	0,164	0,352	0,21	0,016	0,015	0,014	0,111	0,017	0,017	0,016	0,011
7	0,164	0,352	0,21	0,016	0,015	0,014	0,111	0,017	0,017	0,016	0,011
8	0,164	0,352	0,21	0,016	0,015	0,014	0,111	0,017	0,017	0,016	0,011
9	0,164	0,352	0,21	0,016	0,015	0,014	0,111	0,017	0,017	0,016	0,011
10	0,164	0,352	0,21	0,016	0,015	0,014	0,111	0,017	0,017	0,016	0,011

Początkowo PageRank był głównym czynnikiem wpływającym na ranking w wyszukiwarce Google. Specjaliści SEO, znając zasadę działania algorytmu, starali się sztucznie zwiększać jego wartość [161], co wymusiło na wyszukiwarkach zmodyfikowanie i rozszerzenie czynników rankingowych. Algorytm PageRank nadal jest wykorzystywany przez wyszukiwarkę Google, jednak w nowszej i zaktualizowanej formie.

2.3.2. HITS

Algorytm HITS (ang. Hyperlink-Induced Topic Search) jest przeznaczony do automatycznej identyfikacji wartościowych publikacji na dany temat na podstawie konkretnego zapytania w wyszukiwarce. Twórca algorytmu Jon Kleinberg opracował koncepcje autorytetu (ang. authority) i koncentratora (ang. hub), pojęć wzajemnie określonych rekurencyjnie. Dobry autorytet to publikacja, którą cytuje wiele dobrych koncentratorów, a dobry koncentrator to publikacja, która zawiera hiperłącza do wielu dobrych autorytetów [162].

W efekcie działania algorytmu każdej publikacji przyporządkowane są dwie wagi – *a* i *h*, których wartości zawierają się w przedziale od 0 do 1. Wagi te określają, jak dobrym autorytetem i koncentratorem jest konkretna publikacja. Zadaniem koncentratorów było odróżnienie stron autorytatywnych w danym zagadnieniu, które często są cytowane przez strony o zbliżonej tematyce, od stron popularnych, które z kolei są cytowane przez niezwiązane ze sobą publikacje [163]. Koncepcja koncentratorów i autorytetów w algorytmie HITS jest przedstawiona na rysunku 5.

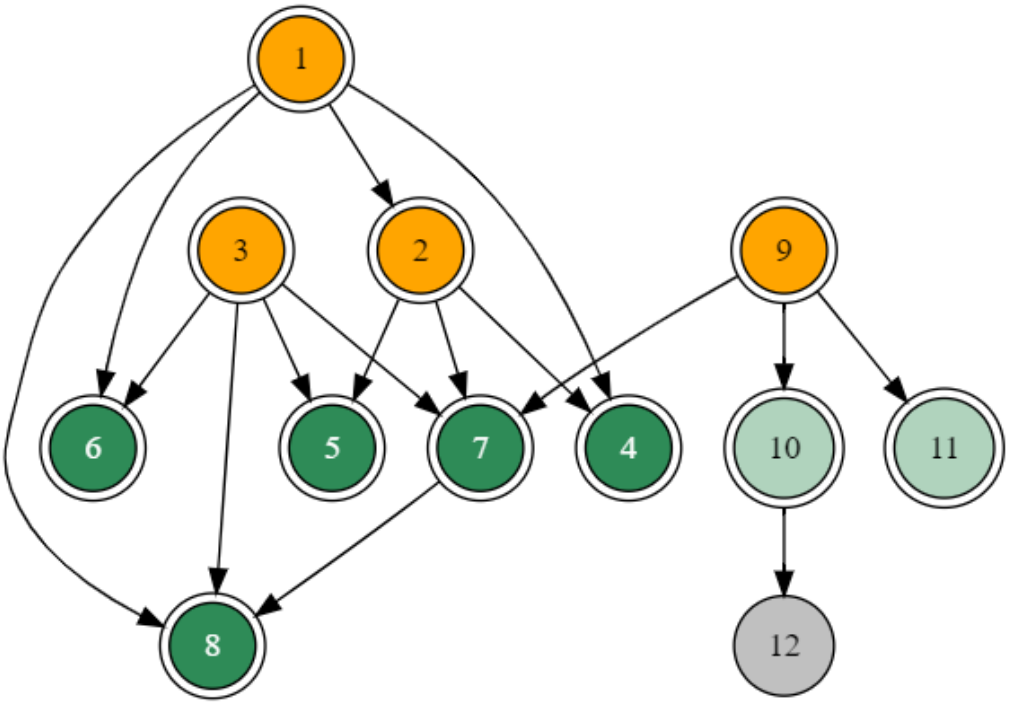


Rysunek 5. Koncepcja koncentratorów i autorytetów w algorytmie HITS, opracowanie własne

Przyporządkowanie stronom WWW odpowiednich wag, osobno dla autorytetu i koncentratora, wymaga przygotowania dwóch zbiorów, bazowego i pierwotnego. Zgodnie z założeniem zbiór bazowy (ang. base set) powinien zawierać wartościowe dokumenty związane z konkretnym zapytaniem w wyszukiwarce oraz jak najwięcej dokumentów sklasyfikowanych jako autorytety i jednocześnie być zbiorem stosunkowo małym [164]. Lista wartościowych dokumentów tworzy zbiór pierwotny (ang. root set), będący podzbiorem zbioru bazowego. Zbiór pierwotny jest budowany na podstawie stron WWW z najwyższym rankingiem, pozyskanych z wyszukiwarki dla konkretnego zapytania.

Kolejnym krokiem jest dołączenie do zbioru pierwotnego wszystkich dokumentów linkujących do zbioru pierwotnego i linkowanych przez zbiór pierwotny, aby nie pominąć dobrych autorytetów i koncentratorów. Z testów przeprowadzonych przez Kleinberga wynika, że optymalny zbiór bazowy powinien zawierać do 200 stron WWW z wyników wyszukiwania, a ponadto powinien objąć do 50 stron WWW linkujących każdą stronę będącą wynikiem wyszukiwania [165]. Tak przygotowany zbiór wynikowy spełnia warunki zbioru bazowego, co umożliwia obliczenie wag autorytetu i koncentratora dla każdej strony WWW.

Przykładowa struktura (zbiór bazowy) linkujących się wzajemnie stron WWW dla konkretnego zapytania w wyszukiwarce, odzwierciedlająca koncepcję koncentratorów i autorytetów, jest przedstawiona na rysunku 6.



Rysunek 6. Przykładowa struktura połączeń pomiędzy stronami WWW, opracowanie własne

Wagi autorytetu i koncentratora są obliczane iteracyjnie do momentu osiągnięcia poziomu równowagi z ustaloną dokładnością. Przed rozpoczęciem iteracji wartość każdej z wag jest inicjowana liczbą 1 [166].

Jeśli wartość *n* to liczba wszystkich dokumentów w zbiorze bazowym, a wartości wag odpowiadające konceptowi autorytetu – *a*, i pojęciu koncentratora – *h*, są wyliczane na podstawie dokumentu – *p*, relacje pomiędzy autorytetem a koncentratorem są wyrażone równaniami (2) i (3):

$$\forall p, a(p) = \sum_{i=1}^{n} h(i) \tag{2}$$

$$\forall p, h(p) = \sum_{i=1}^{n} a(i) \tag{3}$$

Po każdej parze iteracji następuje porównanie wartości z iteracji poprzedniej zgodnie z równaniem (4) i w przypadku, kiedy wartości te są zbieżne, algorytm kończy swoje działanie [167].

$$\sum_{i=1}^{n} h(i)^2 = \sum_{i=1}^{n} a(i)^2 = 1 \tag{4}$$

Strony o numerach 1, 2, 3 i 9 pełnią funkcję koncentratorów, a strony o numerach 4, 5, 6, 7, 8, 10 i 11 odgrywają rolę autorytetów. Strona o numerze 12 uzyskała wartość 0,0 zarówno dla autorytetu, jak i koncentratora, co oznacza, że dla tej konkretnej struktury połączeń strona nie jest nośnikiem wartościowych informacji lub odbiega od tematu zapytania w wyszukiwarce.

Tabela 2. przedstawia wyliczone przez algorytm HITS wartości autorytetu dla każdej strony WWW w pierwszych 10 iteracjach. Stabilizacja wyników następuje w szóstym cyklu, w którym strony o numerach 4, 5, 6, 7, 8 uzyskały najwyższe wartości autorytetu, odpowiednio: 0,140, 0,158, 0,170, 0,192 i 0,195. Strona numer 8 uzyskała najwyższy wynik, co oznacza, że w tej konkretnej strukturze linków jest najbardziej autorytatywna i najprawdopodobniej najlepiej odpowiada na zapytanie skierowane do wyszukiwarki. Obecność strony numer 12 w tej strukturze można uznać za nieistotną, ponieważ uzyskała ona wartość zerową zarówno dla autorytetu, jak i dla koncentratora.

Tabela 2. Wartości autorytetów w kolejnych iteracjach algorytmu HITS, opracowanie własne

Iter.	a(1)	a(2)	a(3)	a(4)	a(5)	a(6)	a(7)	a(8)	a(9)	a(10)	a(11)	a(12)
0	1	1	1	1	1	1	1	1	1	1	1	1
1	0,000	0,062	0,000	0,125	0,125	0,125	0,188	0,188	0,000	0,062	0,062	0,062
2	0,000	0,071	0,000	0,134	0,152	0,161	0,196	0,188	0,000	0,045	0,045	0,009
3	0,000	0,074	0,000	0,138	0,157	0,167	0,195	0,192	0,000	0,038	0,038	0,001
4	0,000	0,075	0,000	0,140	0,158	0,168	0,194	0,194	0,000	0,036	0,036	0,000
5	0,000	0,076	0,000	0,140	0,158	0,169	0,193	0,195	0,000	0,035	0,035	0,000
6	**0,000**	**0,076**	**0,000**	**0,140**	**0,158**	**0,170**	**0,192**	**0,195**	**0,000**	**0,034**	**0,034**	**0,000**
7	0,000	0,076	0,000	0,140	0,158	0,170	0,192	0,195	0,000	0,034	0,034	0,000
8	0,000	0,076	0,000	0,140	0,158	0,170	0,192	0,195	0,000	0,034	0,034	0,000
9	0,000	0,076	0,000	0,140	0,158	0,170	0,192	0,195	0,000	0,034	0,034	0,000
10	0,000	0,076	0,000	0,140	0,158	0,170	0,192	0,195	0,000	0,034	0,034	0,000

Tabela 3. przedstawia wyliczone przez algorytm HITS wartości koncentratora dla każdej strony WWW w pierwszych 10 iteracjach. Stabilizacja wyników następuje w szóstym cyklu, w którym strony o numerach 1, 2 i 3 uzyskały najwyższe wartości koncentratora, odpowiednio: 0,259, 0,219 i 0,319. Strona numer 2 jednocześnie odgrywa rolę autorytetu z wynikiem 0,076, co oznacza, że w strukturze linków mogą się znajdować strony spełniające kryteria zarówno dla autorytetu, jak i koncentratora.

Tabela 3. Wartości koncentratorów w kolejnych iteracjach algorytmu HITS, opracowanie własne

Iter.	h(1)	h(2)	h(3)	h(4)	h(5)	h(6)	h(7)	h(8)	h(9)	h(10)	h(11)	h(12)
0	1	1	1	1	1	1	1	1	1	1	1	1
1	0,235	0,206	0,294	0,000	0,000	0,000	0,088	0,000	0,147	0,029	0,000	0,000
2	0,250	0,218	0,315	0,000	0,000	0,000	0,085	0,000	0,129	0,004	0,000	0,000
3	0,255	0,219	0,318	0,000	0,000	0,000	0,086	0,000	0,121	0,001	0,000	0,000
4	0,257	0,219	0,319	0,000	0,000	0,000	0,086	0,000	0,118	0,000	0,000	0,000
5	0,258	0,219	0,319	0,000	0,000	0,000	0,087	0,000	0,117	0,000	0,000	0,000
6	**0,259**	**0,219**	**0,319**	**0,000**	**0,000**	**0,000**	**0,087**	**0,000**	**0,116**	**0,000**	**0,000**	**0,000**
7	0,259	0,219	0,319	0,000	0,000	0,000	0,087	0,000	0,116	0,000	0,000	0,000
8	0,259	0,219	0,319	0,000	0,000	0,000	0,087	0,000	0,116	0,000	0,000	0,000
9	0,259	0,219	0,319	0,000	0,000	0,000	0,087	0,000	0,116	0,000	0,000	0,000
10	0,259	0,219	0,319	0,000	0,000	0,000	0,087	0,000	0,116	0,000	0,000	0,000

Algorytm HITS do działania wymaga wewnętrznej wyszukiwarki, ponieważ wartości autorytetów i koncentratorów są obliczane indywidualnie dla każdego zapytania. Oznacza to, że ranking zawsze będzie tworzony na podstawie słów kluczowych lub treści,

które są w danym momencie wyszukiwane, a nie podczas indeksowania stron WWW, jak to ma miejsce w przypadku algorytmu PageRank. Wymagana jest również informacja na temat dokumentów wskazujących na zbiór pierwotny, którą trudno uzyskać, mając do dyspozycji tylko ten zbiór. Algorytm HITS równolegle z algorytmem PageRank zapoczątkował rozwój technik analizy powiązań pomiędzy stronami WWW, jednak nie jest powszechnie stosowany przez wyszukiwarki.

2.3.3. MOZ Rank

Założona w 2004 roku firma MOZ, oferująca pod adresem moz.com narzędzia do analityki marketingowej, opracowała własny algorytm nadający stronom internetowym wartość rankingową. Serwis MOZ udostępnia specjalistom SEO szczegółowe informacje o prawie 41 bilionach hiperłączy, w ramach płatnej subskrypcji w cenie od 250 do 10 tys. dolarów miesięcznie[16], w zależności od pakietu.

Założeniem rankingu MOZ jest przewidzenie pozycji danego serwisu internetowego w wyszukiwarkach. Każdej stronie przyznawany jest tzw. autorytet, w przedziale od 0 do 100 punktów. Im wyższy autorytet domeny lub strony, tym wyższe prawdopodobieństwo uzyskania dobrej pozycji w wyszukiwarkach. Ranking MOZ jest obliczany dla całego serwisu internetowego – DA (Domain Authority), oraz dla konkretnej strony – PA (Page Authority) [168].

Algorytm do obliczenia autorytetu oprócz analizy jakości hiperłączy przychodzących bierze pod uwagę wiele dodatkowych czynników. Autorytet jest zawarty w 100-punktowej skali logarytmicznej, co oznacza, że przeskok z 20 na 30 punktów jest znacznie łatwiejszy niż z 70 na 80 punktów. Szczegóły algorytmu są objęte tajemnicą handlową firmy MOZ.

2.3.4. Ahrefs Rank

Twórcy narzędzia Ahrefs opracowali własny algorytm rankingowy oparty na analizie liczby i jakości hiperłączy przychodzących. Algorytm Ahrefs Rank jest obliczany dla całego serwisu internetowego – DR (Domain Rating), oraz dla konkretnej strony – UR (URL Rating) [169].

[16] Na podstawie serwisu moz.com, marzec 2022 r.

DR, wskaźnik jakości domeny, jest obliczany na podstawie jakości hiperłączy zwrotnych do danej domeny, w porównaniu z innymi domenami znajdującymi się w systemie. Wskaźnik ten przyjmuje wartości od 0 do 100 w skali logarytmicznej i im jest on wyższy, tym lepsza jest jakość hiperłączy zwrotnych do danej domeny. DR pomaga specjaliście SEO dokonać optymalnego wyboru strony WWW, z której należy pozyskać hiperłącze zwrotne, podczas działań w ramach optymalizacji poza stroną WWW.

UR, wskaźnik kondycji linkowania adresu URL, informuje o jakości hiperłączy zwrotnych. Wskaźnik ten przyjmuje wartości od 0 do 100 w skali logarytmicznej i im jest on wyższy, tym lepsza jest kondycja linkowanego adresu URL. Do obliczenia UR wykorzystywane są podstawy algorytmu PageRank, co pozwala skutecznie oceniać jakość odnośników prowadzących do analizowanej strony. W odróżnieniu od DR, obliczanego dla całej domeny, UR może być obliczony dla domeny głównej, jak i poszczególnych podstron.

2.4. Podsumowanie

W tym rozdziale omówiono fundamenty dzisiejszych algorytmów rankingowych i zasadę działania algorytmów PageRank i HITS, wykorzystując opracowane specjalnie w tym celu narzędzie programistyczne. Przedstawione wyniki badań pozwalają jednoznacznie stwierdzić, że odkrycie wszystkich czynników wpływających na ranking, zwłaszcza w wyszukiwarce Google, jest praktycznie niemożliwe. W następnym rozdziale omówiono algorytm rankingowy ISOWQ Rank i system rankingowy ISOWQ.

3. Algorytm ISOWQ Rank i system rankingowy ISOWQ

W tym rozdziale zaprezentowano zasady działania algorytmu rankingowego ISOWQ Rank oraz omówiono jego elementy składowe i zastosowanie na przykładowych stronach WWW. Przedstawiono m.in. architekturę systemu rankingowego ISOWQ i sposób implementacji algorytmu ISOWQ Rank w celu nadawania rankingu stronom WWW. Ponadto zaprezentowano wyniki badań porównawczych algorytmu ISOWQ Rank i algorytmu rankingowego MOZ.

3.1. Zasada działania algorytmu ISOWQ Rank

3.1.1. Wstęp

Założenia algorytmów rankingowych, takich jak PageRank i HITS, z których wynika, że o wadze publikacji świadczy liczba odwołań z innych publikacji, są częścią współczesnych systemów rankingowych stosowanych przez najpopularniejsze wyszukiwarki internetowe. Obliczenie rankingu dla strony WWW z wykorzystaniem algorytmu PageRank lub HITS wymaga od wyszukiwarek zebrania jak najwięcej danych o powiązaniach pomiędzy stronami. Im informacja o strukturze hiperłączy jest pełniejsza, tym dokładniejsze są wyniki obliczeń rankingów dla stron WWW. Biorąc pod uwagę liczbę hiperłączy dostępnych w sieci internet, wyrażaną w bilionach, analiza połączeń pomiędzy nimi wymaga użycia zaawansowanych, złożonych z tysięcy serwerów systemów informatycznych, na których budowę mogą sobie pozwolić tylko właściciele największych wyszukiwarek, takich jak Google, Bing, Yahoo!, Yandex czy Baidu.

Uzyskanie wysokiej pozycji rankingowej strony WWW w wyszukiwarce wymaga, oprócz pozyskania jakościowych hiperłączy zewnętrznych, optymalizacji strony pod wyszukiwarki. Biorąc pod uwagę czas – który nie jest znany – między pobraniem strony WWW przez roboty internetowe a obliczeniem rankingu w wyszukiwarce, istotna jest aktualna wiedza na temat jakości optymalizacji w obrębie strony WWW i poza nią. Zwykle po wykonaniu działań związanych z optymalizacją SEO strona WWW jest zgłaszana do indeksacji w wyszukiwarkach za pomocą panelu administracyjnego udostępnianego dla twórców stron. W związku z tym, że informacje o wszystkich czynnikach rankingowych wyszukiwarek nie są znane, pojawiły się narzędzia do analityki marketingowej, takie jak Moz Analytics czy Ahrefs, które ułatwiają analizę struktury hiperłączy.

Udostępniają one własne miary rankingowe ułatwiające ocenę jakości optymalizacji strony WWW pod wyszukiwarki. Narzędzia te z reguły są płatne, a ich algorytmy rankingowe są objęte tajemnicą. Aby umożliwić twórcom stron WWW bezpłatną ocenę optymalizacji w obrębie strony WWW, opracowano algorytm ISOWQ Rank i system rankingowy ISOWQ.

3.1.2. Założenia algorytmu ISOWQ Rank

Algorytm ISOWQ Rank nadaje badanym stronom WWW określoną wartość, oznaczającą ich jakość. Obliczany jest według następującego równania (5):

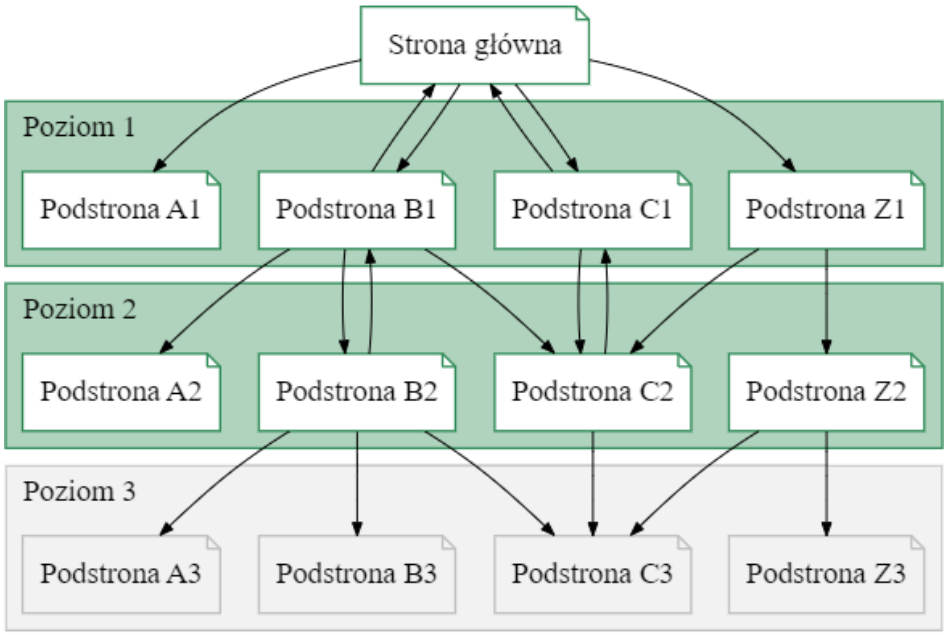
$$IR = \frac{PM+PK+PT}{3} \tag{5}$$

gdzie *IR* to wartość ISOWQ Rank dla strony WWW, *PM* to liczba punktów za wykorzystane technologie i pozycje rankingowe, *PK* to liczba punktów za optymalizację kodu źródłowego, a *PT* to liczba punktów za treść i strukturę tekstu. Wartość ISOWQ Rank jest liczona w skali od 0 do 20 punktów.

W sieci internet dostępne są różne typy stron WWW; do najpopularniejszych należą strony firmowe, blogi, sklepy internetowe i portale. Klienci, decydując się na zakup danego produktu czy usługi, często uzupełniają wiedzę na ich temat na firmowych stronach WWW. To od firmy zależy, czy jej strona WWW będzie zawierała informacje niezbędne do tego, aby pozytywnie wpłynąć na decyzje klientów. Podczas obliczania rankingu przyjęto założenie, że statystyczna strona WWW prezentująca podstawowe informacje biznesowe to struktura połączonych hiperłączy zawierająca stronę główną i co najmniej cztery podstrony. Taka struktura powinna zawierać niezbędne dane o firmie, informacje na temat bieżącej działalności, aktualną ofertę i formularz kontaktowy. Przykład takiej struktury tworzącej firmową stronę WWW jest przedstawiony jest rysunku 7.

Rysunek 7. Podstawowe informacje na firmowej stronie WWW, opracowanie własne

Ranking jest obliczany dla pierwszych 30 adresów URL odnalezionych w kodzie strony WWW, które zostaną poprawnie pobrane z serwera, czyli wtedy, kiedy serwer zwróci kod HTTP 2xx. Przyjęto założenie, że najważniejsze informacje na stronie WWW powinny znajdować się na pierwszych dwóch poziomach linkowania. Na rysunku 8. przedstawiono przykładową strukturę hiperłączy występującą na stronie WWW oraz wyróżniono pierwsze dwa poziomy linkowania, które są uwzględnione podczas wyliczania wartości rankingowej przez algorytm ISOWQ Rank.



Rysunek 8. Przykładowa struktura hiperłączy na stronie WWW, opracowanie własne

Przyjęto zasadę, że punktacja zostanie proporcjonalnie obniżona, jeśli w analizowanej stronie WWW jest mniej niż wymagane cztery podstrony, a także kiedy pojawią się odnośniki wewnętrzne do nieistniejących stron, czyli wtedy, gdy serwer zwróci kod błędu HTTP 4xx lub 5xx. Pseudokod 1. przedstawia funkcję obliczającą współczynnik korygujący (*LR*) ranking dla strony WWW w przypadku, kiedy nie występuje na niej zakładana minimalna liczba podstron, lub w przypadku, kiedy wykryte zostały hiperłącza do nieistniejących podstron.

```
# <1.00, 3.00>
Wejście:
  CRAWL - liczba stron pobranych poprawnie z serwera WWW
  MIN_PAGES – minimalne CRAWL dla statystycznej strony WWW, domyślnie 5
  ERRORS - liczba stron zgłaszających błąd serwera HTTP 4xx lub 5xx
Wyjście:
  LR - wartość w przedziale <1.00, 3.00>
```

```
Utwórz zmienną LR ← 1

If ERRORS > 0:
  LR ← LR + (ERRORS / (CRAWL + ERRORS))
If CRAWL < MIN_PAGES:
  LR ← LR + ((MIN_PAGES - CRAWL) * 0.25)

If LR > 3:
  LR ← 3

Zwróć LR
```

Pseudokod 1. Funkcja obliczająca współczynnik korygujący – *LR*

Podczas obliczeń trzech głównych parametrów algorytmu – *PM*, *PK* i *PT*, średnie wartości rankingowe dla wszystkich podstron są dzielone przez współczynnik *LR*, co w konsekwencji może obniżyć ranking dla strony WWW. Wartość rankingowa dla strony WWW jest obliczana na podstawie stron poprawnie pobranych z serwera WWW, dlatego wprowadzenie do obliczeń współczynnika *LR* ma na celu wyeliminowanie anomalii, kiedy ranking jest wyliczany dla strony WWW zawierającej jedynie stronę główną i wiele hiperłączy do nieistniejących podstron.

W tabeli 4. przedstawiono zmianę wartości punktacji w zależności od współczynnika *LR*. Punktacja w pierwszym wierszu tabeli nie została obniżona, ponieważ liczba pobranych stron (*C*) wyniosła co najmniej wymaganą wartość (*D*) i nie zostały zgłoszone błędy serwera (*E*). W piątym wierszu tabeli punktacja została obliczona dla strony WWW składającej się ze strony głównej i jednej podstrony (*C*), dodatkowo zawierającej cztery hiperłącza do nieistniejących podstron (*E*), dlatego współczynnik *LR* spowodował jej obniżenie z wartości 7,00 na 2,89.

Tabela 4. Wpływ współczynnika korygującego *LR* na wartość rankingową, opracowanie własne

Lp.	C (strony pobrane poprawnie)	D (wymagana liczba stron)	E (strony zgłaszające błąd serwera)	LR (współczynnik korygujący ranking)	Przykładowe korekty punktacji o początkowej wartości 7,00 w zależności od LR
1	**30**	**5**	**0**	**1,00**	**7,00 / 1,00 = 7,00**
2	25	5	5	1,16	7,00 / 1,16 = 6,03
3	20	5	8	1,29	7,00 / 1,29 = 5,43
4	4	5	0	1,25	7,00 / 1,25 = 5,60
5	2	5	4	2,42	7,00 / 2,42 = 2,89

Czynniki wpływające na obliczenie trzech głównych parametrów algorytmu – *PM*, *PK* i *PT*, przez lata się zmieniały, począwszy od 2011 roku, kiedy algorytm zastosowano po raz pierwszy w systemie rankingowym ISOWQ. Liczba czynników rankingowych

i ich wartości zmieniały się w czasie w zależności od ówcześnie dostępnej wiedzy na temat optymalizacji w obrębie strony WWW i poza nią, zmieniających się technologii związanych z projektowaniem stron, a także w wyniku badań empirycznych mających na celu odkrycie istotności poszczególnych parametrów, które mogą wpłynąć na ranking w wyszukiwarkach. Takie zmiany miały miejsce m.in. po ukryciu informacji na temat wartości PageRank przez wyszukiwarkę Google w 2016 roku, zamknięciu katalogu DMOZ w 2017 roku czy wprowadzeniu do algorytmu ISOWQ Rank informacji na temat wartości rankingowych MOZ DA i MOZ PA.

3.1.3. Punktacja za wykorzystane technologie i pozycje rankingowe

Obliczenie punktacji za użyte technologie i pozycje rankingowe – *PM*, przebiega w dwóch etapach. Na pierwszym etapie do obliczenia punktacji dla strony WWW jako całości brane są pod uwagę takie czynniki jak wartości rankingowe MOZ i Alexa Rank, liczba odnośników przychodzących, zastosowanie wtyczek społecznościowych i szyfrowania SSL, fizyczna lokalizacja serwera WWW, występowanie jawnych adresów e-mail na stronie i rejestracja adresu IP serwera hostującego w bazach DNSbl. Na tym etapie strona WWW może uzyskać od −3,00 do 35,00 punktów.

Na drugim etapie obliczana jest punktacja dla każdej podstrony, gdzie analizowane są takie czynniki jak wykorzystanie wtyczek społecznościowych, narzędzi Google, narzędzi do publikacji treści multimedialnych czy udostępnianych dokumentów. Na tym etapie każda podstrona może uzyskać od −0,25 do 12,00 punktów, po czym obliczana jest wartość średnia z punktacji wszystkich podstron. Punktacja na drugim etapie jest korygowana o współczynnik *LR*.

Punktacja końcowa jest sumą punktów uzyskanych na pierwszym i drugim etapie. Choć teoretycznie zakres punktacji za użyte technologie i pozycje rankingowe mieści się w zakresie od −3,25 do 47,00 punktów, przyjęto regułę, że w razie uzyskania wartości ujemnej punktacja jest ustalana na 0 punktów, natomiast wartość maksymalna to 20 punktów. Przykładowo, jeśli strona WWW uzyska −2,45 lub 27,85 punktu, to wynik końcowy zostanie ustalony, odpowiednio, na 0 lub 20 punktów.

3.1.3.1. Etap pierwszy – punktacja dla strony WWW jako całości

Punktacja za wartości rankingowe MOZ DA i MOZ PA

Parametry rankingowe MOZ DA (ang. Domain Authority) i MOZ PA (ang. Page Authority), udostępniane przez firmę MOZ, określają autorytet strony WWW, na który wpływ ma jakość optymalizacji w obrębie strony i poza nią. Algorytm wyliczania wartości tych parametrów nie jest publicznie znany, jednak zakłada się, że im większy autorytet domeny lub strony, tym większe prawdopodobieństwo uzyskania dobrej pozycji w wyszukiwarkach. Przyjęto założenie, że wartości rankingowe MOZ DA i MOZ PA są ważnym elementem oceny jakości strony WWW.

Pseudokod 2. przedstawia funkcję obliczającą punktację za wartości rankingowe MOZ DA i MOZ PA w przedziale od 0,00 do 12,00 punktów.

```
# <0.00, 12.00>
Wejście:
  MOZ_DA - wartość MOZ DA dla domeny
  MOZ_PA - wartość MOZ PA dla strony głównej
Wyjście:
  P - wartość w przedziale <0.00, 12.00>

Utwórz zmienną P ← 0

If MOZ_DA > 0 or MOZ_PA > 0:
  P ← ((MOZ_DA + MOZ_PA) / 2) * 0.12
  If P > 12: P ← 12

Zwróć P
```

Pseudokod 2. Funkcja obliczająca punktację za wartości rankingowe MOZ DA i MOZ PA

Punktacja za wartość rankingową Alexa Rank

Ranking Alexa jest interpretowany jako wskaźnik popularności strony w internecie. Im ten wskaźnik jest niższy, tym bardziej popularna od pozostałych jest dana strona WWW. Przyjmuje się, że jeśli wartość Alexa Rank jest poniżej 100 000, oznacza to, że strona jest bardzo popularna wśród internautów.

Witryna internetowa Alexa.com po prawie 30 latach działalności, 1 maja 2022 roku, została zamknięta przez jej właściciela, firmę Amazon, a zamknięcie dostępu do bazy danych za pośrednictwem API planuje się na 15 grudnia 2022 roku[17].

[17] Alexa.com, https://support.alexa.com/hc/en-us/articles/4411466276375, maj 2022 r.

Pseudokod 3. przedstawia funkcję obliczającą punktację za wartość rankingową Alexa Rank w przedziale od 0,00 do 10,00 punktów.

```
# <0.00, 10.00>
Wejście:
  ALEXA_RANK – wartość Alexa Rank dla strony głównej
Wyjście:
  P - wartość w przedziale <0.00, 10.00>

Utwórz zmienną P ← 0

If ALEXA_RANK > 0 and ALEXA_RANK <= 1000000:
  P ← 10 - (10 * (ALEXA_RANK / 1000000))

Zwróć P
```

Pseudokod 3. Funkcja obliczająca punktację za wartość rankingową Alexa Rank

Punktacja za liczbę hiperłączy zewnętrznych

Parametr MOZ EUID, udostępniany przez firmę MOZ, zawiera informację o liczbie wysokiej jakości hiperłączy przychodzących do strony WWW (ang. link equity). Hiperłącze uznane za wysokiej jakości powinno mieć wysoki autorytet w rankingu MOZ i być powiązane tematycznie z linkowaną stroną WWW. Przyjęto założenie, że liczba wysokiej jakości odnośników kierujących do strony WWW jest ważnym czynnikiem oceny jakości strony.

Pseudokod 4. przedstawia funkcję obliczającą punktację za liczbę hiperłączy zewnętrznych w przedziale od 0,00 do 5,00 punktów.

```
# <0.00, 5.00>
Wejście:
  MOZ_EUID - wartość MOZ EUID dla strony głównej
Wyjście:
  P - wartość w przedziale <0.00, 5.00>

Utwórz zmienną P ← 0

P ← MOZ_EUID * 0.01
If P > 5:
  P ← 5

Zwróć P
```

Pseudokod 4. Funkcja obliczająca punktację za liczbę hiperłączy zewnętrznych

Punktacja za wtyczki społecznościowe

Wtyczki społecznościowe jako interaktywne elementy na stronie WWW pozwalają na zwiększenie zasięgu i zainteresowania profilami społecznościowymi ich twórców. Wtyczki społecznościowe umożliwiają szybkie udostępnianie treści ze strony WWW na profilach osób, które ją odwiedzają, co pozwala bezpłatnie je promować [170]. Założono, że informacja o wykorzystaniu wtyczek społecznościowych na stronie WWW jest ważnym czynnikiem oceny jej jakości.

Pseudokod 5. przedstawia funkcję obliczającą punktację za wykorzystanie wtyczek społecznościowych w przedziale od 0,00 do 1,00 punktu.

```
# <0.00, 1.00>
Wejście:
  SOCIAL_MEDIA(PLUGIN) - tablica z listą wykorzystanych wtyczek
        społecznościowych (Facebook, Twitter, LinkedIn, Google)
Wyjście:
  P - wartość w przedziale <0.00, 1.00>

Utwórz zmienną P ← 0

For each PLUGIN € SOCIAL_MEDIA:
  If PLUGIN = True: # if used
    P ← P + 0.25

Zwróć P
```

Pseudokod 5. Funkcja obliczająca punktację za wykorzystanie wtyczek społecznościowych

Punktacja za polecenia na portalach społecznościowych

Przyjęto założenie, że informacja o liczbie poleceń strony WWW na portalach społecznościowych wpływa na ranking w wyszukiwarkach. Biorąc pod uwagę znaczenie portali społecznościowych, polecenie strony WWW poprawia jej wizerunek i wiarygodność [73].

Pseudokod 6. przedstawia funkcję obliczającą punktację za liczbę poleceń na portalach społecznościowych w przedziale od 0,00 do 2,00 punktów.

```
# <0.00, 2.00>
Wejście:
  SOCIAL_MEDIA(SHARES) - tablica z liczbą poleceń na portalach
                społecznościowych (Facebook, Twitter, Google)
Wyjście:
  P - wartość w przedziale <0.00, 2.00>

Utwórz zmienną P ← 0

For each SHARES € SOCIAL_MEDIA:
  If SHARES >= 10 and SHARES < 100:
    P ← P + (SHARES * 0.01)
```

```
  If SHARES >= 100 and SHARES < 1000:
    P ← P + (SHARES * 0.001) + 1
  If SHARES >= 1000:
    P ← P + 2
If P > 2:
  P ← 2

Zwróć P
```

Pseudokod 6. Funkcja obliczająca punktację za liczbę poleceń na portalach społecznościowych

Punktacja za liczbę znaków w nazwie domeny

Przyjęto założenie, że krótkie nazwy domen łatwiej jest zapamiętać, zaprezentować graficznie czy wpisać do przeglądarki internetowej. Założono, że im liczba znaków w nazwie domeny jest mniejsza, tym większe jest prawdopodobieństwo wyższych pozycji rankingowych w SERP. Ustalono, że punkty zostaną przyznane domenom o nazwach składających się z mniej niż 14 znaków, natomiast wystąpienie znaku łącznika w nazwie domeny proporcjonalnie obniży punktację.

Pseudokod 7. przedstawia funkcję obliczającą punktację za liczbę znaków w nazwie domeny w przedziale od 0,00 do 2,00 punktów.

```
# <0.00, 2.00>
Wejście:
  COUNT_CHAR - liczba znaków w nazwie domeny
  COUNT_DASH - liczba wystąpień znaku łącznika w nazwie domeny
Wyjście:
  P - wartość w przedziale <0.00, 2.00>

Utwórz zmienną P ← 0

If COUNT_CHAR <= 13:
  P ← (13 - COUNT_CHAR) * 0.2
  If COUNT_DASH > 0:
    P ← P * (1 - (COUNT_DASH * 0.1))
  If P > 2:
    P ← 2

Zwróć P
```

Pseudokod 7. Funkcja obliczająca punktację za liczbę znaków w nazwie domeny

Punktacja za szyfrowanie SSL

Certyfikat SSL (ang. Secure Sockets Layer) gwarantuje bezpieczeństwo danych w trakcie ich przesyłania pomiędzy przeglądarką a serwerem WWW, wykorzystując do tego algorytmy szyfrujące [171]. Przyjęto założenie, że strona WWW powinna być dostępna

poprzez szyfrowane połączenie, w celu zapewnienia bezpieczeństwa udostępnianych informacji.

Pseudokod 8. przedstawia funkcję obliczającą punktację za wykorzystanie na stronie WWW szyfrowania SSL, w przedziale od 0,00 do 2,00 punktów.

```
# <0.00, 2.00>
Wejście:
  SSL - czy strona WWW stosuje szyfrowanie SSL? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 2.00>

Utwórz zmienną P ← 0

If SSL = True:
  P ← 2

Zwróć P
```

Pseudokod 8. Funkcja obliczająca punktację za wykorzystanie szyfrowania SSL

Punktacja za fizyczną lokalizację serwera WWW

Fizyczna lokalizacja serwera WWW ma istotny wpływ na szybkość ładowania stron WWW, co przekłada się na wartość wskaźnika TTFB [172] (ang. Time to First Byte), wykorzystywanego przez wyszukiwarki internetowe do ustalania rankingu. Założono, że utrzymanie strony WWW na serwerach zlokalizowanych na terenie kraju, do którego należy domena najwyższego poziomu – ccTLD (ang. country code top-level domain), obniża wskaźnik TTFB, co ma pozytywny wpływ na jej ranking w wyszukiwarkach na danym obszarze geograficznym.

Pseudokod 9. przedstawia funkcję obliczającą punktację za fizyczną lokalizację serwera WWW w przedziale od 0,00 do 1,00 punktu.

```
# <0.00, 1.00>
Wejście:
  HOST_CCTLD - czy serwis WWW jest utrzymywany na serwerze w kraju
               zgodnie z ccTLD? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 1.00>

Utwórz zmienną P ← 0

If HOST_CCTLD = True:
  P ← 1

Zwróć P
```

Pseudokod 9. Funkcja obliczająca punktację za fizyczną lokalizację serwera WWW

Punktacja za rejestrację serwera WWW w bazach DNSbl

O jakości i wiarygodności stron WWW świadczy to, czy są utrzymywane na serwerach WWW zarejestrowanych w bazach DNSbl jako źródła rozprzestrzeniania się spamu [173]. Zwykle poczta elektroniczna wysłana z adresu e-mail znajdującego się na takim serwerze jest blokowana przez serwery odbiorców, co negatywnie wpływa na wizerunek firmy promującej swoje usługi na stronie WWW. Przyjęto regułę, że punktacja zostanie proporcjonalnie obniżona w zależności od liczby baz DNSbl zwracających adres IP serwera, który utrzymuje stronę WWW.

Pseudokod 10. przedstawia funkcję obliczającą punktację za rejestrację serwera WWW w bazach DNSbl w przedziale od −2,00 do 0,00 punktów.

```
# <-2.00, 0.00>
Wejście:
  COUNT_DNSbl - liczba baz DNSbl zwracających adres IP serwera WWW
Wyjście:
  P - wartość w przedziale <-2.00, 0.00>

Utwórz zmienną P ← 0

If COUNT_DNSbl > 0:
  P ← (0 - COUNT_DNSbl) * 0.1
If P < -2:
  P ← -2

Zwróć P
```

Pseudokod 10. Funkcja obliczająca punktację za rejestrację serwera WWW w bazach DNSbl

Punktacja za publikowanie adresów e-mail

Publikowanie adresów e-mail na stronie WWW umożliwia robotom internetowym ich odczytanie i późniejsze wykorzystanie w systemach do wysyłania spamu. Taki dostępny publicznie adres często trafia do wielu baz danych, z których w praktyce nie można go wypisać. Założono, że ponieważ istnieją technologie ułatwiające publikowanie adresów e-mail na stronie w sposób uniemożliwiający botom ich odczytanie [174], informacja o odnalezionych adresach e-mail w kodzie zostanie wykorzystana do obniżenia punktacji.

Pseudokod 11. przedstawia funkcję obliczającą punktację za adresy e-mail odnalezione w kodzie strony WWW w przedziale od −1,00 do 0,00 punktów.

```
# <-1.00, 0.00>
Wejście:
  COUNT_EMAIL - liczba adresów e-mail odnalezionych w kodzie strony WWW
Wyjście:
  P - wartość w przedziale <-1.00, 0.00>

Utwórz zmienną P ← 0
```

```
If COUNT_EMAIL > 0:
  P ← -1

Zwróć P
```

Pseudokod 11. Funkcja obliczająca punktację za adresy e-mail odnalezione w kodzie strony WWW

3.1.3.2. Etap drugi – punktacja dla podstron strony WWW

Punktacja za wtyczki społecznościowe

W celu obliczenia punktacji badane jest wykorzystanie wtyczek społecznościowych oraz zastosowanie narzędzi z portalu AddThis[18]. Serwis AddThis umożliwia integrację strony z portalami społecznościowymi bez konieczności korzystania z ich wtyczek, za pomocą własnych narzędzi programistycznych, umieszczanych w kodzie HTML [175]. Przyjęto regułę, że użycie wtyczek za pośrednictwem portali społecznościowych będzie wyżej punktowane niż wykorzystanie do tego celu narzędzi AddThis. Przyjęto również zasadę, że punktacja zostanie obniżona, kiedy wtyczka lub narzędzie AddThis nie zostaną wykryte w kodzie strony WWW.

Pseudokod 12. przedstawia funkcję obliczającą punktację za wykorzystanie wtyczek społecznościowych w przedziale od −0,25 do 5,25 punktu.

```
# <-0.25, 5.25>
Wejście:
  SOCIAL_MEDIA(PLUGIN) - tablica z listą wykorzystanych wtyczek
  społecznościowych (Facebook, Twitter, LinkedIn, Google, AddThis)
Wyjście:
  P - wartość w przedziale <-0.25, 5.25>

Utwórz zmienne P ← 0 i A ← False

For each PLUGIN € SOCIAL_MEDIA:
  If PLUGIN = True: # if used
    If ADDTHIS = True:  # plugin via AddThis service
      A ← True
      P ← P + 0.25
    elif:
      P ← P + 1.25

If P = 0:
  P ← -0.25
If P > 0 and A = True:
  P ← P + 0.25

Zwróć P
```

Pseudokod 12. Funkcja obliczająca punktację za wykorzystanie wtyczek społecznościowych

[18] AddThis, https://www.addthis.com/get, maj 2022 r.

Punktacja za wykorzystanie narzędzi Google

Firma Google udostępnia projektantom stron WWW narzędzia programistyczne, takie jak Google Analytics, umożliwiające pozyskanie informacji o ruchu na stronie, Google Maps, przeznaczone do prezentowania map i zdjęć lotniczych powierzchni Ziemi, czy Google AdSense, które pozwala wyświetlać kontekstowe reklamy tekstowe, banery oraz reklamy wideo. Uruchomiony w 2011 roku portal społecznościowy Google Plus[19] oferował nowe kanały komunikacji pomiędzy użytkownikami [176], a do zintegrowania go ze stroną WWW trzeba było stosować specjalne biblioteki programistyczne.

Pseudokod 13. przedstawia funkcję obliczającą punktację za wykorzystanie narzędzi Google w przedziale od 0,00 do 2,75 punktu.

```
# <0.00, 2.75>
Wejście:
  GOOGLE(TOOL) - tablica z listą wykorzystanych narzędzi Google
                 (Analytics, AdSense, Maps, Plus)
Wyjście:
  P - wartość w przedziale <0.00, 2.75>

Utwórz zmienną P ← 0

for each TOOL € GOOGLE:
  if TOOL in (Analytics, AdSense):
    P ← P + 1
  if TOOL = Maps:
    P ← P + 0.5
  if TOOL = Plus:
    P ← P + 0.25

Zwróć P
```

Pseudokod 13. Funkcja obliczająca punktację za wykorzystanie narzędzi Google

Punktacja za stosowane technologie do publikowania treści multimedialnych

Treści multimedialne umieszczone na stronie WWW wzmacniają siłę przekazu zawartych na niej informacji. Popularny serwis YouTube wykorzystuje do wyświetlania filmów technologie HTML5 i FLV [177], a najnowsza wersja technologii Silverlight firmy Microsoft umożliwia publikowanie treści multimedialnych przy współpracy z akceleratorami graficznymi 3D. Choć technologia Silverlight nie jest już wspierana przez firmę Microsoft[20], nadal istnieją wtyczki do przeglądarek internetowych umożliwiających odtwarzanie treści multimedialnych zapisanych w tym formacie.

[19] Portal społecznościowy Google Plus zamknięto 2 kwietnia 2019 r.
[20] Microsoft, https://support.microsoft.com/help/4511036/silverlight-end-of-support, maj 2022 r.

Pseudokod 14. przedstawia funkcję obliczającą punktację za stosowane technologie do publikowania treści multimedialnych w przedziale od 0,00 do 1,00 punktu.

```
# <0.00, 1.00>
Wejście:
  YT - czy strona WWW publikuje treści z portalu YouTube? (T/F)
  SL - czy strona WWW publikuje treści w formacie Silverlight? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 1.00>

Utwórz zmienną P ← 0

If YT = True:
  P ← 0.85
If SL = True:
  P ← P + 0.15

Zwróć P
```

Pseudokod 14. Funkcja obliczająca punktację za stosowane technologie do publikowania treści multimedialnych

Punktacja za udostępnianie dokumentów biurowych

Umieszczenie dokumentów biurowych na stronie WWW, zazwyczaj w formacie PDF, często jest podyktowane chęcią przekazania dodatkowych informacji związanych z tematyką strony. Dotyczy to również formatów biurowych pakietu Microsoft 365 czy Apache OpenOffice, które umożliwiają zapisanie informacji w arkuszu kalkulacyjnym, bazie danych czy prezentacji multimedialnej.

Pseudokod 15. przedstawia funkcję obliczającą punktację za udostępnianie dokumentów biurowych w przedziale od 0,00 do 1,50 punktu.

```
# <0.00, 1.50>
Wejście:
  DOC(TYPE) - tablica z listą dokumentów umieszczonych na stronie WWW
              (Microsoft Office, Microsoft PowerPoint, Apache OpenOffice)
Wyjście:
  P - wartość w przedziale <0.00, 1.50>

Utwórz zmienną P ← 0

For each TYPE € DOC:
  P ← P + 0.5

Zwróć P
```

Pseudokod 15. Funkcja obliczająca punktację za udostępnianie dokumentów w formatach biurowych

Punktacja za komunikację przez komunikatory internetowe

Komunikatory internetowe jako osobne aplikacje lub rozszerzenia dla przeglądarek pozwalają na kontakt w formie czatu. Przyjęto regułę, że umieszczenie w kodzie HTML strony WWW informacji o możliwości wykorzystania do kontaktu komunikatorów, takich jak Skype, Jabber czy ICQ, będzie punktowane.

Pseudokod 16. przedstawia funkcję obliczającą punktację za umieszczenie w kodzie strony WWW informacji o możliwości kontaktu za pomocą komunikatora internetowego w przedziale od 0,00 do 0,50 punktu.

```
# <0.00, 0.50>
Wejście:
  MESSENGER - czy w kodzie HTML strony WWW jest informacja o
              możliwości kontaktu za pomocą komunikatora? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 0.50>

Utwórz zmienną P ← 0

If MESSENGER = True:
  P ← 0.5

Zwróć P
```

Pseudokod 16. Funkcja obliczająca punktację za komunikowanie się poprzez komunikatory internetowe

Punktacja za brak hiperłączy wychodzących

Liczba hiperłączy wychodzących (ang. outbound links) ma znaczenie dla algorytmów rankingowych opartych na analizie struktury połączeń pomiędzy stronami WWW. Założono, że brak hiperłączy wychodzących wpływa pozytywnie na pozycję rankingową w wyszukiwarkach, i informacja ta zostanie wykorzystana do obliczenia punktacji.

Pseudokod 17. przedstawia funkcję obliczającą punktację za brak hiperłączy wychodzących w przedziale od 0,00 do 1,00 punktu.

```
# <0.00, 1.00>
Wejście:
  COUNT_OUT_LINK - liczba hiperłączy wychodzących
Wyjście:
  P - wartość w przedziale <0.00, 1.00>

Utwórz zmienną P ← 0

If COUNT_OUT_LINK = 0:
  P ← 1

Zwróć P
```

Pseudokod 17. Funkcja obliczająca punktację za brak hiperłączy wychodzących

3.1.3.3. Punktacja końcowa

Ostateczna liczba punktów za wykorzystane technologie i pozycje rankingowe jest wyznaczana na podstawie wyników uzyskanych na obydwu etapach obliczeń. Wynik z drugiego etapu dodatkowo jest korygowany o współczynnik *LR*.

Pseudokod 18. przedstawia funkcję obliczającą końcową punktację za wykorzystane technologie i pozycje rankingowe – *PM*, w przedziale od 0,00 do 20,00 punktów.

```
# <0.00, 20.00>
Wejście:
  HP - punktacja dla strony głównej
  AP - średnia wartość punktowa dla wszystkich podstron
  LR - wartość współczynnika korygującego
Wyjście:
  PM - wartość w przedziale <0.00, 20.00>

Utwórz zmienną PM ← 0

If AP > 0:
  PM ← HP + (AP / LR)
elif:
  PM ← HP + (AP * LR)

If PM < 0:
  PM ← 0
elif PM > 20:
  PM ← 20

Zwróć PM
```

Pseudokod 18. Funkcja obliczająca punktację za wykorzystane technologie i pozycje rankingowe – *PM*

3.1.4. Punktacja za optymalizację kodu źródłowego

Obliczenie punktacji za optymalizację kodu źródłowego pod kątem wyszukiwarek – *PK*, przebiega w dwóch etapach. Na pierwszym etapie do obliczenia punktacji dla strony WWW jako całości wykorzystywane są informacje o niepowtarzalności tytułów i opisów stron zawartych w znacznikach META oraz zastosowaniu mikroformatów, liczba odnalezionych w kodzie HTML adresów e-mail i liczba adresów URL zwracających kod błędu. Na tym etapie strona WWW może uzyskać od 0,00 do 4,88 punktu.

Na drugim etapie obliczana jest punktacja dla każdej podstrony, gdzie analizowane są takie czynniki jak wykorzystanie znaczników HTML pod kątem SEO, optymalizacja wielkości kodu, relacja rozmiaru tekstu do kodu źródłowego, użycie stylów kaskadowych i kodu JavaScript w zewnętrznych plikach, stosunek liczby odnośników zewnętrznych do wewnętrznych. Na tym etapie każda podstrona może uzyskać od −8,40 do 20,55 punktu,

po czym obliczana jest wartość średnia z punktacji wszystkich podstron. Na drugim etapie punktacja jest korygowana o współczynnik *LR*.

Punktacja końcowa jest sumą punktów uzyskanych na pierwszym i drugim etapie. Choć teoretycznie liczba punktów za optymalizację kodu źródłowego mieści się w zakresie od −8,40 do 25,43, przyjęto regułę, że w razie uzyskania wartości ujemnej punktacja jest ustalana na 0 punktów, natomiast wartość maksymalna to 20 punktów. Przykładowo, jeśli strona WWW uzyska −6,45 lub 22,85 punktu, to wynik końcowy zostanie ustalony, odpowiednio, na 0 lub 20 punktów.

3.1.4.1. Etap pierwszy – punktacja dla strony WWW jako całości

Punktacja za stosowanie mikroformatów

Mikroformaty, jako uzupełnienie składni języka HTML, pozwalają wyróżnić dane strukturalne, dzięki którym boty są w stanie odpowiednio, jak też płynnie, interpretować treść strony oraz jej poszczególne elementy, takie jak grafika czy nagłówki. Dane strukturalne pozwalają rozszerzyć wyniki wyszukiwania lokalnego dla firm o dodatkowe informacje na temat godzin pracy, numerów telefonów czy opinii klientów, a wyniki z produktami mogą być uzupełnione informacją o aktualnych cenach i dostępności [178]. Prawidłowo wdrożone dane strukturalne pozwalają wyszukiwarkom wzbogacić wyniki wyszukiwań, dzięki czemu strona WWW może się pojawić w wynikach z elementami rozszerzonymi, np. w kartach informacyjnych (ang. rich snippets).

Poniżej przedstawiono praktyczne zastosowanie danych strukturalnych w kodzie strony WWW. Pierwszy kod HTML (listing 1.) opisuje dane osobowe w sposób typowy dla stron niestosujących mikroformatów, natomiast drugi kod HTML (listing 2.) opisuje je z zastosowaniem formatu hCard, umożliwiającego prezentację danych kontaktowych przedsiębiorstw, pracowników firm czy lokalizację miejsc.

```
<div>
  <div>Jan Kowalski</div>
  <div>Jan</div>
  <div>Nazwa Firmy</div>
  <div>123-456-789</div>
  <a href="https://nazwafirmy.pl">https://nazwafirmy.pl</a>
</div>
```

Listing 1. Kod HTML przed zastosowaniem mikroformatów

```
<head profile="http://www.w3.org/2006/03/hcard">
</head>
<div class="vcard">
  <div class="fn">Jan Kowalski</div>
  <div class="nickname">Jan</div>
  <div class="org">Nazwa Firmy</div>
  <div class="tel">123-456-789</div>
  <a class="url" href="https://nazwafirmy.pl">https://nazwafirmy.pl</a>
</div>
```

Listing 2. Kod HTML po zastosowaniu mikroformatów

Pseudokod 19. przedstawia funkcję obliczającą punktację za stosowanie danych strukturalnych w przedziale od 0,00 do 1,00 punktu.

```
# <0.00, 1.00>
Wejście:
  MICROFORMAT - czy strona WWW stosuje mikroformaty? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 1.00>

Utwórz zmienną P ← 0

If MICROFORMAT = True:
  P ← 1

Zwróć P
```

Pseudokod 19. Funkcja obliczająca punktację za stosowanie mikroformatów

Punktacja za poprawność składni kodu HTML

Przeglądarki internetowe, a także programy i urządzenia wspomagające, takie jak czytniki ekranu, linijki brajlowskie czy przełączniki, opierają się na informacjach zawartych w kodzie HTML. Prawidłowo napisany kod HTML, oparty na standardach W3C, powinien być zgodny z deklaracją znacznika DOCTYPE, wolny od błędów i poprawny semantycznie. Wszystkie elementy treści wprowadzone z użyciem języka znaczników powinny mieć pełne znaczniki początkowe i końcowe, elementy powinny być zagnieżdżane zgodnie ze specyfikacją, nie mogą mieć zduplikowanych atrybutów, a jeśli mają atrybut ID, to powinien on być niepowtarzalny [179].

Pseudokod 20. przedstawia funkcję obliczającą punktację za brak błędów w składni kodu HTML w przedziale od 0,00 do 0,40 punktu.

```
# <0.00, 0.40>
Wejście:
  HTML_SYNTAX_OK - czy składnia kodu HTML jest poprawna? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 0.40>
```

```
Utwórz zmienną P ← 0

If HTML_SYNTAX_OK = True:
  P ← 0.4

Zwróć P
```

Pseudokod 20. Funkcja obliczająca punktację za brak błędów w składni kodu HTML

Punktacja za ukrycie adresów e-mail w kodzie HTML

Ukrywanie adresu e-mail przed botami w kodzie HTML strony WWW jest ważnym elementem ochrony przed niechcianą pocztą – spamem. Istnieje wiele technik umożliwiających maskowanie adresu e-mail na stronie WWW, choćby wykorzystanie stylów CSS czy funkcji języka JavaScript [180]. Prawidłowe umieszczenie adresu e-mail na stronie WWW jest tym elementem optymalizacji w obrębie strony WWW, który może mieć wpływ na zmniejszenie poziomu otrzymywanego spamu.

Pseudokod 21. przedstawia funkcję obliczającą punktację za ukrycie adresów e-mail w kodzie HTML w przedziale od 0,00 do 0,40 punktu.

```
# <0.00, 0.40>
Wejście:
  EMAIL - czy w kodzie HTML wykryto adres e-mail? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 0.40>

Utwórz zmienną P ← 0

If not EMAIL = True:
  P ← 0.4

Zwróć P
```

Pseudokod 21. Funkcja obliczająca punktację za ukrycie adresów e-mail w kodzie HTML

3.1.4.2. Etap drugi – punktacja dla podstron strony WWW

Punktacja za wykryte słowa kluczowe

Optymalizacja w obrębie strony WWW jest związana z odpowiednim umieszczeniem najważniejszych słów kluczowych, tak aby wyszukiwarki internetowe mogły je kategoryzować i klasyfikować [181]. Optymalizacja polega na umieszczeniu słów kluczowych w znacznikach TITLE, META DESCRIPTION, H1–H6, B i STRONG, zwanych dalej znacznikami „TDHB”. Istotne jest to, aby słowa kluczowe umieszczone w tych

znacznikach znajdowały się także w strefach tekstowych na stronie WWW, zwykle ujętych w znacznikach P i DIV oraz w znacznikach definiujących nagłówki, listy i tabele [182]. Przyjęto założenie, że o wadze danego słowa znajdującego się w znacznikach „TDHB" decydować będzie liczba jego wystąpień w tekście na stronie WWW. Im liczba wystąpień danego słowa będzie większa, tym wyższa będzie jego ranga względem pozostałych słów. Założono, że dane słowo jest kluczowe dla strony WWW, jeśli jego gęstość (ang. keyword density) w tekście wynosi co najmniej 1%. Zakładając, że konkretne słowa kluczowe mogą być odmieniane w zależności od kontekstu, przyjęto regułę, że dopuszczalne podobieństwo pomiędzy słowami ujętymi w znacznikach „TDHB" a słowami występującymi w tekście strony WWW jest ustalane za pomocą algorytmu Levenshteina, jako miary do ustalania odległości między słowami kluczowymi [183]. Przyjęto też zasadę, że słowa są do siebie podobne, jeśli są identyczne po dokonaniu maksymalnie jednego przekształcenia algorytmem Levenshteina. Ponadto przyjęto regułę, że do obliczenia punktacji wystarczy wiedza na temat pierwszych 30 słów kluczowych wykrytych na stronie WWW.

Pseudokod 22. przedstawia funkcję obliczającą punktację za wykryte słowa kluczowe w przedziale od −0,50 do 4,00 punktów.

```
# <-0.50, 4.00>
Wejście:
  V1(W_TDHB) - tablica z listą słów w znacznikach "TDHB"
  V2(W_TEXT) - tablica z listą słów w strefach tekstowych
Wyjście:
  P - wartość w przedziale <-0.50, 4.00>

Utwórz zmienną P ← 0
Utwórz tablicę KEYS ← []

For each W_TDHB € V1:
  For each W_TEXT € V2:
    If density(W_TEXT) >= 1: # gęstość > 1%
      If W_TDHB = W_TEXT:
        KEYS[] ← W_TDHB
      elif W_TDHB = Levenshtein(1, W_TEXT): # odległość Levenshteina
        KEYS[] ← W_TDHB
  If count(KEYS) > 30:
    break

P ← -0.5 + count(KEYS) * 0.15
If P > 4:
  P ← 4

Zwróć P
```

Pseudokod 22. Funkcja obliczająca punktację za wykryte słowa kluczowe

Punktacja za wykorzystanie znacznika A

Struktura hiperłączy ma istotne znaczenie dla wyszukiwarek internetowych. Nawet niewielkie strony WWW składające się przynajmniej z kilku podstron powinny tworzyć spójną strukturę. Wykorzystuje się w tym celu linkowanie wewnętrzne (ang. internal linking), ułatwiające odwiedzającym poruszanie się po poszczególnych podstronach i szybkie zapoznawanie się z zawartymi na nich informacjami. Linkowanie wewnętrzne odgrywa ważną rolę w budowaniu – z myślą o botach wyszukiwarek – mapy witryny [184]. Ważną funkcję podczas linkowania pełni atrybut TITLE znacznika A, ponieważ definiuje on tytuł linkowanego elementu, a także ułatwia narzędziom dla osób niedowidzących jego poprawną interpretację. Przykład definicji hiperłącza wykorzystującej znacznik A z atrybutem TITLE jest przedstawiony na listingu 3.

```
<a href="https://taxmobile.pl" title="Księgowość dla firm">TaxMobile</a>
```

Listing 3. Definicja hiperłącza z atrybutem TITLE wykorzystująca znacznik A

Hiperłącza wychodzące (ang. outbound links) to łącza, które wskazują na treści dostępne poza granicami serwisu WWW i odsyłają boty wyszukiwarek do wskazanego adresu. Ważnym elementem optymalizacji w obrębie strony WWW jest maksymalne ograniczenie na niej liczby hiperłączy wychodzących, tak aby zachowała wysoką pozycję rankingową w wyszukiwarkach. Założono, że optymalna liczba wszystkich hiperłączy na stronie wynosi od 6 do 100, łączy wewnętrznych – powyżej 5, a łączy wychodzących – od 0 do 5.

Pseudokod 23. przedstawia funkcję obliczającą punktację za wykorzystanie znacznika A w przedziale od −2,65 do 3,40 punktu.

```
# <-2.65, 3.40>
Wejście:
  A_ALL   - liczba wszystkich hiperłączy w kodzie HTML
  A_OUT   - liczba wszystkich hiperłączy wychodzących (outbound links)
  A_IN    - liczba wszystkich hiperłączy wewnętrznych (internal linking)
  A_TITLE - liczba wszystkich hiperłączy z atrybutem TITLE
Wyjście:
  P - wartość w przedziale <-2.65, 3.40>

Utwórz zmienne P ← 0 i T ← 0

# punktacja za liczbę wszystkich hiperłączy
# <-0.50, 0.30>
If A_ALL > 100: P ← P - 0.5
If A_ALL <= 75: P ← P + 0.1
If A_ALL <= 50: P ← P + 0.1
If A_ALL <= 25: P ← P + 0.1

# punktacja za liczbę hiperłączy wychodzących (outbound links)
# <-0.50, 0.30>
```

```
If A_OUT > 20: P ← P - 0.5
If A_OUT <= 5: P ← P + 0.1
If A_OUT <= 1: P ← P + 0.1
If A_OUT = 0: P ← P + 0.1

# punktacja za liczbę hiperłączy wewnętrznych (internal linking)
# <-0.15, 0.30>
If A_IN = 0: P ← P - 0.15
If A_IN > 5: P ← P + 0.1
If A_IN > 10: P ← P + 0.1
If A_IN > 15: P ← P + 0.1

# punktacja za relację hiperłączy in-out
# <-1.00, 1.00>
T ← (A_IN - A_OUT) * 0.1
If T < -1:
  T ← -1
If T > 1:
  T ← 1
P ← P + T

# punktacja za wykorzystanie atrybutu TITLE
# <-0.50, 1.50>
If A_ALL > 0:
  If A_TITLE > 0:
    T ← (A_TITLE / A_ALL) * 1.5
    If T > 1.5:
      T ← 1.5
    P ← P + T
  elif:
    P ← P - 0.5

Zwróć P
```

Pseudokod 23. Funkcja obliczająca punktację za wykorzystanie znacznika A

Punktacja za wykorzystanie znacznika IMG

Obrazy umieszczone na stronie WWW często są uzupełnieniem informacji zawartych w tekście. Botom wyszukiwarek trudno zrozumieć, co przedstawia konkretny obraz, dlatego podczas optymalizacji w obrębie strony WWW wykorzystuje się atrybut ALT znacznika IMG [185] do umieszczenia dodatkowych informacji, ułatwiających zdekodowanie zawartości pliku graficznego [186].

Podczas optymalizacji strony WWW dodaje się również informacje na temat wymiarów obrazów za pomocą atrybutów WIDTH i HEIGHT, co ma ułatwić ich poprawne wyświetlanie w przeglądarkach. Przykład definicji obrazu wykorzystującej znacznik IMG z atrybutami ALT, WIDTH i HEIGHT jest przedstawiony na listingu 4.

```
<img src="liscie.jpg" alt="Liście na drzewach" width="800" height="600">
```

Listing 4. Definicja obrazu z atrybutami ALT, WIDTH i HEIGHT wykorzystująca znacznik IMG

Pseudokod 24. przedstawia funkcję obliczającą punktację za wykorzystanie znacznika IMG w przedziale od −0,65 do 2,25 punktu.

```
# <-0.65, 2.25>
Wejście:
  IMG_ALL - liczba wszystkich obrazów w kodzie HTML
  IMG_ALT - liczba wszystkich obrazów z atrybutem ALT
  IMG_DIM - liczba wszystkich obrazów z atrybutem WIDTH i HEIGHT
Wyjście:
  P - wartość w przedziale <-0.65, 2.25>

Utwórz zmienne P ← 0 i T ← 0

# punktacja za wykorzystanie atrybutu ALT
# <-0.50, 1.50>
If IMG_ALL > 0:
  If IMG_ALT > 0:
    T ← (IMG_ALT / IMG_ALL) * 1.5
    If T > 1.5:
      T ← 1.5
    P ← P + T
  elif:
    P ← P - 0.5

# punktacja za wykorzystanie atrybutów WIDTH i HEIGHT
# <-0.15, 0.75>
If IMG_ALL > 0:
  If IMG_DIM > 0:
    T ← (IMG_DIM / IMG_ALL) * 0.75
    If T > 0.75:
      T ← 0.75
    P ← P + T
  elif:
    P ← P - 0.15

Zwróć P
```

Pseudokod 24. Funkcja obliczająca punktację za wykorzystanie znacznika IMG

Punktacja za wykorzystanie znaczników HTML

Prawidłowe wykorzystanie znaczników HTML na stronie WWW odpowiedzialnych za prezentowanie treści ma istotne znaczenie w procesie optymalizacji w obrębie strony WWW. Znaczniki H1–H6, pełniące funkcję nagłówków, informują wizualnie użytkownika i roboty wyszukiwarek o konstrukcji danej strony. Najwyżej w hierarchii znajduje się znacznik H1, odgrywający rolę tytułu, który powinien zawierać najważniejsze słowa kluczowe dotyczące strony WWW [187]. W odróżnieniu od znaczników H2–H6 znacznik H1 powinien wystąpić na stronie tylko raz.

Wykorzystanie dodatkowych znaczników, takich jak KBD – do oznaczenia fragmentu tekstu wprowadzanego z klawiatury, oraz SAMP lub CODE – do poinformowania botów wyszukiwarek, że objęta nimi zawartość to kod komputerowy, wzbogacają formę treści prezentowanych na stronie WWW. Choć wyszukiwarki potrafią rozpoznać zawartość umieszczoną w ramkach, w których wyświetlane są fragmenty innego dokumentu, stosowanie znacznika IFRAME z punktu widzenia optymalizacji w obrębie strony WWW nie jest zalecane [188]. Na listingu 5. przedstawiono przykład wykorzystania znaczników formatujących tekst na stronie WWW.

```
<!DOCTYPE html>
<head>
<meta charset="utf-8">
</head>
<body>
<table>
  <tr>
    <td width="20%" valign="top">
      <h1>Nagłówek 1</h1>
      <h2>Nagłówek 2</h2>
      <h3>Nagłówek 3</h3>
      <h4>Nagłówek 4</h4>
      <h5>Nagłówek 5</h5>
      <h6>Nagłówek 6</h6>
    </td>
    <td width="50%" valign="top">

<p>Aby skopiować tekst (Windows),
naciśnij <kbd>Ctrl</kbd> + <kbd>C</kbd></p>
      <p>Informacja systemowa:</p>
      <p><samp> Pliku nie odnaleziono.<br>
      Naciśnij klawisz F1, aby kontynuować.</samp></p>

      <p>Znacznik <code>button</code> umożliwia obsługę
      przycisku na stronie WWW.</p>

      <iframe src="https://pl.isowq.org" title="ISOWQ"></iframe>
    </td>
  </tr>
</table>
</body>
</html>
```

Listing 5. Przykład wykorzystania znaczników formatujących tekst na stronie WWW

Wynik działania kodu HTML z listingu 5. jest przedstawiony na rysunku 9.

Rysunek 9. Rezultat działania kodu HTML z listingu 5. w przeglądarce internetowej, opracowanie własne

Pseudokod 25. przedstawia funkcję obliczającą punktację za wykorzystanie znaczników HTML w przedziale od −0,50 do 1,65 punktu.

```
# <-0.50, 1.65>
Wejście:
  H[1-6] - liczba znaczników H<1-6> w kodzie HTML
  TK - liczba znaczników <kbd> w kodzie HTML
  TS - liczba znaczników <samp> w kodzie HTML
  TC - liczba znaczników <code> w kodzie HTML
  TI - liczba znaczników <iframe> w kodzie HTML
Wyjście:
  P - wartość w przedziale <-0.50, 1.65>

Utwórz zmienną P ← 0

If H1 > 1: P ← P - 0.25
If H1 = 1: P ← P + 0.25
If H2 > 0: P ← P + 0.25
If H3 + H4 + H5 + H6 > 0: P ← P + 0.25

If TK > 0: P ← P + 0.15
If TK > 1: P ← P + 0.15

If TC > 0: P ← P + 0.15
If TC > 1: P ← P + 0.15

If TS > 0: P ← P + 0.15
If TS > 1: P ← P + 0.15

If TI > 0: P ← P - 0.25

Zwróć P
```

Pseudokod 25. Funkcja obliczająca punktację za wykorzystanie znaczników HTML

Punktacja za wielkość tekstu zawartego w znacznikach P i A

Na wielkość kodu HTML składa się wiele elementów; jednym z nich jest tekst widoczny w przeglądarce, zwykle zawarty w znacznikach: P – definiującym akapit, oraz A – definiującym hiperłącze. Podczas optymalizacji w obrębie strony WWW istotne jest, aby wielkość kodu HTML, który zawiera znaczniki opisujące elementy dodatkowe, takie jak banery reklamowe oraz kod języka JavaScript czy stylów CSS, była optymalna w stosunku do wielkości kodu zawierającego właściwą treść – zazwyczaj tekst. Przyjęto zasadę, że punktacja zostanie obliczona tylko dla stron o wielkości do 50 kilobajtów.

Pseudokod 26. przedstawia funkcję obliczającą punktację za wielkość tekstu zawartego w znacznikach P i A w przedziale od 0,00 do 1,50 punktu.

```
# <0.00, 1.50>
Wejście:
  SIZE_HTML - rozmiar kodu HTML w bajtach
  SIZE_TEXT_IN_PA - rozmiar tekstu w znacznikach P i A w bajtach
Wyjście:
  P - wartość w przedziale <0.00, 1.50>

Utwórz zmienną P ← 0

If SIZE_HTML > 0 and SIZE_HTML <= 50000 and SIZE_TEXT_IN_PA > 0:
  P ← (SIZE_TEXT_IN_PA / SIZE_HTML) * 1.5
  If P > 1.5:
    P ← 1.5

Zwróć P
```

Pseudokod 26. Funkcja obliczająca punktację za wielkość tekstu zawartego w znacznikach P i A

Punktacja za strefy tekstowe

Tekst na stronie WWW pełni ważną funkcję informacyjną zarówno dla jej odbiorcy, jak i dla wyszukiwarek internetowych. Ważnym elementem optymalizacji w obrębie strony WWW jest wzbogacenie jej zawartości o merytoryczne informacje. Przyjęto założenie, że strefa tekstowa to miejsce na stronie WWW zawierające co najmniej 80 znaków, a liczba unikalnych stref tego rodzaju na stronie powinna być większa od 10.

Pseudokod 27. przedstawia funkcję obliczającą punktację za liczbę unikalnych stref tekstowych wykrytych na stronie WWW w przedziale od 0,00 do 0,50 punktu.

```
# <0.00, 0.50>
Wejście:
  TEXT_ZONES - liczba unikalnych stref tekstowych w kodzie HTML
Wyjście:
  P - wartość w przedziale <0.00, 0.50>
```

```
Utwórz zmienną P ← 0

If TEXT_ZONES > 10:
  P ← 0.5

Zwróć P
```

Pseudokod 27. Funkcja obliczająca punktację za liczbę unikalnych stref tekstowych wykrytych na stronie

Punktacja za wykorzystanie znaczników TITLE i META

Odpowiednie wykorzystanie znaczników TITLE i META ma istotne znaczenie w trakcie procesu optymalizacji w obrębie strony WWW. Znaczniki META same w sobie nie mają bezpośredniego wpływu na pozycję strony, jednak stanowią bardzo ważny element mechanizmu, na którym opierają się roboty wyszukiwarek.

Znacznik TITLE, opisujący tytuł strony WWW, to jeden z tych elementów, które mają największy wpływ na trafność wyszukiwania [189]. Optymalna długość tytułu strony powinna się zawierać w przedziale od 60 do 80 znaków.

Znacznik META DESCRIPTION to element kodu strony, którego celem jest opisanie jej zawartości. Stanowi jej skróconą reklamę tekstową, a także zawiera najważniejsze słowa kluczowe odnoszące się do tematyki umieszczonych na niej treści [190]. Optymalna długość opisu strony powinna się zawierać w przedziale od 140 do 180 znaków.

Znacznik META KEYWORDS był wykorzystywany przez wyszukiwarki podczas ustalania pozycji rankingowych, jednak obecnie nie ma on większego znaczenia[21]. Znaczniki META AUTHOR i META COPYRIGHT informują, kto jest autorem zawartych na stronie WWW informacji oraz do kogo należy własność praw autorskich. Znaczniki META GENERATOR i META DUBLIN CORE pozwalają określić nazwę edytora HTML, który wykorzystano do utworzenia strony WWW, oraz opisać stronę za pomocą metadanych standardu Dublin Core. Znaczniki META GOOGLE-SITE-VERIFICATION i META ISOWQ służą do weryfikacji strony WWW w narzędziu Google Search Console oraz w systemie rankingowym ISOWQ. Przykład wykorzystania znaczników TITLE i META w kodzie strony WWW jest przedstawiony na listingu 6.

```
<!DOCTYPE html>
<head>
<meta charset="utf-8">
<title>Tytuł strony WWW (60 - 80 znaków)</title>
<meta name="description" content="Opis strony WWW (140 - 180 znaków)">
<meta name="keywords" content="słowo kluczowe 1, słowo kluczowe 2">
<meta name="author" content="Autor strony WWW">
```

[21] Wyszukiwarka Google od 2009 roku ignoruje META KEYWORDS.

```
<meta name="copyright" content="Prawa autorskie do strony WWW">
<meta name="generator" content="System CMS strony WWW">
<meta name="google-site-verification" content="unique-string">
<meta name="DC.title" content="Tytuł strony WWW">
<meta name="DC.description" content="Opis strony WWW">
<meta http-equiv="refresh" content="10; url=https://www.isowq.org">
</head>
<body>
<p>Treść strony WWW</p>
</body>
</html>
```

Listing 6. Przykład wykorzystania znaczników TITLE i META w kodzie strony WWW

Znacznik META HTTP-EQUIV REFRESH umożliwia skonfigurowanie prostej formy przekierowania. Jego stosowanie nie jest zalecane [191]. Wyszukiwarka Google zaleca, by do informowania o nowym adresie URL strony WWW korzystać z przekierowania 301 po stronie serwera WWW.

Pseudokod 28. przedstawia funkcję obliczającą punktację za wykorzystanie znaczników TITLE i META w przedziale od –1,10 do 2,00 punktu.

```
# <-1.10, 2.00>
Wejście:
  LEN_TITLE - liczba znaków w znaczniku TITLE
  LEN_DESC - liczba znaków w znaczniku META DESCRIPTION
  LEN_KEYS - liczba znaków w znaczniku META KEYWORDS
  M_AUTHOR - czy użyty jest znacznik META AUTHOR? (T/F)
  M_COPYRIGHT - czy użyty jest znacznik META COPYRIGHT? (T/F)
  M_GENERATOR - czy użyty jest znacznik META GENERATOR? (T/F)
  M_GOOGLE - czy użyty jest znacznik META GOOGLE-SITE-VERIFICATION? (T/F)
  M_ISOWQ - czy użyty jest znacznik META ISOWQ? (T/F)
  M_DC - czy użyty jest znacznik META DUBLIN CORE? (T/F)
  M_REFRESH - czy użyty jest znacznik META HTTP-EQUIV REFRESH? (T/F)
Wyjście:
  P - wartość w przedziale <-1.10, 2.00>

Utwórz zmienną P ← 0

# TITLE
If LEN_TITLE >= 60 and LEN_TITLE <= 80:
  P ← P + 0.75
elif:
  P ← P - 0.1
If LEN_TITLE = 0:
  P ← P - 0.25

# META DESCRIPTION
If LEN_DESC >= 140 and LEN_DESC <= 180:
  P ← P + 0.75
elif:
  P ← P - 0.1
If LEN_DESC = 0:
  P ← P - 0.25
```

```
# META KEYWORDS
If LEN_KEYS > 80 and LEN_KEYS > LD:
  P ← P - 0.15

# META AUTHOR … META HTTP-EQUIV REFRESH
If M_AUTHOR = True:
  P ← P + 0.2
If M_COPYRIGHT = True:
  P ← P + 0.1
If M_GENERATOR = True:
  P ← P + 0.1
If M_GOOGLE = True:
  P ← P + 0.05
If M_ISOWQ = True:
  P ← P + 0.01
If M_DC = true:
  P ← P + 0.04
If M_REFRESH = True:
  P ← P – 0.25

Zwróć P
```

Pseudokod 28. Funkcja obliczająca punktację za wykorzystanie znaczników TITLE i META

Punktacja za wersję języka HTML

HTML5, będący rozwinięciem języków HTML 4 i XHTML 1, to obecny standard języka opisującego dokument hipertekstowy [192], a wraz z technologiami CSS i JavaScript jest podstawą do budowy nowoczesnych stron WWW [193]. Język HTML5 wprowadza nowe typy znaczników, umożliwiające rysowanie wykresów oraz odtwarzanie plików audio i wideo bez instalowania dodatkowych wtyczek [194]. Strona WWW stworzona za pomocą HTML5 będzie jednakowo przetwarzana przez wszystkie popularne przeglądarki internetowe, takie jak Google Chrome, Microsoft Edge, Firefox, Opera i Safari. Ze względu na rosnącą wydajność przeglądarek umożliwia budowę systemów do obrazowania wirtualnego [195].

Pseudokod 29. przedstawia funkcję obliczającą punktację za wersję języka HTML wykorzystaną do budowy strony WWW w przedziale od −0,25 do 0,30 punktu.

```
# <-0.25, 0.30>
Wejście:
  DOCTYPE - czy strona WWW wykorzystuje znacznik DOCTYPE? (T/F)
  HTML_VERSION - wersja języka HTML wykorzystana do budowy strony WWW
Wyjście:
  P - wartość w przedziale <-0.25, 0.30>

Utwórz zmienną P ← 0

If DOCTYPE = true:
  P ← 0.05
```

```
  If HTML_VERSION = XHTML:
    P ← P + 0.15
  If HTML_VERSION = HTML5:
    P ← P + 0.25
elif:
  P ← -0.25

Zwróć P
```

Pseudokod 29. Funkcja obliczająca punktację za wykorzystaną wersję języka HTML

Punktacja za wielkość kodu HTML

Wielkość kodu HTML wpływa na transfer danych, i choć serwery WWW wykorzystują metody ich kompresji, przekłada się to na szybkość ładowania stron WWW w przeglądarkach [196]. Przyjęto założenie, że optymalna wielkość kodu HTML strony powinna się zawierać w przedziale od 0,5 do 25 kilobajtów, a strony o wielkości powyżej 150 kilobajtów powodują nadmierne zużycie przepustowości łącza.

Pseudokod 30. przedstawia funkcję obliczającą punktację za wielkość kodu HTML w przedziale od −0,35 do 0,50 punktu.

```
# <-0.35, 0.50>
Wejście:
  SIZE - wielkość kodu HTML w bajtach
Wyjście:
  P - wartość w przedziale <-0.35, 0.50>

Utwórz zmienną P ← 0

If SIZE >= 500 and SIZE <= 25000: P ← P + 0.25
If SIZE >= 500 and SIZE <= 10000: P ← P + 0.25
If SIZE >= 150000: P ← P - 0.1
If SIZE >= 250000: P ← P - 0.25

Zwróć P
```

Pseudokod 30. Funkcja obliczająca punktację za wielkość kodu HTML

Punktacja za optymalizację wielkości kodu HTML

Optymalizacja w obrębie strony WWW wiąże się również ze zmodyfikowaniem kodu HTML tak, aby maksymalnie zmniejszyć jego wielkość, nie tracąc pierwotnej informacji. Modyfikacja taka obejmuje m.in. usunięcie z kodu zbędnych komentarzy, nadmiarowych znaków odstępu czy pustych linii [197]. Przyjęto zasadę, że punktacja zostanie obliczona tylko dla stron WWW o wielkości do 50 kilobajtów.

Pseudokod 31. przedstawia funkcję obliczającą punktację za wielkość kodu HTML w przedziale od 0,00 do 2,00 punktów.

```
# <0.00, 2.00>
Wejście:
  SIZE - wielkość kodu HTML w bajtach
  COMPRESS - rozmiar kodu HTML po kompresji w bajtach
Wyjście:
  P - wartość w przedziale <0.00, 2.00>

Utwórz zmienną P ← 0

If SIZE > 0 and SIZE <= 50000:
  P ← 0.8 - (COMPRESS / SIZE)
  If P >= -0.2 and P <= 0:
    P ← (abs(P) * 10) * (1 + (round(SIZE / 5000) * 0.1)
    If P > 2:
      P ← 2
  elif:
    P ← 0

Zwróć P
```

Pseudokod 31. Funkcja obliczająca punktację za optymalizację wielkości kodu HTML

Punktacja za optymalizację kodu w znacznikach STYLE i SCRIPT

Umieszczenie definicji stylów CSS czy kodu JavaScript bezpośrednio w kodzie strony WWW wpływa na jej rozmiar i szybkość ładowania się w przeglądarkach. Elementem optymalizacji w obrębie strony WWW jest przeniesienie stylów CSS i kodu JavaScript z kodu HTML do zewnętrznych plików [198]. Zastosowanie tej techniki powoduje, że pliki te mogą być załadowane do pamięci tymczasowej przeglądarki, co przełoży się na wzrost szybkości ładowania strony WWW. Przyjęto regułę, że w przypadku wykrycia kodu w znacznikach STYLE lub SCRIPT punktacja zostanie obniżona.

Pseudokod 32. przedstawia funkcję obliczającą punktację za optymalizację kodu w znacznikach STYLE i SCRIPT w przedziale od −0,60 do 0,00 punktów.

```
# <-0.60, 0.00>
Wejście:
  SIZE_CSS - rozmiar kodu w znacznikach STYLE w bajtach
  SIZE_JS - rozmiar kodu w znacznikach SCRIPT w bajtach
Wyjście:
  P - wartość w przedziale <-0.60, 0.00>

Utwórz zmienną P ← 0

If (SIZE_CSS > 500): P ← P - 0.15
If (SIZE_JS > 1500): P ← P - 0.15
If (SIZE_JS > 3000): P ← P - 0.15
```

```
If (SIZE_JS > 5000): P ← P - 0.15

Zwróć P
```

Pseudokod 32. Funkcja obliczająca punktację za optymalizację kodu w znacznikach STYLE i SCRIPT

Punktacja za wykorzystanie technologii Flash

Technologia Flash, uznana za przestarzałą i już niewspierana przez firmę Adobe, utrudnia (a w wielu przypadkach uniemożliwia) botom wyszukiwarek poprawne zdekodowanie wszystkich elementów wchodzących w skład strony WWW. Przydatnym dla projektantów stron WWW narzędziem jest biblioteka swfobject, która ułatwia obsługę obiektów Flash [199] przez umożliwienie wyświetlania treści alternatywnych w przypadku niewykrycia obsługi takich obiektów w przeglądarce. Alternatywą dla technologii Flash jest język HTML5, który wprowadził znaczniki umożliwiające obsługę multimediów [200].

Pseudokod 33. przedstawia funkcję obliczającą punktację za wykorzystanie biblioteki swfobject i technologii Flash w przedziale od −0,15 do 0,10 punktu.

```
# <-0.15, 0.10>
Wejście:
  FLASH - czy strona WWW wykorzystuje technologię Flash? (T/F)
  SWFOBJECT - czy strona WWW wykorzystuje bibliotekę swfobject? (T/F)
Wyjście:
  P - wartość w przedziale <-0.15, 0.10>

Utwórz zmienną P ← 0

If FLASH = True:
  P ← -0.15
  If SWFOBJECT = True:
    P ← P + 0.25

Zwróć P
```

Pseudokod 33. Funkcja obliczająca punktację za wykorzystanie technologii Flash

Punktacja za poprawność składni kodu HTML

Strona WWW zawierająca błędy składni kodu HTML może nie być poprawnie wyświetlana w przeglądarkach, pomimo że mają one wbudowane mechanizmy próbujące naprawić uszkodzony kod [201]. Punktacja jest uzależniona od liczby znalezionych błędów, wielkości kodu HTML i tekstu.

Pseudokod 34. przedstawia funkcję obliczającą punktację za poprawność składni kodu HTML w przedziale od −0,50 do 0,50 punktu.

```
# <-0.50, 0.50>
Wejście:
  SIZE_HTML - wielkość kodu HTML w bajtach
  SIZE_TEXT - wielkość tekstu w bajtach
  SIZE_ERRORS - liczba błędów w składni kodu HTML
Wyjście:
  P - wartość w przedziale <-0.50, 0.50>

Utwórz zmienną P ← 0

If SIZE_ERRORS > 0
  If (SIZE_HTML - SIZE_TEXT) > 0:
    P ← 0 - ((SIZE_ERRORS / (SIZE_HTML - SIZE_TEXT)) * 100) * 5
  If P < -0.5:
    P ← -0.5
elif:
  P ← 0.5

Zwróć P
```

Pseudokod 34. Funkcja obliczająca punktację za poprawność składni kodu HTML

Punktacja za konstrukcję adresu URL

Prosty i czytelny adres, nazywany również adresem przyjaznym (ang. friendly URL), w przejrzysty sposób wskazuje użytkownikowi, do którego miejsca serwisu się odnosi. Przyjazny adres URL tworzy ciąg znaków, które układają się w logiczną całość, nie zawiera przypadkowych liczb, liter czy znaków interpunkcyjnych. Przyjazne adresy ułatwiają logiczne uporządkowanie kategorii tematycznych na stronie WWW, co dla botów wyszukiwarek jest ułatwieniem podczas indeksacji [202]. Opracowanie prostych i zrozumiałych adresów jest ważnym elementem optymalizacji w obrębie strony WWW.

Pseudokod 35. przedstawia funkcję obliczającą punktację za konstrukcję adresu URL strony WWW w przedziale od −0,15 do 0,50 punktu.

```
# <-0.15, 0.50>
Wejście:
  URL_FRIENDLY - czy konstrukcja adresu URL strony WWW
                 jest czytelna dla użytkownika? (T/F)
Wyjście:
  P - wartość w przedziale <-0.15, 0.50>

Utwórz zmienną P ← 0

If URL_FRIENDLY = True:
  P ← 0.5
elif:
  P ← -0.15

Zwróć P
```

Pseudokod 35. Funkcja obliczająca punktację za konstrukcję adresu URL strony WWW

Punktacja za wykorzystanie znacznika META ROBOTS

Znacznik META ROBOTS jest stosowany głównie do blokowania indeksacji stron o niskiej wartości, duplikatów, stron logowania czy innych stron nieistotnych z punktu widzenia optymalizacji w obrębie strony WWW [203]. Zmiana nazwy znacznika pozwala stosować go oddzielnie dla każdej wyszukiwarki. Na listingu 7. przedstawiono przykład wykorzystania znacznika META ROBOTS dla wyszukiwarek Google, Bing i Yahoo!.

```
<!DOCTYPE html>
<head>
<meta name="robots" content="index, nofollow">
<meta name="googlebot" content="noarchive">
<meta name="bingbot" content="nosnippet">
<meta name="slurp" content="noarchive">
</head>
<body>
<p>Treść strony WWW<p>
</body>
</html>
```

Listing 7. Przykład wykorzystania znacznika META ROBOTS

Pseudokod 36. przedstawia funkcję obliczającą punktację za wykorzystanie znacznika META ROBOTS w przedziale od 0,00 do 0,35 punktu.

```
# <0.00, 0.35>
Wejście:
  META_ROBOTS - czy użyto znacznika META ROBOTS? (T/F)
  BOT_GOOGLE - czy użyto znacznika dla wyszukiwarki Google? (T/F)
  BOT_BING - czy użyto znacznika dla wyszukiwarki Bing? (T/F)
  BOT_YAHOO - czy użyto znacznika dla wyszukiwarki Yahoo!? (T/F)
Wyjście:
  P - wartość w przedziale <0.00, 0.35>

Utwórz zmienną P ← 0

If META_ROBOTS = True:
  P ← 0.05
  For each BOT form (BOT_GOOGLE, BOT_BING, BOT_YAHOO):
    If BOT = true:
      P ← P + 0.1

Zwróć P
```

Pseudokod 36. Funkcja obliczająca punktację za wykorzystanie znacznika META ROBOTS

Punktacja uzupełniająca za optymalizację kodu HTML

Ważnym uzupełnieniem optymalizacji w obrębie strony WWW jest przeniesienie definicji stylów CSS i kodu języków JavaScript i VisualBasic do osobnych plików [204] oraz zakodowanie adresów e-mail przed botami. Umieszczenie w kodzie HTML adresu URL do favikony[22] spowoduje wyświetlenie jej w wynikach wyszukiwania [205] i dodanie adresu URL do kanału RSS (ang. RDF Site Summary) oraz umożliwi szybki dostęp do często zmieniającej się treści na stronie WWW, a wdrożenie danych strukturalnych pozwoli wyszukiwarkom trafniej interpretować jej zawartość.

Pseudokod 37. przedstawia funkcję obliczającą punktację za optymalizację kodu HTML w przedziale od −1,00 do 1,00 punktu.

```
# <-1.00, 1.00>
Wejście:
  USE_CSS - czy strona WWW używa osobnych plików ze stylami CSS? (T/F)
  USE_JS - czy strona WWW używa osobnych plików z kodem JavaScript? (T/F)
  USE_VB - czy strona WWW używa osobnych plików z kodem VisualBasic? (T/F)
  USE_FAVICON - czy strona WWW korzysta z pliku ikony (favicon)? (T/F)
  USE_RSS - czy strona WWW udostępnia kanał RSS? (T/F)
  USE_MICROFORMATS - czy strona WWW stosuje mikroformaty? (T/F)
  COUNT_EMAIL - liczba adresów e-mail wykrytych w kodzie HTML strony WWW
Wyjście:
  P - wartość w przedziale <-1.00, 1.00>

Utwórz zmienne P ← 0 i T ← 0

if USE_CSS = true: P ← P + 0.1
if USE_JS = true: P ← P + 0.1
if USE_VB = true: P ← P + 0.1
if USE_FAVICON = true: P ← P + 0.05
if USE_RSS = true: P ← P + 0.15
if USE_MICROFORMATS = true: P ← P + 0.5
if COUNT_EMAIL > 0:
  T ← 0 - (COUNT_EMAIL * 0.25)
  if T < -1:
    T ← -1
  P ← P + T

Zwróć P
```

Pseudokod 37. Funkcja obliczająca punktację uzupełniającą za optymalizację kodu HTML

3.1.4.3. Punktacja końcowa

Istotnym elementem optymalizacji w obrębie strony WWW jest odpowiednie opracowanie tytułów i opisów na wszystkich jej podstronach. Informacje ujęte w znacznikach

[22] Ikona, która pojawia się przed adresem URL w polu adresowym przeglądarki internetowej.

TITLE i META DESCRIPTION powinny być unikalne na wszystkich podstronach i zawierać najważniejsze słowa kluczowe. Przyjęto regułę, że w przypadku, kiedy wszystkie tytuły lub opisy stron będą unikalne, końcowa punktacja dla strony WWW jako całości zostanie proporcjonalnie zwiększona w przedziale od 0,00 do 3,08 punktu (pseudokod 38.).

```
# <0.00, 3.08>
Wejście:
  T - czy wszystkie znaczniki TITLE są unikalne? (T/F)
  D - czy wszystkie znaczniki META DESCRIPTION są unikalne? (T/F)
  C - liczba stron poprawnie pobranych z serwera WWW (max. 30)
  A - średnia liczba punktów uzyskanych przez podstrony (max. 20.55)
  LR - wartość współczynnika korygującego
Wyjście:
  P - wartość w przedziale <0.00, 3.08>

Utwórz zmienną P ← 0

If A > 0:
  If T = True:
    P ← P + (0.0025 * C * (A / LR))
  If D = True:
    P ← P + (0.0025 * C * (A / LR))

Zwróć P
```

Pseudokod 38. Funkcja obliczająca punktację za unikalne znaczniki TITLE i META DESCRIPTION

Ostateczna liczba punktów za optymalizację kodu źródłowego jest wyznaczana na podstawie wyników uzyskanych na obydwu etapach obliczeń. Wynik z drugiego etapu dodatkowo jest korygowany o współczynnik *LR*.

Pseudokod 39. przedstawia funkcję obliczającą końcową punktację za optymalizację kodu źródłowego – *PK*, w przedziale od 0,00 do 20,00 punktów.

```
# <0.00, 20.00>
Wejście:
  HP - punktacja dla strony głównej
  AP - średnia wartość punktowa dla wszystkich podstron
  LR - wartość współczynnika korygującego
  UP - punktacja za unikalne znaczniki TITLE i META DESCRIPTION
Wyjście:
  PK - wartość w przedziale <0.00, 20.00>

Utwórz zmienną PK ← 0

If AP > 0:
  PK ← HP + (AP / LR)
elif:
  PK ← HP + (AP * LR)

PK ← PK + UP

If PK < 0:
```

```
  PK ← 0
elif PK > 20:
  PK ← 20

Zwróć PK
```

Pseudokod 39. Funkcja obliczająca punktację za optymalizację kodu źródłowego – *PK*

3.1.5. Punktacja za treść i strukturę tekstu

Na punktację za treść i strukturę tekstu – *PT*, wpływa rozmiar i formatowanie widocznego tekstu na stronie (wykorzystanie nagłówków, wyróżnień itp.). Brane są pod uwagę również testy czytelności, ułatwiające określenie stopnia trudności rozumienia danego tekstu.

Choć teoretycznie zakres punktacji za treść i strukturę tekstu mieści się w zakresie od −0,50 do 22,66 punktu, przyjęto regułę, że w razie uzyskania wartości ujemnej punktacja jest ustalana na 0 punktów, natomiast maksymalna możliwa wartość to 20 punktów. Przykładowo, jeśli strona WWW uzyska −0,45 lub 20,25 punktu, to wynik końcowy zostanie ustalony, odpowiednio, na 0 lub 20 punktów.

3.1.5.1. Punktacja dla podstron strony WWW

W pierwszej kolejności obliczana jest całkowita liczba liter i słów występujących w tekście na stronie WWW. Przyjęto założenie, że optymalna wielkość tekstu to co najmniej 600 liter ze znakami odstępu i 100 słów. Jeśli warunek ten zostanie spełniony, w kolejnych etapach do obliczenia punktacji zostanie wykorzystana wiedza na temat użytych znaczników formatujących tekst oraz testy czytelności.

Znaczniki H1–H6, pełniące funkcję nagłówków, powinny wskazywać, czego dotyczy dana sekcja lub akapit. Prawidłowa struktura tekstu powinna zawierać jeden znacznik H1 i co najmniej jeden znacznik H2. Znacznik H3 zazwyczaj jest stosowany wtedy, kiedy zagadnienia wyszczególnione przez znacznik H2 są na tyle złożone, że zasadne jest podzielenie ich na jeszcze mniejsze części [206]. Służy do tego znacznik P, który dzieli treść na pojedyncze bloki tekstowe – akapity, i jest najpowszechniej używanym znacznikiem opisującym strukturę tekstu na stronie WWW.

Dostępne w języku HTML znaczniki B i STRONG służą do pogrubienia tekstu, jednak ich znaczenie nie jest takie samo. Chcąc nadać większego znaczenia konkretnemu wyrażeniu, co dla wyszukiwarek jest dodatkową informacją ułatwiającą zrozumienie treści na stronie WWW, należy zastosować znacznik STRONG. Znacznik B nie nadaje

specjalnego znaczenia objętemu nim wyrażeniu, lecz jedynie formatuje jego wygląd przez pogrubienie czcionki.

Wyróżnić tekst na stronie WWW w formie pochylenia pozwala znacznik I. Nie ma on większego znaczenia dla botów wyszukiwarek, jednak dzięki niemu użytkownik strony WWW może zwrócić szczególną uwagę na wyróżniony w ten sposób fragment tekstu. Zmienne matematyczne i fragmenty kodu programistycznego można również wyróżnić przez pochylenie z użyciem znacznika VAR.

Zastosowanie znacznika LI, łącznie ze znacznikami OL i UL, umożliwia umieszczenie na stronie WWW list uporządkowanych, mających pewną numerację, oraz list nieuporządkowanych, w których kolejność umieszczanych po sobie elementów nie jest istotna. Znacznik LI wykorzystuje się do tworzenia układów nawigacyjnych, nazywanych „okruszkami chleba" (ang. breadcrumbs) [207], które w dużej mierze ułatwiają użytkowanie strony WWW.

Do oznaczenia skrótu lub akronimu wykorzystywany jest znacznik ACRONYM oraz wprowadzony w języku HTML5 znacznik ABBR. Znaczniki te pozwalają na oznaczenie w tekście formy skróconej, np. „prof." (*profesor*), bądź utworzenie wyrazu z pierwszych liter lub pierwszych zgłosek, najczęściej sylab, kilku wyrazów, będących zwykle jakąś nazwą, np. „ISOWQ" (*International Studies of Website Quality*) [208].

Składnia języka HTML zawiera znacznik CITE, umożliwiający odniesienie się w tekście do źródła, np. do tytułu książki, oraz znacznik Q, przeznaczony do umieszczenia w treści krótkich cytatów, które nie zawierają żadnych akapitów. Cytowanie dłuższych fragmentów tekstu, obejmujących kilka akapitów, wymaga stosowania znacznika BLOCKQUOTE.

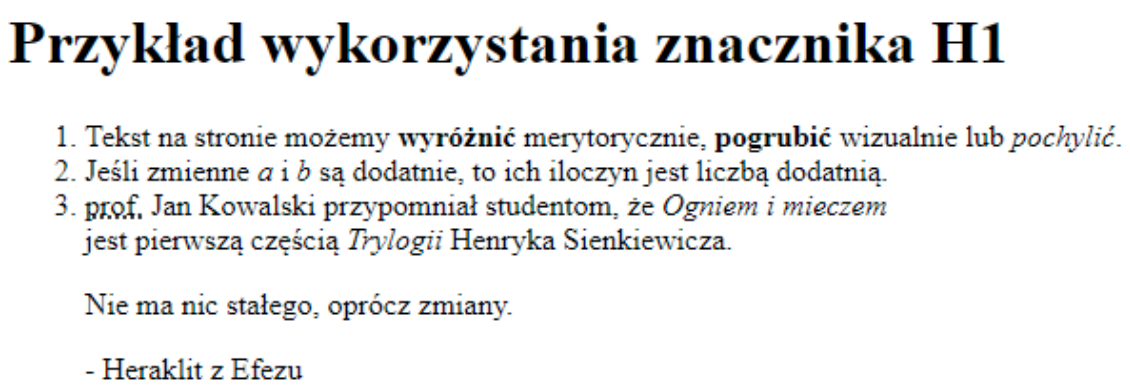
Na listingu 8. przedstawiono przykład wykorzystania znaczników formatujących tekst na stronie WWW.

```
<!DOCTYPE html>
<head>
<meta charset="utf-8">
</head>
<body>

<h1>Przykład wykorzystania znacznika H1</h1>

<ol>
  <li>
    Tekst na stronie możemy <strong>wyróżnić</strong> merytorycznie,
    <b>pogrubić</b> wizualnie lub <i>pochylić</i>.
  </li>
  <li>
```

```
    Jeśli zmienne <var>a</var> i <var>b</var> są dodatnie,
    to ich iloczyn jest liczbą dodatnią.
  </li>
  <li>
    <abbr title="profesor">prof.</abbr> Jan Kowalski przypomniał studentom,
    że <cite>Ogniem i mieczem</cite><br>jest pierwszą częścią
    <cite>Trylogii</cite> Henryka Sienkiewicza.
  </li>
</ol>

<blockquote>
  <p>Nie ma nic stałego, oprócz zmiany.</p>
  <footer>- Heraklit z Efezu</footer>
</blockquote>

</body>
</html>
```

Listing 8. Przykład wykorzystania znaczników formatujących tekst na stronie WWW

Wynik działania kodu HTML z listingu 8. jest przedstawiony na rysunku 10.



Rysunek 10. Rezultat działania kodu z listingu 8. w przeglądarce internetowej, opracowanie własne

Umieszczając treść na stronie WWW, należy kontrolować proporcję wielkości tekstu i wielkości całego kodu HTML. Strony o poprawnie skonstruowanej strukturze, stosujące różne formy wyróżniania informacji zawartych w treści, zazwyczaj są wysoko oceniane przez algorytmy wyszukiwarek.

Ważnym elementem oceny jakości tekstu prezentowanego na stronie WWW jest informacja o jego czytelności, czyli trudności zrozumienia zawartych w nim treści [209]. Taką ocenę, opartą na liczbie liter, słów i sylab, uzyskuje się dla konkretnego języka na podstawie testów czytelności [210]. Indeks czytelności Flescha oznacza, że im jest niższy, tym tekst jest trudniejszy do zrozumienia. Maksymalna wartość tego indeksu wynosi 120 i dotyczy tekstów najłatwiejszych, czyli takich, w których każde zdanie zawiera tylko dwa jednosylabowe słowa [211]. Indeks czytelności Gunninga-Foga ma na celu

określenie przystępności tekstu i wyraża liczbę lat edukacji potrzebnych do jego zrozumienia. Jeśli wartość tego indeksu wynosi 6, oznacza to, że tekst jest prosty i będzie zrozumiały dla uczniów szkoły podstawowej, natomiast wartość 16 oznacza, że tekst jest trudny i zrozumiały dla studentów studiów magisterskich [212].

Do oceny czytelności tekstu wykorzystuje się również indeksy takie jak Automated Readability Index, Coleman-Liau Index, SMOG Index, Dale-Chall Readability Formula czy Spache Readability Formula [213]. Testy czytelności oparte na sylabach odnoszą się głównie do języka angielskiego, dlatego przy obliczaniu punktacji odgrywają one rolę uzupełniającą.

Pseudokod 40. przedstawia funkcję obliczającą punktację za treść i strukturę tekstu, w przedziale od −0,50 do 20,60 punktu.

```
# <-0.50, 20.60>
Wejście:
  CW - liczba słów na stronie WWW
  CL - liczba liter na stronie WWW
  TG - tablica z liczbą wystąpień znaczników w kodzie HTML
  RT - tablica z punktacją za testy czytelności tekstu
  ST - liczba znaków w całym tekście
  SZ - liczba znaków w strefach tekstowych
Wyjście:
  P - wartość w przedziale <-0.50, 20.60>

Utwórz zmienne P ← 0 i T ← 0

# rozmiar tekstu <-0.50, 10.00>
T ← (CW / 100) * 600
If T > 0:
  T ← CL / T
elif:
  T ← 0
If T > 1:
  T ← 1
T ← (CW / 100) * 2 * T
If T > 10:
  T ← 10
P ← T

If CW >= 100:
  # znaczniki HTML <-0.20, 3.10>
  # H<1-6>
  If TG[H1] > 1: P ← P - 0.2
  If TG[H1] = 1: P ← P + 0.2
  If TG[H2] > 0: P ← P + 0.2
  If TG[H3] + TG[H4] + TG[H5] + TG[H6] > 1: P ← P + 0.2
  If TG[H3] + TG[H4] + TG[H5] + TG[H6] > 3: P ← P + 0.2

  # <li>
  If TG[LI] > 0: P ← P + 0.2
  If TG[LI] > 1: P ← P + 0.2
```

```
  # <b>, <strong>
  If TG[B] + TG[STRONG] > 0: P ← P + 0.2
  If TG[B] + TG[STRONG] > 1: P ← P + 0.2

  # <i>
  If TG[I] > 0: P ← P + 0.2
  If TG[I] > 1: P ← P + 0.2

  # <acronym>, <abbr>, <var>
  If TG[ACRONYM] > 0 or TG[ABBR] > 0: P ← P + 0.2
  If TG[ACRONYM] > 1 or TG[ABBR] > 1: P ← P + 0.2
  If TG[VAR] > 0: P ← P + 0.1

  # <q>, <blockquote>, <cite>
  If TG[Q] + TG[BLOCKQUOTE] + TG[CITE] > 0: P ← P + 0.2
  If TG[Q] + TG[BLOCKQUOTE] + TG[CITE] > 1: P ← P + 0.2

  # <p>
  If TG[P] > 1: P ← P + 0.1
  If TG[P] > 2: P ← P + 0.1

  # strefy tekstowe <0.00, 4.00>
  T = (SZ / ST) * 4.5
  If T > 4: T ← 4
  P = P + T

  # testy czytelności <0.00, 3.50>
  If RT[FKRE] >= 50 and RT[FKRE] >= 80: P ← P + 0.5
  If RT[FKGL] >= 5 and RT[FKGL] >= 12: P ← P + 0.75
  If RT[GFC] >= 5 and RT[GFC] >= 12: P ← P + 0.25
  If RT[CLI] >= 5 and RT[CLI] >= 14: P ← P + 0.5
  If RT[SI] >= 5 and RT[SI] >= 12: P ← P + 0.25
  If RT[DC] >= 5 and RT[DC] >= 12: P ← P + 0.25
  If RT[SPACHE] >= 2 and RT[SPACHE] >= 12: P ← P + 0.25
  If RT[ARI] >= 5 and RT[ARI] >= 12: P ← P + 0.75
elif:
  P ← P - 0.5

Zwróć P
```

Pseudokod 40. Funkcja obliczająca punktację za treść i strukturę tekstu

3.1.5.2. Punktacja końcowa

Ostateczna liczba punktów za treść i strukturę tekstu jest wyznaczana na podstawie wyniku uzyskanego dla strony WWW, skorygowanego o współczynnik *LR*. W przypadku gdy strona uzyska powyżej 4,55 punktu, wartość ta zostanie powiększona o 10%.

Pseudokod 41. przedstawia funkcję obliczającą końcową punktację za treść i strukturę tekstu – *PT*, w przedziale od 0,00 do 20,00 punktów.

```
# <0.00, 20.00>
Wejście:
  AP - średnia wartość punktowa dla wszystkich podstron
  LR - wartość współczynnika korygującego
```

```
Wyjście:
  PT - wartość w przedziale <0.00, 20.00>

Utwórz zmienną PT ← 0

If AP > 0:
  PT ← AP / LR

If PT > 4.55:
  PT ← PT * 1.1

If PT < 0:
  PT ← 0
elif PT > 20:
  PT ← 20

Zwróć PT
```

Pseudokod 41. Funkcja obliczająca końcową punktację za treść i strukturę tekstu – *PT*

3.1.6. Pseudokod algorytmu ISOWQ Rank

Zasadę działania algorytmu rankingowego ISOWQ Rank przedstawia pseudokod 42. Na wejściu przekazywane są informacje techniczne dla strony głównej, w tym dane dotyczące serwera WWW oraz analiza techniczna podstron. Na wyjściu otrzymywana jest liczba z dokładnością do dwóch miejsc po przecinku, o wartości w przedziale od 0,00 do 20,00.

W pierwszej kolejności obliczany jest współczynnik korygujący *LR* (pseudokod 1.), a następnie punktacja głównych czynników algorytmu – *PM*, *PK* i *PT*, odpowiednio za wykorzystane technologie i pozycje rankingowe (pseudokod 18.), za optymalizację kodu źródłowego (pseudokod 39.) oraz za treść i strukturę tekstu (pseudokod 41.).

```
Algorytm ISOWQ Rank

Wejście:
  P(S) - tablica z analizą techniczną strony głównej
  T(U) - tablica z analizą techniczną podstron
Wyjście:
  IR - wartość rankingowa w przedziale <0.00, 20.00>

Utwórz zmienne PM, PK, PT, IR
Wyzeruj PM, PK, PT, IR

# PM - punktacja za wykorzystane technologie i pozycje rankingowe
# PK - punktacja za optymalizację kodu źródłowego
# PT - punktacja za treść i strukturę tekstu
# IR - ISOWQ Rank - (PM + PK + PT) / 3

Utwórz zmienne HP, AP, LR, UP
Wyzeruj zmienne HP, AP, LR, UP

Utwórz tablicę V
Wyzeruj tablicę V
```

```
# HP - punktacja dla strony głównej
# AP - średnia wartość punktowa dla wszystkich podstron
# LR - wartość współczynnika korygującego
# UP - punktacja za unikalne znaczniki TITLE i META DESCRIPTION

LR ← oblicz_współczynnik_korygujący_LR(T)          # <1.00, 3.00>

#
# PM - punktacja za wykorzystane technologie i pozycje rankingowe
# Etap pierwszy - punktacja dla strony WWW jako całości
#

For each Si € P oblicz:
  przeanalizuj:
    wartości rankingowe MOZ DA i MOZ PA             # <0.00, 12.00>
    wartość rankingową Alexa Rank                   # <0.00, 10.00>
    liczbę hiperłączy zewnętrznych (MOZ EUID)       # <0.00, 5.00>
    wykorzystanie wtyczek społecznościowych         # <0.00, 1.00>
    liczbę poleceń na portalach społecznościowych   # <0.00, 2.00>
    liczbę znaków w nazwie domeny                   # <0.00, 2.00>
    wykorzystanie szyfrowania SSL                   # <0.00, 2.00>
    fizyczną lokalizację serwera WWW                # <0.00, 1.00>
    rejestrację serwera WWW w bazach DNSbl          # <-2.00, 0.00>
    adresy e-mail odnalezione w kodzie strony WWW   # <-1.00, 0.00>
  wynik wstaw do HP

#
# Etap drugi - punktacja dla podstron strony WWW
#

For each Ui € T oblicz punktację dla każdej krotki:
  przeanalizuj:
    wtyczki społecznościowe                         # <-0.25, 5.25>
    wykorzystanie narzędzi Google                   # <0.00, 2.75>
    publikowanie treści multimedialnych             # <0.00, 1.00>
    udostępnianie dokumentów biurowych              # <0.00, 1.50>
    komunikację przez komunikatory internetowe      # <0.00, 0.50>
    liczbę hiperłączy wychodzących                  # <0.00, 1.00>
  wynik wstaw do Vi

Wstaw do AP wartość średnią tablicy V
PM ← oblicz_punktację_PM(HP, AP, LR)

#
# PK - punktacja za optymalizację kodu źródłowego
# Etap pierwszy - punktacja dla strony WWW jako całości
#

Wyzeruj zmienne HP, AP
Wyzeruj tablicę V

For each Si € P oblicz:
  przeanalizuj:
    stosowanie mikroformatów                        # <0.00, 1.00>
    poprawność składni kodu HTML                    # <0.00, 0.40>
    liczbę adresów e-mail w kodzie HTML             # <0.00, 0.40>
  wynik wstaw do HP

#
```

```
# Etap drugi - punktacja dla podstron strony WWW
#

For each Ui € T oblicz punktację dla każdej krotki:
  przeanalizuj:
    wykryte słowa kluczowe                          # <-0.50, 4.00>
    wykorzystanie znacznika A                       # <-2.65, 3.40>
    wykorzystanie znacznika IMG                     # <-0.65, 2.25>
    wykorzystanie znaczników HTML                   # <-0.50, 1.65>
    wielkość tekstu zawartego w znacznikach P i A   # <0.00, 1.50>
    strefy tekstowe                                 # <0.00, 0.50>
    wykorzystanie znaczników TITLE i META           # <-1.10, 2.00>
    wersję języka HTML                              # <-0.25, 0.30>
    wielkość kodu HTML                              # <-0.35, 0.50>
    optymalizację wielkości kodu HTML               # <0.00, 2.00>
    optymalizację kodu w znacznikach STYLE i SCRIPT # <-0.60, 0.00>
    wykorzystanie technologii Flash                 # <-0.15, 0.10>
    poprawność składni kodu HTML                    # <-0.50, 0.50>
    konstrukcję adresu URL                          # <-0.15, 0.50>
    wykorzystanie znacznika META ROBOTS             # <0.00, 0.35>
    dodatkową optymalizację kodu HTML               # <-1.00, 1.00>
  wynik wstaw do Vi

Wstaw do AP wartość średnią tablicy V

For each Ui € T oblicz:
  przeanalizuj:
    unikalne znaczniki TITLE i META DESCRIPTION     # <0.00, 3.08>
  wynik wstaw do UP

PK ← oblicz_punktację_PK(HP, AP, LR, UP)

#
# PT - punktacja za treść i strukturę tekstu
# Punktacja dla podstron strony WWW
#

Wyzeruj zmienne AP
Wyzeruj tablicę V

For each Ui € T oblicz punktację dla każdej krotki:
  przeanalizuj:
    rozmiar tekstu                                  # <-0.50, 10.00>
    znaczniki HTML (nagłówki, listy, cytowania)     # <-0.20, 3.10>
    strefy tekstowe                                 # <0.00, 4.00>
    testy czytelności                               # <0.00, 3.50>
  wynik wstaw do Vi

Wstaw do AP wartość średnią tablicy V

PT ← oblicz_punktację_PT(AP, LR)

IR ← ((PM + PK + PT) / 3)
Zwróć IR
```

Pseudokod 42. Algorytm rankingowy ISOWQ Rank

3.1.7. Podsumowanie

Algorytm ISOWQ Rank pozwala ocenić jakość strony WWW na podstawie analizy kodu HTML i struktury tekstu, pozycji rankingowych, wykorzystanych technologii oraz parametrów technicznych serwera WWW. W odróżnieniu od algorytmów opartych na analizie struktury hiperłączy, wymagających do działania rozbudowanych zasobów technicznych (serwery, centra danych), algorytm ISOWQ Rank umożliwia szybkie obliczenie wartości rankingowej.

Algorytm ISOWQ Rank jest wykorzystywany przez system rankingowy ISOWQ do nadawania rankingu stronom WWW. System ten grupuje strony WWW w przedziałach co 5 punktów, odpowiednio: 0,00–4,99, 5,00–9,99, 10,00–14,99 i 15,00–20,00, wyróżniając je, począwszy od 5 punktów, dodatkowym oznaczeniem graficznym.

W praktyce wynik powyżej 5 punktów oznacza, że strona spełnia podstawowe wymogi techniczne i jest obecna w popularnych rankingach. Przykładowo serwis melex.com.pl[23] uzyskał 8,73 punktu ISOWQ Rank, co należy uznać za dobry wynik. Na podstawie przeprowadzonej analizy można zauważyć, że serwis ten do publikowania materiałów wideo wykorzystuje witrynę YouTube. Warto zaznaczyć, że jego kod jest zoptymalizowany na poziomie 87%, a atrybut TITLE występuje w 79% znaczników A (hiperłączy), co jest dobrą wartością. Niestety, serwis ten nie używa m.in. takich technologii jak wtyczki społecznościowe i szyfrowanie SSL, atrybut ALT występuje tylko w 25% znaczników IMG, a publikowane adresy e-mail nie są zakodowane.

Uzyskanie powyżej 10 punktów oznacza, że strona do komunikacji z otoczeniem stosuje techniki marketingowe, ma wiele odnośników przychodzących, zawiera dużo tekstu i ma wysoką pozycję w rankingach. Przykładem takiego serwisu jest alivia.org.pl[24], który uzyskał 14,21 punktu ISOWQ Rank, co jest bardzo dobrym wynikiem. Używa wtyczek społecznościowych, szyfrowania SSL i jest oparty na nowoczesnym systemie CMS (ang. Content Management System) – WordPress, który umożliwia publikację materiałów tekstowych i wideo zoptymalizowanych pod kątem wyszukiwarek internetowych. Zadbano również o marketing, o czym świadczy wykorzystanie witryny YouTube, duża liczba udostępnień w witrynie Facebook oraz wysoka pozycja w rankingach MOZ i Alexa Rank. Serwis zawiera na każdej podstronie dużo unikatowego tekstu. Optymalizacja kodu jest na poziomie 91%, a atrybut ALT występuje w 73% znaczników IMG, co jest dobrym

[23] ISOWQ, https://www.isowq.org/website/melex.com.pl/1439980/, maj 2022 r.
[24] ISOWQ, https://www.isowq.org/website/alivia.org.pl/1441262/, maj 2022 r.

wynikiem. Dodatkowa optymalizacja pod względem wykorzystania atrybutu TITLE w znacznikach A oraz zakodowanie adresów e-mail zwiększyłyby szansę na przekroczenie granicy 15 punków ISOWQ Rank.

Wyniki powyżej 15 punktów są zarezerwowane dla serwisów WWW, które wykorzystują dominujące technologie związane z web marketingiem, są dobrze zoptymalizowane według wytycznych SEO i znajdują się na wysokich miejscach w rankingach. W praktyce poziom 15 punktów jest przekraczany bardzo rzadko i udaje się to przeważnie dużym witrynom. Przykładem jest serwis lazienkaplus.pl[25], który uzyskał 15,39 punktu ISOWQ Rank. Używa on wtyczek społecznościowych, mikroformatów, ma wysokie pozycje w rankingach MOZ i Alexa Rank, ma dobrze zoptymalizowany kod (93%), wykorzystanie atrybutu TITLE w znaczniku A osiąga 91%, a atrybutu ALT w znaczniku IMG – 93%.

3.2. Implementacja algorytmu ISOWQ Rank

Do wdrożenia algorytmu rankingowego ISOWQ Rank wykorzystano skryptowy język programowania PHP, stosowany powszechnie do budowy aplikacji webowych oraz do generowania witryn internetowych po stronie serwera WWW. Język PHP jest intensywnie rozwijany od ponad 25 lat i dziś jest liderem (prawie 80%[26] udziału) wśród języków wykorzystywanych do tworzenia oprogramowania dla stron WWW. Obecnie oferuje wiele możliwości, takich jak kompilowanie programu do postaci kodu maszynowego przed jego wykonaniem, dostęp do silników baz danych – MySQL, SQLite, PostgreSQL i Oracle, obsługa formatu XML, programowanie współbieżne czy zaawansowane funkcje kryptograficzne [214]. Programista języka PHP może jednocześnie aktualizować zarówno część związaną ze stroną internetową systemu, jak i część odpowiedzialną za poprawne funkcjonowanie robotów internetowych.

System ISOWQ zawiera prawie tysiąc plików z kodem źródłowym w języku PHP, obejmujących kod robotów internetowych, systemu analitycznego, strony WWW i panelu administracyjnego oraz bibliotek programistycznych. Na listingu 9. przedstawiono fragment struktury plików w systemie ISOWQ.

```
ISOWQ\INC
├─class
│    allowCrawl.class.php         # analiza meta i robots.txt
│    cache.class.php              # system pamięci podręcznej
│    chip_download.class.php      # obsługa pobrań plików z audytami
```

[25] ISOWQ, https://www.isowq.org/website/lazienkaplus.pl/1322181/, maj 2022 r.
[26] W3Tech, https://w3techs.com/technologies/details/pl-php, maj 2022 r.

```
|    crawler.class.php            # lista zadań dla botów ISOWQ
|    daemon.class.php             # obsługa współbieżności
|    decodeHTMLlinks.class.php    # analiza hiperłączy
|    decodeHTMLtech.class.php     # analiza techniczna
|    decodeHTMLtext.class.php     # analiza treści i struktury tekstu
|    members.class.php            # obsługa użytkowników systemu
|    mysqli.class.php             # obsługa bazy danych
|    pdf.class.php                # obsługa plików w formacie PDF
|    points.class.php             # obliczanie rankingu ISOWQ Rank
|    portal.class.php             # obsługa strony WWW systemu ISOWQ
|    process.class.php            # obsługa procesów systemowych
|    proxy.class.php              # obsługa serwerów Proxy
|    report.class.php             # prezentacja audytu na stronie WWW
|    reportTLD.class.php          # prezentacja zbiorcza dla ccTLD
|    sitemap.class.php            # mapa strony dla wyszukiwarek
|    social.class.php             # obsługa mediów społecznościowych
|    sqliteAnalysis.class.php     # obsługa baz danych SQLite
|    sqliteTechnical.class.php    # obsługa bazy technicznej SQLite
|    toUTF8.class.php             # konwersja treści na format UTF8
|
└─daemon
     analyzer.php                 # obsługa botów ISOWQ
     bridge.php                   # obsługa kolejki botów ISOWQ
```

Listing 9. Fragment struktury plików systemu ISOWQ w języku PHP

Konfigurację interpretera PHP odpowiednio zmodyfikowano, aby kod botów mógł obsługiwać wiele wątków i procesów operujących na współdzielonych danych. Lista bibliotek programistycznych aktywowanych w konfiguracji interpretera języka PHP jest przedstawiona na listingu 10.

```
[PHP Modules]
bz2          # obsługa archiwum w formacie bzip2
calendar     # obsługa dat i kalendarzy
Core         # podstawowe funkcje języka PHP
ctype        # weryfikacja łańcuchów zgodnych z ustawieniami narodowymi
curl         # protokoły komunikacji z serwerami
date         # obsługa daty i czasu
dom          # obsługa drzewa DOM
exif         # metainformacje zawarte w plikach
FFI          # obsługa kodu C++ wewnątrz kodu PHP
fileinfo     # informacja o typach plików
filter       # filtry danych
ftp          # obsługa protokołu FTP
gd           # funkcje graficzne
gettext      # obsługa wielu języków w kodzie PHP
gmp          # biblioteka matematyczna
hash         # funkcje haszujące
iconv        # konwersja tekstu w różnych systemach kodowania
igbinary     # serializacja danych
imagick      # biblioteka graficzna ImageMagick
intl         # formatowanie tekstu zgodne z ustawieniami narodowymi
json         # obsługa formatu JSON
libxml       # obsługa formatu XML
mbstring     # obsługa tekstu w formacie UTF-8
```

```
memcached    # obsługa bazy dla danych tymczasowych
msgpack      # serializacja obiektów
mysqli       # obsługa bazy danych MySQL
mysqlnd      # natywna obsługa bazy danych MySQL
openssl      # obsługa algorytmów szyfrujących
pcntl        # obsługa programowania współbieżnego
pcre         # biblioteka PCRE
PDO          # standard komunikacji z bazami danych
pdo_mysql    # PDO dla bazy danych MySQL
pdo_sqlite   # PDO dla bazy danych SQLite
Phar         # obsługa archiwum w formacie PHAR
posix        # funkcje standardu POSIX
readline     # obsługa linii poleceń
Reflection   # API do obsługi interfejsów, klas i metod
session      # obsługa sesji
shmop        # obsługa pamięci współdzielonej
SimpleXML    # obsługa formatu XML
sockets      # obsługa gniazd
sodium       # biblioteka kryptograficzna
SPL          # obsługa biblioteki SPL
sqlite3      # obsługa baz danych SQLite
standard     # funkcje języka PHP
sysvmsg      # obsługa wiadomości (System V)
sysvsem      # obsługa semaforów (System V)
sysvshm      # obsługa pamięci współdzielonej (System V)
tidy         # funkcje analizujące składnię kodu HTML
tokenizer    # obsługa interfejsu tokenizera wbudowanego w silnik Zend
xml          # obsługa formatu XML
xmlreader    # funkcje odczytu danych w formacie XML
xmlwriter    # funkcje zapisu danych w formacie XML
xsl          # obsługa formatu XSL
Zend OPcache # akcelerator kodu PHP
zip          # obsługa archiwum w formacie ZIP
zlib         # obsługa biblioteki zlib
```

Listing 10. Biblioteki programistyczne aktywowane w interpreterze języka PHP

Podczas projektowania systemu i implementacji algorytmu ISOWQ Rank dużym wyzwaniem programistycznym było opracowanie metod umożliwiających detekcję w kodzie strony WWW stref tekstowych i wykorzystanych technologii. W początkowej fazie budowy systemu ISOWQ analizowano kod HTML losowo wybranych stron WWW, aby nauczyć boty poprawnie interpretować ewentualne błędy występujące w jego składni. Biorąc pod uwagę liczbę stron WWW dostępnych w sieci internet, detekcja większości wariantów wykorzystania na nich konkretnych technologii i metod prezentowania treści jest bardzo trudna. Próbami rozwiązania problemów związanych z ekstrakcją danych z dokumentów hipertekstowych są metody analizy kodu HTML umożliwiające pozyskanie wiedzy zawartej w jego strukturze [215], detekcję struktur tekstowych [216] i wtyczek w systemach CMS [217], wyodrębnienie tytułów [218], obrazów [219] i struktur logicznych z tabel [220], klasyfikację znaczników HTML [221], fragmentację [222],

filtrowanie [223], klasyfikację [224] i kategoryzację [225] treści oraz analizę wizualną strony [226].

W systemie ISOWQ ekstrakcja tekstu ze strony WWW odbywa się przez przekształcenie kodu HTML w obiekt drzewa DOM (ang. Document Object Model). W tak powstałym obiekcie wyszukiwane są elementy nagłówków – znaczniki H1–H6, list – OL, UL i LI, tabel – TR, TH i TD, akapitów – P, i znaczniki DIV, definiujące działy lub sekcje w dokumencie HTML. Następnie w wyszukanym ciągu znaków usuwane są nadmiarowe znaki odstępu, tabulacji i przejścia do nowej linii. Zastosowanie tej metody nie wpływa znacząco na wykorzystanie zasobów obliczeniowych systemu i okazało się optymalne dla większości stron WWW.

System ISOWQ pobiera dane z zewnętrznych baz za pośrednictwem interfejsu API (ang. Application Programming Interface), np. z serwisu MOZ, oraz analizuje dane publikowane na ogólnodostępnych stronach WWW. O ile dostęp do danych za pomocą interfejsu API nie jest obarczony wysokim ryzykiem otrzymania błędnych wyników, to ekstrakcja danych z kodu HTML, którego konstrukcja może się w każdej chwili zmienić, może zwrócić niepoprawny wynik. Pomimo okresowej aktualizacji kodu systemu ISOWQ nie można w całości wyeliminować tego ryzyka.

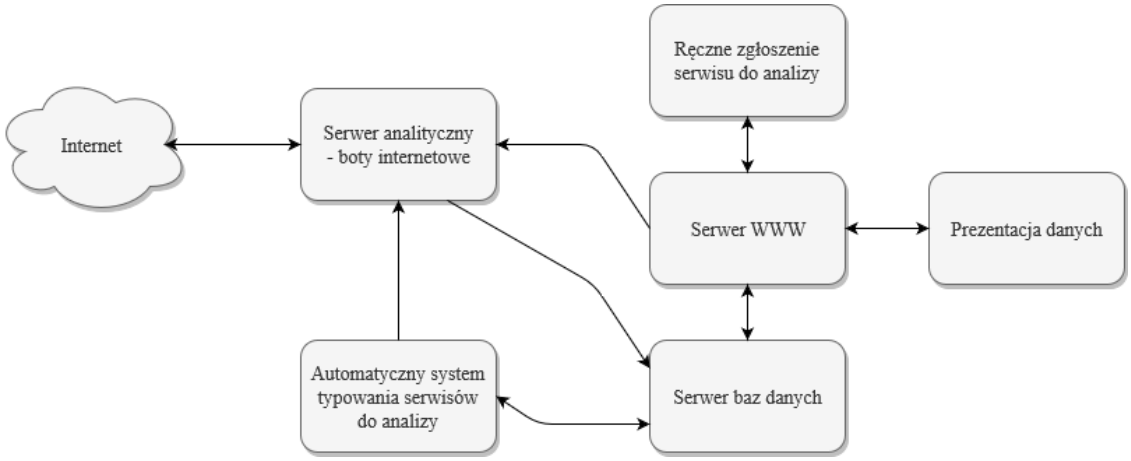
3.3. Architektura systemu rankingowego ISOWQ

W latach 2010–2011 zaprojektowano i wdrożono system rankingowy ISOWQ, jako propozycję nowego narzędzia do technicznej analizy stron internetowych. Po 10 latach funkcjonowania systemu baza danych zawiera szczegółowe analizy ponad 1,3 mln stron WWW, co daje ponad 26 mln adresów URL.

W pierwszych latach działania systemu ISOWQ analizowane serwisy internetowe były utrzymywane pod domenami narodowymi należącymi do krajów Unii Europejskiej, krajów kandydujących, Rosji i innych członków Wspólnoty Niepodległych Państw oraz Stanów Zjednoczonych, a także pod europejską domeną .eu. Dzisiaj głównym zadaniem systemu jest analiza serwisów WWW działających pod wszystkimi (243) domenami narodowymi (ccTLD), jak również publiczne udostępnianie raportów z tych analiz na stronie WWW. Dane udostępniane przez system ISOWQ można w dowolny sposób wykorzystywać w innych systemach informatycznych czy poddawać dalszym analizom.

3.3.1. Budowa systemu

System ISOWQ składa się z dwóch niezależnych segmentów. W skład pierwszego segmentu wchodzi podsystem odpowiedzialny za automatyczne typowanie i analizowanie kolejnych serwisów WWW, a zadaniem drugiego jest prezentacja danych i obsługa użytkownika systemu. Użytkownik ma możliwość dodania do systemu dodatkowych adresów stron WWW i zlecenia analizy. Oba segmenty mają bezpośredni dostęp do serwera baz danych. Architekturę systemu ISOWQ przedstawia rysunek 11.



Rysunek 11. Architektura systemu ISOWQ, opracowanie własne

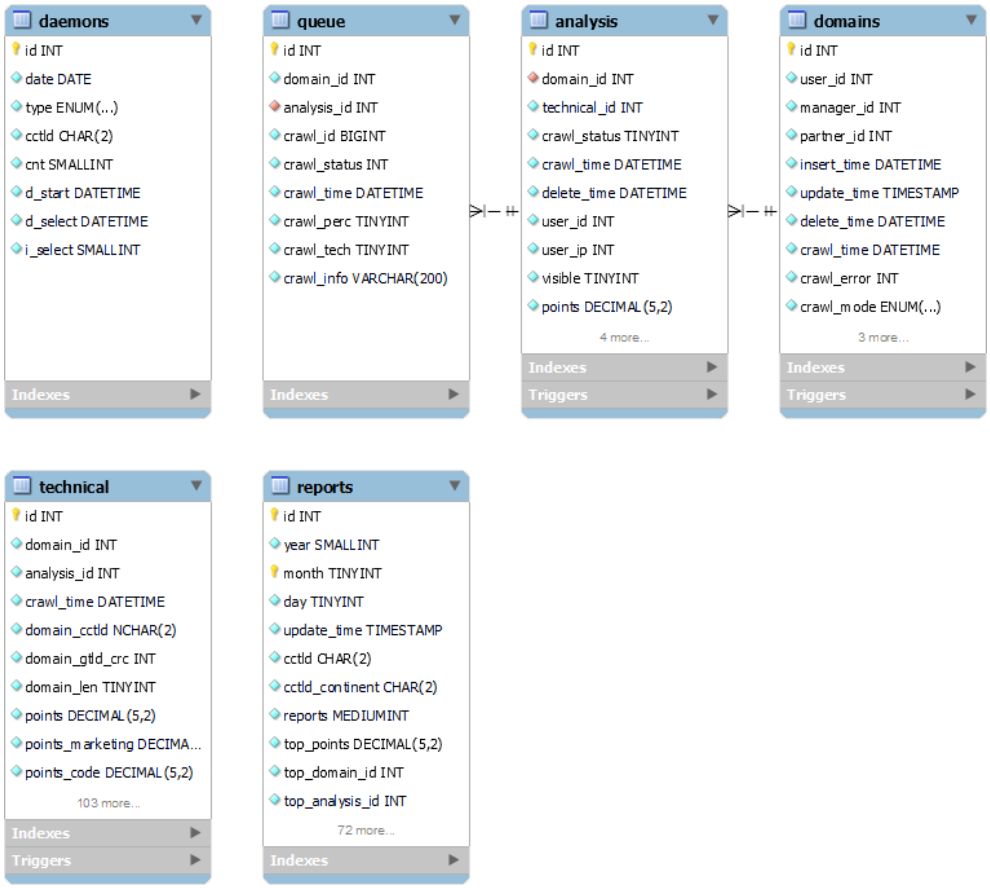
Do budowy systemu wykorzystano dwuprocesorowe serwery IBM eServer x345 i Dell PowerEdge 1950 z pamięcią RAM w rozmiarze, odpowiednio, 8 i 16 GB oraz nośnikami danych o pojemności, odpowiednio, 350 GB oraz 1 i 2 TB, które ze względów bezpieczeństwa skonfigurowano w macierzy RAID. Na serwerach zainstalowano systemy OpenBSD, FreeBSD i Ubuntu z oprogramowaniem baz danych MySQL i SQLite oraz serwerem WWW Apache z obsługą języka PHP.

3.3.2. Struktura bazy danych

3.3.2.1. MySQL

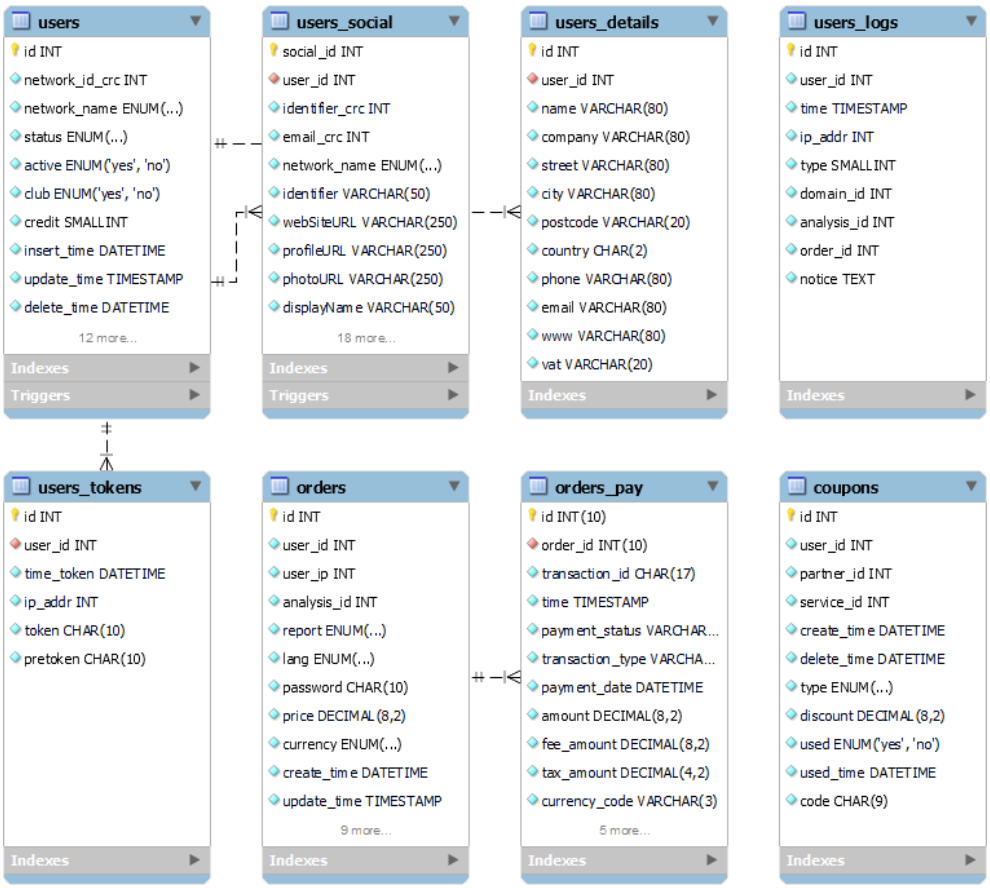
Strukturę dla relacyjnej bazy danych MySQL zaprojektowano w narzędziu MySQL Workbench firmy Oracle. Do składowania danych analitycznych zastosowano silnik InnoDB, obsługujący transakcje i stosowanie kluczy obcych, oraz silnik MyISAM, przeznaczony do obsługi raportów i danych archiwalnych. Rysunek 12. przedstawia strukturę

tabel, w których przechowywane są dane analizowanych stron WWW i raportów zbiorczych.



Rysunek 12. Struktura bazy danych SQL do obsługi botów systemu ISOWQ, opracowanie własne

Na rysunku 13. przedstawiona jest struktura tabel z danymi użytkowników. Wszystkie operacje wykonywane przez użytkowników systemu są rejestrowane. Informacje zawarte w tabelach bazy MySQL zajmują na nośnikach danych serwera 1,6 GB. Początkowo do składowania danych archiwalnych i kodów HTML stron WWW wykorzystywano silnik Archive, umożliwiający kompresowanie danych. Jednak z powodu znacznego przyrostu wielkości bazy MySQL podjęto decyzję o zmianie metody gromadzenia danych na format plików SQLite, co poprawiło wydajność systemu.



Rysunek 13. Struktura bazy danych SQL do obsługi użytkowników systemu ISOWQ, opracowanie własne

3.3.2.2. SQLite

W formacie SQLite zapisywane są informacje o każdej przeprowadzonej analizie strony WWW. Aby zwiększyć wydajność systemu, przyjęto regułę, że dane będą zapisywane w osobnych plikach dla każdej domeny internetowej. W systemie utworzono ponad 950 tys. plików w formacie SQLite, zajmujących na nośnikach danych serwera ponad 243 GB. Decyzja o wykorzystaniu formatu SQLite okazała się optymalna. Dane dotyczące konkretnej strony WWW mogą być łatwo przeniesione, odczytane przez dowolnego klienta SQLite i poddane ponownej analizie.

3.3.3. Analiza zgromadzonych danych

Na podstawie zgromadzonych danych można generować wiele zestawień, np. dotyczących wykorzystania na stronach internetowych multimediów, użycia wtyczek społecznościowych czy wersji języka znaczników HTML w konkretnych grupach domen ccTLD. Tabela 5. przedstawia takie zestawienie dla europejskich domen narodowych i terytoriów zależnych za 2016 rok. Oznaczenie kolumn w tabeli jest następujące: *ccTLD* – dwuliterowa nazwa domeny narodowej, *AVG* – średnia z kolumn *HTML5*, *YouTube* i *SM* (Social

Media) wyrażona w procentach, *Analizy* – liczba wykonanych analiz stron WWW w danej domenie narodowej.

Tabela 5. Ranking domen narodowych (ccTLD), Europa, 2016 rok

Miejsce	ccTLD	AVG	HTML5	YouTube	SM	Analizy
1	gg (Guernsey)	51,18	83,60	19,22	50,70	98
2	no (Norwegia)	45,22	78,19	19,29	38,19	1375
3	ba (Bośnia i H.)	39,86	77,49	33,57	8,20	170
4	mc (Monaco)	39,75	77,49	33,57	8,20	69
5	al (Albania)	39,12	65,29	25,37	26,71	153
20	fr (Francja)	25,09	51,32	9,16	14,80	2404
33	pl (Polska)	21,67	39,70	8,58	16,73	2391
34	lv (Łotwa)	21,57	39,62	11,06	14,04	1456
48	de (Niemcy)	18,89	36,69	10,17	9,82	2812

Z powyższej tabeli wynika, że serwisy internetowe w domenie .gg, które zostały przeanalizowane w 2016 roku, wykorzystywały język znaczników HTML5 w prawie 84%, a serwisy w domenie .pl – tylko w 22%. Tabela 6. przedstawia wyniki dla analiz przeprowadzonych w 2021 roku. Oznaczenie kolumn jest identyczne jak w tabeli 5.

Tabela 6. Ranking domen narodowych (ccTLD), Europa, 2021 rok

Miejsce	ccTLD	AVG	HTML5	YouTube	SM	Analizy
1	va (Watykan)	66,75	75,83	64,15	60,26	48
2	es (Hiszpania)	59,19	95,42	35,43	46,73	63
3	si (Słowenia)	56,39	90,23	34,62	44,32	63
4	li (Lichtenstein)	55,29	98,33	19,97	47,57	63
5	gi (Gibraltar)	54,68	92,74	26,99	44,31	63
13	lv (Łotwa)	51,49	88,57	26,84	39,06	64
16	pl (Polska)	49,61	91,99	17,45	39,38	64
48	de (Niemcy)	37,84	76,38	16,23	20,91	63
52	sm (San Marino)	30,14	55,69	15,37	19,37	62

Z powyższej tabeli wynika, że serwisy internetowe w domenie .li, które zostały przeanalizowane w 2021 roku, wykorzystywały język znaczników HTML5 w ponad 98%, a serwisy w domenie .pl – w prawie 92%.

Następnym zestawieniem zawartym w tabeli 7. jest procentowy udział niezakodowanych adresów e-mail w kodzie strony oraz procent rejestracji adresów IP serwerów hostujących, jako nośników spamu, w bazach DNSbl.

Tabela 7. Ranking DNSbl i niezakodowanych adresów e-mail dla domen narodowych (ccTLD), 2016 rok

Miejsce	ccTLD	DNSbl	E-mail	Analizy
1	tn (Tunezja)	81,28	57,56	144
2	et (Etiopia)	75,86	52,01	91
3	ne (Niger)	70,16	39,05	30
4	bt (Bhutan)	63,16	66,39	137
5	sn (Senegal)	53,69	53,11	140
122	pl (Polska)	9,89	55,79	2391
131	fr (Francja)	8,96	39,70	2404
169	lv (Łotwa)	6,87	60,63	1456
173	de (Niemcy)	6,75	60,53	2812

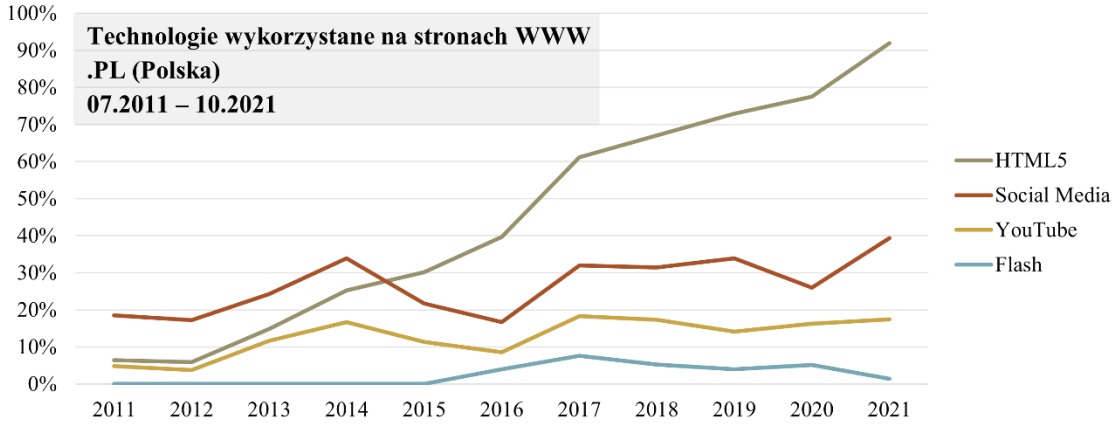
Z powyższego zestawienia wynika, że serwery zaklasyfikowane jako nośnik spamu działają głównie w krajach afrykańskich, natomiast problem niezakodowanych adresów e-mail na stronach WWW ma wymiar globalny. Tabela 8. przedstawia wyniki analiz przeprowadzonych w 2021 roku.

Tabela 8. Ranking DNSbl dla domen narodowych (ccTLD), 2021 rok

Miejsce	ccTLD	DNSbl	E-mail	Analizy
1	et (Etiopia)	60,99	69,44	61
2	dz (Algieria)	56,70	77,71	63
3	rw (Rwanda)	48,94	70,30	63
4	tn (Tunezja)	47,36	62,84	62
5	bo (Boliwia)	47,16	63,50	63
119	pl (Polska)	9,88	72,13	64
127	lv (Łotwa)	9,12	76,54	64
151	uk (Wielka Brytania)	7,76	58,47	62
242	tf (Fr. Terytoria Pd. i Antarktyczne)	0,00	14,72	42

Z powyższego zestawienia wynika, podobnie jak z danych z 2016 roku, że serwery zaklasyfikowane jako nośnik spamu działają głównie w krajach afrykańskich, pomimo odnotowanego w 2021 roku procentowego spadku.

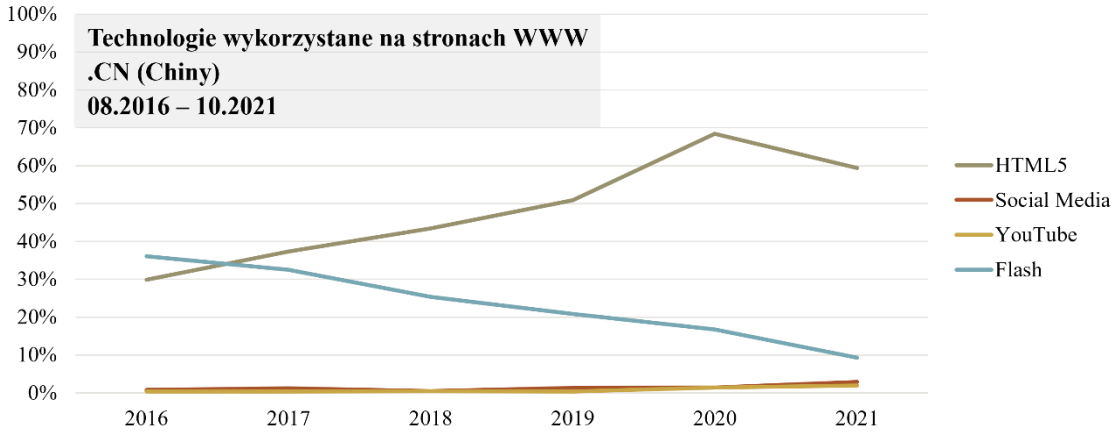
Na podstawie zgromadzonych danych można odpowiedzieć na pytanie, jak zmieniał się w czasie udział wykorzystania konkretnych technologii na stronach WWW. Analiza stron w domenie .pl dotycząca stosowania języka znaczników HTML5, wtyczek społecznościowych, serwisu YouTube czy technologii Flash jest przedstawiona na rysunku 14.



Rysunek 14. Udział technologii w analizowanych serwisach WWW w domenie .pl w okresie od 07.2011 do 10.2021, na podstawie danych uzyskanych z systemu ISOWQ, opracowanie własne

Z powyższego rysunku wynika, że na stronach internetowych w domenie .pl udział języka HTML5 w porównaniu z pozostałymi wersjami języka (HTML 3, HTML 4, XHTML itp.) z roku na rok rośnie, co oznacza, że jest on coraz częściej stosowany do budowy nowych lub aktualizacji istniejących stron WWW. Wykorzystanie wtyczek społecznościowych i serwisu YouTube wynosi, odpowiednio, 40% i prawie 20%, co można uznać za dobry wynik. Spadek użycia technologii Flash, niemalże do zera, jest spowodowany zakończeniem jej obsługi prawie we wszystkich przeglądarkach internetowych na początku 2021 roku.

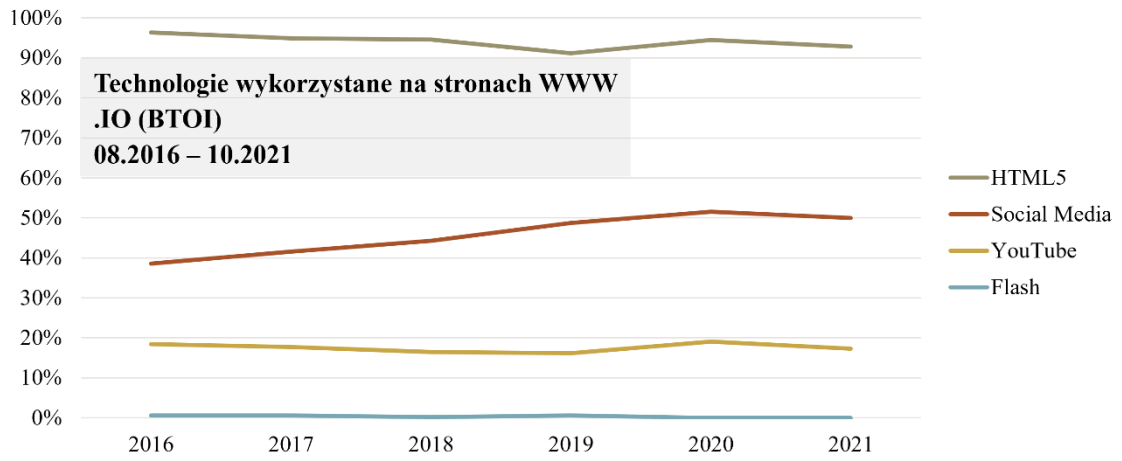
Kolejny przykład, przedstawiony na rysunku 15., dotyczy użycia technologii na stronach WWW w domenie .cn. Wykorzystanie języka znaczników HTML5 rośnie z roku na rok, jednak wykorzystanie wtyczek społecznościowych i serwisu YouTube jest na pograniczu błędu statystycznego. Sytuacja ta może być spowodowana blokadą portali społecznościowych przez ośrodki władzy w Chinach lub brakiem zainteresowania tymi mediami ze strony tamtejszych użytkowników. Stosowanie technologii Flash sukcesywnie zanika, podobnie jak w przypadku serwisów WWW w domenie .pl.



Rysunek 15. Udział technologii w analizowanych serwisach WWW w domenie .cn w okresie od 08.2016 do 10.2021, na podstawie danych uzyskanych z systemu ISOWQ, opracowanie własne

Interesujące wyniki dotyczą wykorzystania technologii na stronach WWW w domenie .io, należącej do Brytyjskiego Terytorium Oceanu Indyjskiego. W odróżnieniu od domen .pl i .cn, które zazwyczaj są rejestrowane przez przedsiębiorstwa działające na terytorium, odpowiednio, Polski i Chin, domenę .io wykorzystują m.in. firmy zajmujące się nowymi technologiami, działające na wielu rynkach, takie jak Onion Corporation (onion.io) czy Blynk (blynk.io).

Rysunek 16. przedstawia użycie technologii na stronach internetowych w domenie .io. Ponad 90% przeanalizowanych stron WWW stosuje znaczniki języka HTML5, 50% korzysta z wtyczek społecznościowych, a prawie 20% prezentuje treści multimedialne za pomocą portalu YouTube, co można uznać za bardzo dobry wynik. Wykorzystanie technologii Flash jest na granicy błędu statystycznego.



Rysunek 16. Udział technologii w analizowanych serwisach WWW w domenie .io w okresie od 08.2016 do 10.2021 na podstawie danych uzyskanych z systemu ISOWQ, opracowanie własne

System ISOWQ umożliwia szczegółową analizę danych zebranych przez ponad 10 lat jego działania. Dane te oprócz tego, że zapewniają informacje na temat wykorzystanych technologii, pozwalają ustalić, które znaczniki języka HTML są najpowszechniej stosowane, jakie biblioteki programistyczne w języku JavaScript stosowane są najczęściej, a także jak szybko następuje aktualizacja języka PHP na serwerach WWW.

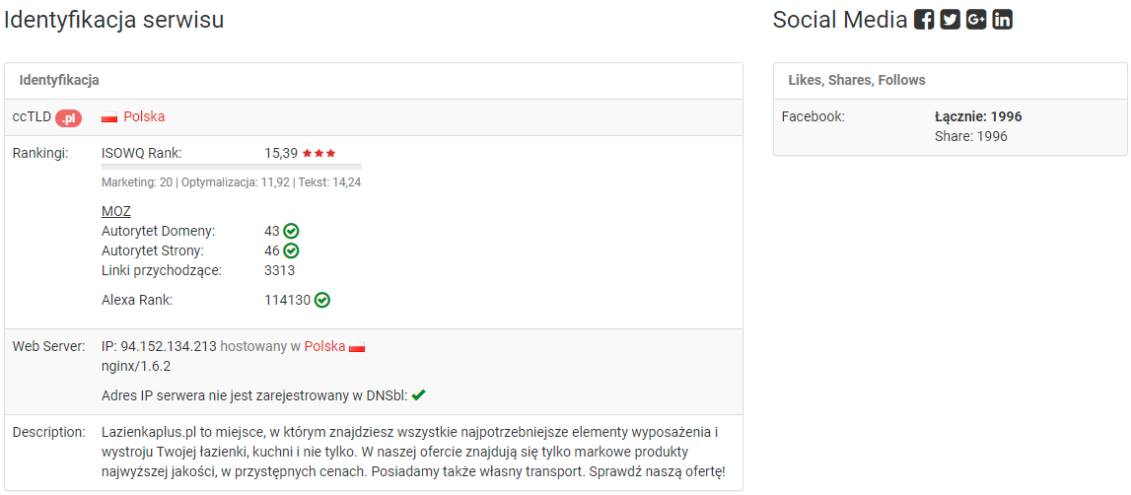
3.4. Elementy analizy technicznej stron internetowych

Na stronie internetowej systemu ISOWQ dostępne są techniczne analizy serwisów WWW przeprowadzone od 2011 roku. Dane te są, w formie raportów, prezentowane zbiorczo dla całej strony WWW i osobno dla każdej podstrony. W niniejszym rozdziale

przedstawiono analizę techniczną[27] wykonaną 28 sierpnia 2018 roku dla strony utrzymywanej pod adresem www.lazienkaplus.pl. Strona uzyskała 15,39 punktu w rankingu ISOWQ Rank, co jest wynikiem bardzo dobrym i świadczy o wysokich pozycjach w rankingach MOZ i Alexa oraz dobrej optymalizacji kodu HTML i treści.

3.4.1. Informacje zbiorcze dla całego serwisu WWW

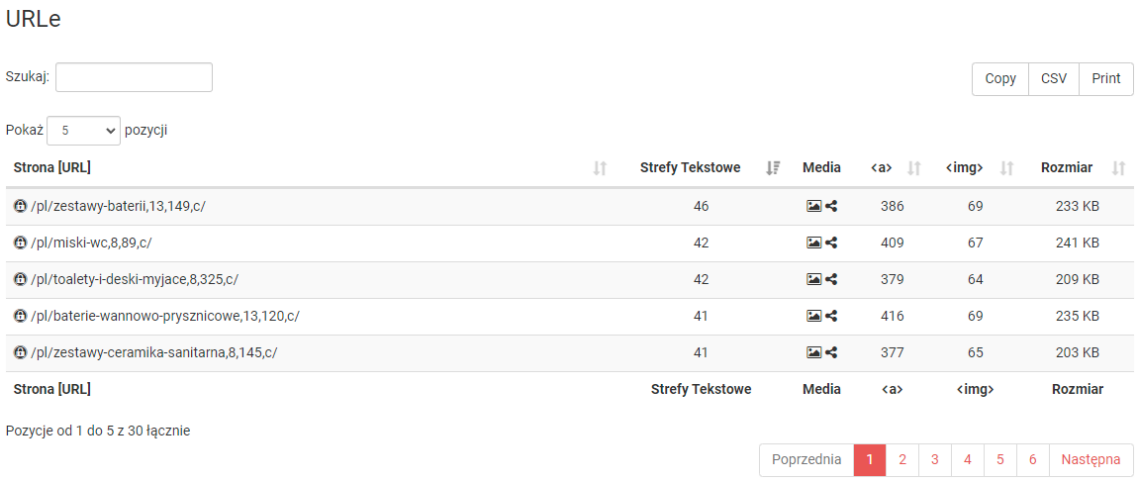
Pierwsza część raportu (rysunek 17.) zawiera informacje o lokalizacji i parametrach technicznych serwera WWW, uzyskanej punktacji w rankingach MOZ i Alexa, liczbie hiperłączy przychodzących oraz zawartości znacznika META DESCRIPTION na stronie głównej. Uzupełnieniem są dane na temat liczby poleceń w mediach społecznościowych.



Rysunek 17. Informacja o stronie WWW i liczbie poleceń w mediach społecznościowych, opracowanie własne

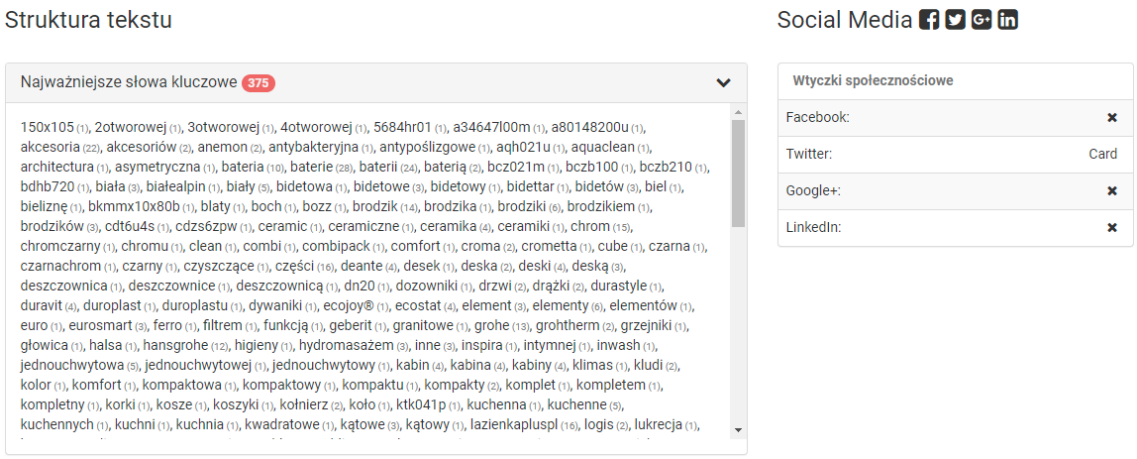
W następnej części raportu znajduje się lista wszystkich adresów URL, które zostały poprawnie pobrane przez boty. W przypadku wykrycia błędnych adresów zostaną one ujęte na osobnej liście. Lista (rysunek 18.) zawiera dodatkowe informacje o każdej stronie: liczbę stref tekstowych, znaczników A i IMG oraz rozmiar kodu HTML w kilobajtach.

[27] ISOWQ, https://www.isowq.org/website/lazienkaplus.pl/1322181/, maj 2022 r.



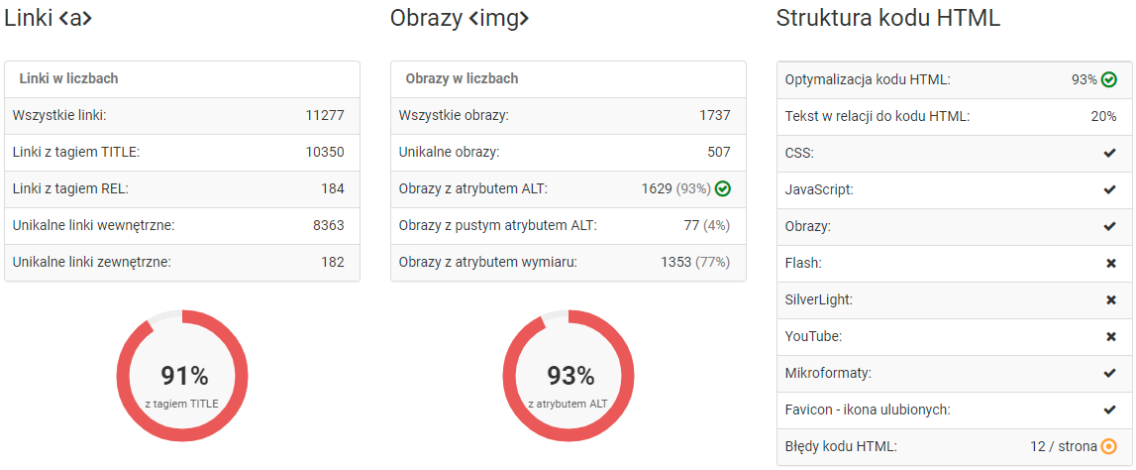
Rysunek 18. Lista adresów URL pobranych przez boty systemu ISOWQ, opracowanie własne

Pod listą pobranych adresów URL prezentowane są dane na temat najważniejszych słów kluczowych występujących na wszystkich stronach serwisu WWW, a także dane dotyczące wykorzystania wtyczek społecznościowych (rysunek 19.).



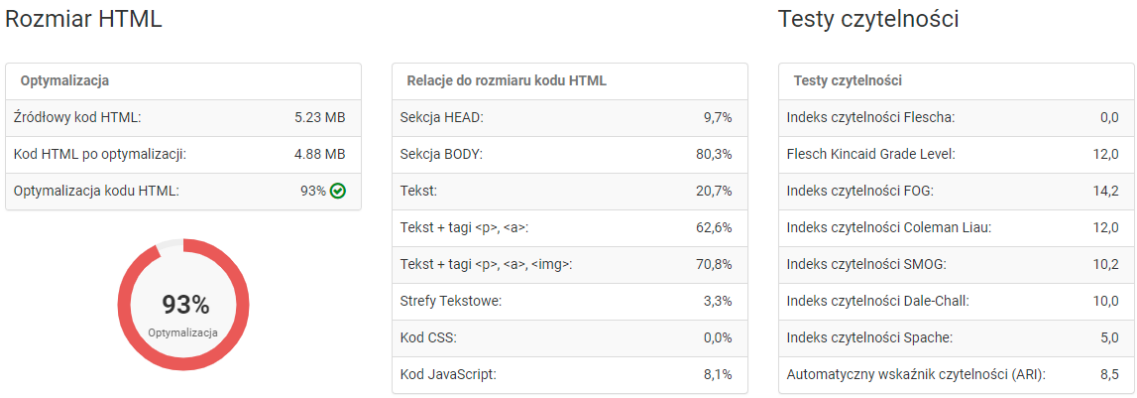
Rysunek 19. Lista słów kluczowych i wykorzystanych wtyczek społecznościowych, opracowanie własne

Kolejna część raportu (rysunek 20.) zawiera dane statystyczne na temat optymalizacji znaczników A i IMG oraz kodu HTML. Dodatkowo w tej części przedstawione są informacje o wykorzystaniu stylów CSS i języka JavaScript, mikroformatów, technologii YouTube, Silverlight i Flash, a także o użyciu ikony favicon i średniej liczbie błędów składni w kodzie HTML.



Rysunek 20. Informacje o optymalizacji znaczników A i IMG oraz kodu HTML, opracowanie własne

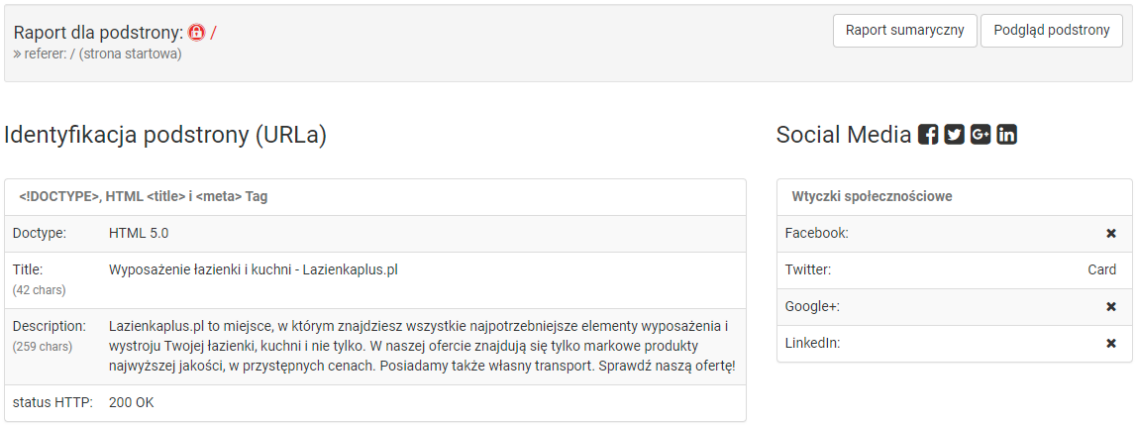
W ostatniej części raportu znajdują się informacje na temat rozmiaru kodu HTML, jego optymalizacji i względnych rozmiarów poszczególnych sekcji kodu (rysunek 21.). Uzupełnieniem są dane na temat testów czytelności.



Rysunek 21. Informacje o rozmiarze kodu HTML i wynikach testów czytelności, opracowanie własne

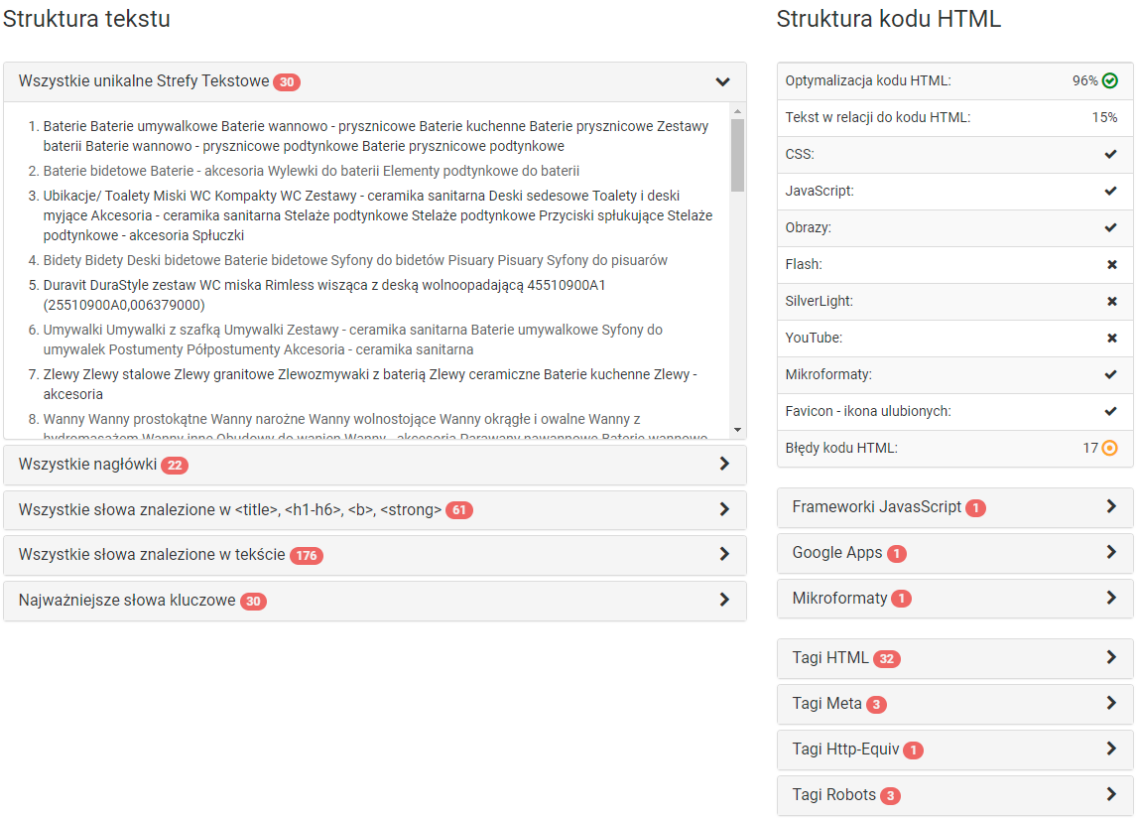
3.4.2. Szczegółowe informacje dla podstrony

Dostęp do raportu technicznego dla każdej podstrony jest możliwy bezpośrednio z listy adresów URL (rysunek 18.). Na rysunku 22. przedstawiono wyniki analizy technicznej dla strony głównej serwisu www.lazienkaplus.pl. W tej części raportu prezentowane są dane na temat zastosowanego języka znaczników HTML, zawartości znaczników META oraz rodzaju wykorzystanej wtyczki społecznościowej.



Rysunek 22. Informacje o znacznikach i wykorzystanych wtyczkach społecznościowych, opracowanie własne

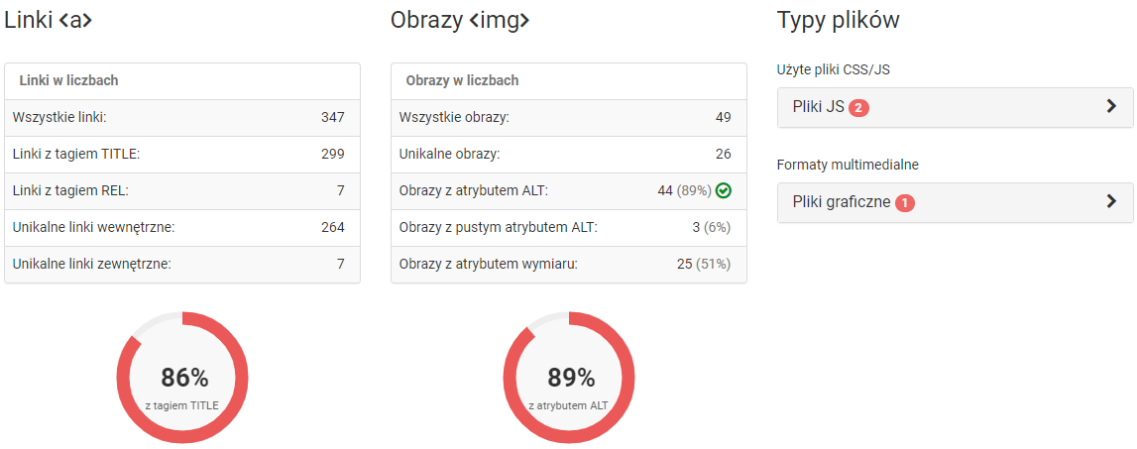
W kolejnej części raportu prezentowane są treści w wykrytych strefach tekstowych i nagłówkach H1–H6, a także lista słów kluczowych wykrytych w znacznikach TITLE, H1–H6, B i STRONG oraz lista wszystkich słów w treści strony z liczbą ich wystąpień (rysunek 23.).



Rysunek 23. Informacje o strukturze tekstu i kodu HTML, opracowanie własne

Informacje o strukturze kodu HTML obejmują listę wykorzystanych bibliotek programistycznych języka JavaScript i narzędzi Google oraz listę zastosowanych mikroformatów i znaczników META i ROBOTS. Dane obejmują również listę wszystkich znaczników języka HTML w kodzie z liczbą ich wystąpień.

W ostatniej części raportu znajdują się informacje na temat optymalizacji znaczników A i IMG (rysunek 24.) oraz udostępniana jest lista wszystkich hiperłączy wewnętrznych i zewnętrznych, a także wyniki testów czytelności.



Rysunek 24. Informacje o optymalizacji znaczników A i IMG, opracowanie własne

3.5. Podsumowanie

W tym rozdziale przedstawiono zasadę działania algorytmu rankingowego ISOWQ Rank oraz omówiono jego założenia i metodykę przydzielania punktacji za wykorzystane technologie, pozycje rankingowe, optymalizację kodu źródłowego, treść i strukturę tekstu. Zaprezentowano też pseudokod algorytmu oraz sposób jego implementacji.

Omówiono architekturę systemu ISOWQ – jego budowę i strukturę baz danych. Przedstawiono analizę zgromadzonych danych na temat wykorzystania języka znaczników HTML5, portalu YouTube do publikowania treści wideo oraz wtyczek mediów społecznościowych na stronach internetowych począwszy od 2011 roku. Omówiono także przykładowy raport techniczny dla strony WWW, z podziałem na informacje zbiorcze dla całego serwisu i jego strony głównej.

W następnym rozdziale omówiono wyniki badań porównawczych algorytmów rankingowych. Badanie polegało na wyznaczeniu współczynnika korelacji τ-Kendalla pomiędzy wynikami uzyskanymi za pomocą algorytmów ISOWQ Rank i MOZ.

4. Badanie porównawcze algorytmów rankingowych

W niniejszym rozdziale przedstawiono wyniki badań porównawczych algorytmów rankingowych. Do oszacowania relacji pomiędzy punktacjami uzyskanymi za pomocą algorytmów ISOWQ Rank i MOZ wykorzystano współczynnik τ-Kendalla. Współczynnik ten pozwala mierzyć stopień podobieństwa dwóch uporządkowanych zbiorów danych [227].

4.1. Wybór stron WWW do przeprowadzenia badań

Przy wyborze stron WWW biorących udział w badaniu kierowano się regułą, że powinny reprezentować różną tematykę i zawierać co najmniej 10 podstron. Wybrano następujące witryny internetowe:

- mlyny-rozdrabniacze.pl – sklep internetowy z damską odzieżą,
- link2europe.pl – agencja pośrednictwa pracy,
- brightmedia.pl – agencja reklamowa,
- melex.com.pl – producent pojazdów elektrycznych,
- machineryzone.pl – międzynarodowa giełda maszyn budowlanych,
- alivia.org.pl – fundacja onkologiczna „Alivia”,
- argos.org.pl – fundacja dla zwierząt „ARGOS”,
- palacporaj.pl – rezerwacje noclegów w pałacu „Poraj”,
- 4wsk.pl – 4 Wojskowy Szpital Kliniczny,
- wrobywatel.pl – inicjatywa społeczna „WrObywatel”.

Adresy stron WWW zweryfikowano pod kątem ich dostępności dla robotów systemu ISOWQ, a następnie dodano do kolejki w celu przeprowadzenia analizy. Badanie wykonano w dniach od 1 do 10 kwietnia 2019 roku.

4.2. Wstępna analiza danych

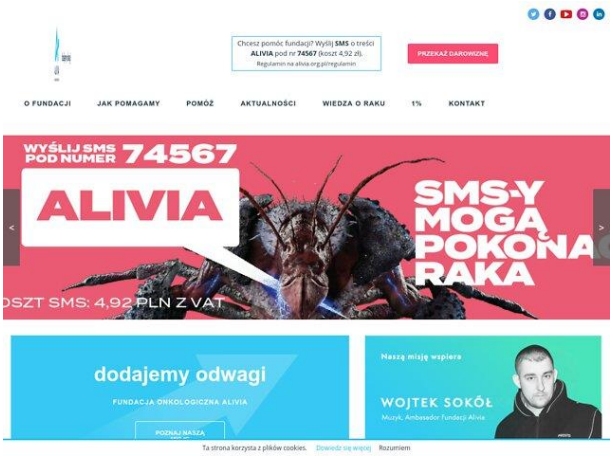
Przeprowadzona analiza obejmowała m.in. wykorzystanie następujących technologii:

- HTML 5 (+31 typów dodatkowych, m.in. HTML 2, 3, 4 i XHTML 1 i 2):
  - analiza kodu – META, HTTP-EQUIV , tagi HTML, schematy, mikroformaty,
- media społecznościowe (Facebook, Twitter, Google, LinkedIn),
- usługi Google, tj. Ads (Adwords), Adsense, Analytics, Maps:

- framework JS (jQuery, Bootstrap + 24 dodatkowe),
- multimedia (YouTube, Flash, Silverlight),
- szyfrowanie SSL, analiza niezakodowanych adresów e-mail w kodzie strony oraz rejestracja adresów IP serwerów hostujących (e-mail i DNSbl).

Najwyższą wartość rankingową w badaniu otrzymała strona WWW dostępna pod adresem alivia.org.pl, która w rankingu ISOWQ Rank uzyskała 14,21 punktu[28] (19,78 punktu za wykorzystane technologie i pozycje rankingowe, 7,08 punktu za optymalizacje kodu źródłowego i 15,83 punktu za treść i strukturę tekstu), a w rankingu MOZ – 45 i 43 punkty, odpowiednio za autorytet domeny (MOZ DA) i autorytet strony (MOZ PA).

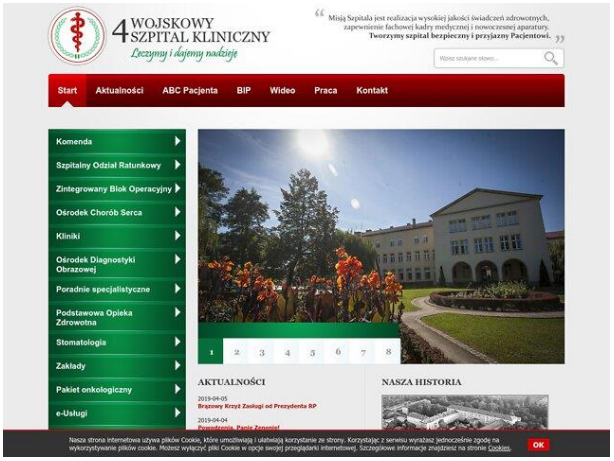
Serwis wykorzystuje technologię HTML5, ma rozbudowaną strukturę tekstu, korzysta z systemu CMS – Wordpress, wtyczek społecznościowych i szyfrowania SSL, a także publikuje treści multimedialne z portalu YouTube i ma kod HTML dobrze zoptymalizowany pod wyszukiwarki. Na rysunku 25. przedstawiono zrzut ekranu strony WWW alivia.org.pl wykonany 4 kwietnia 2019 roku.



Rysunek 25. Zrzut ekranu strony alivia.org.pl wykonany 4 kwietnia 2019 roku, opracowanie własne

Drugi wynik w badaniu otrzymała strona WWW dostępna pod adresem 4wsk.pl, która w rankingu ISOWQ Rank uzyskała 11,01 punktu[29] (11,38 punktu za wykorzystane technologie i pozycje rankingowe, 8,37 punktu za optymalizacje kodu źródłowego i 13,29 punktu za treść i strukturę tekstu), a w rankingu MOZ – 31 punktów za autorytet domeny (MOZ DA) i 34 punkty za autorytet strony (MOZ PA).

[28] ISOWQ, https://www.isowq.org/website/alivia.org.pl/1441262/, maj 2022.
[29] ISOWQ, https://www.isowq.org/website/4wsk.pl/1443221/, maj 2022.



Rysunek 26. Zrzut ekranu strony 4wsk.pl wykonany 8 kwietnia 2019 roku, opracowanie własne

Serwis ma rozbudowaną strukturę tekstu, stosuje szyfrowanie SSL i ma dobrze zoptymalizowany kod HTML. W kodzie serwisu nie wykryto wykorzystania wtyczek społecznościowych ani treści multimedialnych z portalu YouTube. Serwis jest zbudowany z zastosowaniem technologii XHTML 1.1. Na rysunku 26. przedstawiono zrzut ekranu strony WWW 4wsk.pl wykonany 8 kwietnia 2019 roku.

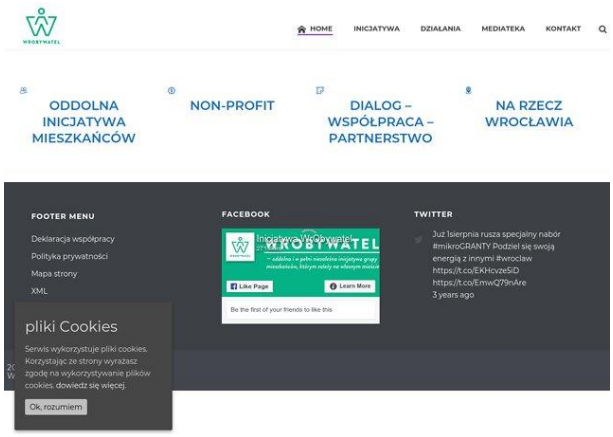
Na dalszych miejscach znalazły się strony internetowe z następującą punktacją rankingową ISOWQ Rank:

- machineryzone.pl z wynikiem 10,34 punktu[30],
- argos.org.pl z wynikiem 9,57 punktu[31],
- melex.com.pl z wynikiem 8,73 punktu[32],
- link2europe.pl z wynikiem 9,04 punktu[33],
- palacporaj.pl z wynikiem 5,86 punktu[34],
- mlyny-rozdrabniacze.pl z wynikiem 5,64 punktu[35].

Przedostatnie miejsce w badaniu otrzymała strona WWW dostępna pod adresem wrobywatel.pl, która w rankingu ISOWQ Rank[36] otrzymała 3,82 punktu (5,86 punktu za wykorzystane technologie i pozycje rankingowe, 4,03 punktu za optymalizacje kodu

[30] ISOWQ, https://www.isowq.org/website/machineryzone.pl/1440087/, maj 2022 r.
[31] ISOWQ, https://www.isowq.org/website/argos.org.pl/1441579/, maj 2022 r.
[32] ISOWQ, https://www.isowq.org/website/melex.com.pl/1439980/, maj 2022 r.
[33] ISOWQ, https://www.isowq.org/website/link2europe.pl/1439533/, maj 2022 r.
[34] ISOWQ, https://www.isowq.org/website/palacporaj.pl/1439800/, maj 2022 r.
[35] ISOWQ, https://www.isowq.org/website/mlyny-rozdrabniacze.pl/1439111/, maj 2022 r.
[36] ISOWQ, https://www.isowq.org/website/wrobywatel.pl/1444142/, maj 2022 r.

źródłowego i 1,58 punktu za treść i strukturę tekstu), a w rankingu MOZ – 14 punktów za autorytet domeny (MOZ DA) i 21 punkty za autorytet strony (MOZ PA).



Rysunek 27. Zrzut ekranu strony wrobywatel.pl wykonany 10 kwietnia 2019 roku, opracowanie własne

Serwis ma niską punktację w rankingach MOZ i Alexa, nie stosuje szyfrowania SSL, a atrybut TITLE znaczników A jest opisany zaledwie w 1%. Choć serwis udostępnia multimedia za pomocą portalu YouTube i korzysta z wtyczek społecznościowych, na każdej podstronie zawiera średnio trzy strefy tekstowe, co powoduje obniżenie punktacji zarówno za treść, jak i optymalizację kodu. Na rysunku 27. przedstawiono zrzut ekranu strony WWW wrobywatel.pl wykonany 10 kwietnia 2019 roku.

Najniższy wynik w przeprowadzonym badaniu uzyskała strona WWW dostępna pod adresem brightmedia.pl, która w rankingu ISOWQ Rank uzyskała 3,05 punktu[37] (5,48 punktu za wykorzystane technologie i pozycje rankingowe, 3,68 punktu za optymalizacje kodu źródłowego oraz 0,00 punktów za treść i strukturę tekstu), a w rankingu MOZ – 34 punkty za autorytet domeny (MOZ DA) i 27 punktów za autorytet strony (MOZ PA).

Strona nie stosuje szyfrowania SSL i przyjaznych adresów, a w kodzie nie wykryto wykorzystania wtyczek społecznościowych. Na uwagę zasługuje to, że atrybut ALT znacznika IMG jest opisany w 100%, a atrybut TITLE znaczników A – w 91%, co jest bardzo dobrym wynikiem. Na rysunku 28. przedstawiono zrzut ekranu strony WWW brightmedia.pl wykonany 1 kwietnia 2019 roku.

[37] ISOWQ, https://www.isowq.org/website/brightmedia.pl/1439337/, maj 2022 r.

Rysunek 28. Zrzut ekranu strony brightmedia.pl wykonany 1 kwietnia 2019 roku, opracowanie własne

Niekonwencjonalna metoda umieszczenia tekstu w kodzie serwisu brightmedia.pl uniemożliwia jego poprawną detekcję i analizę, co ma odzwierciedlenie w punktacji za treść i optymalizację kodu. Oto fragment kodu źródłowego tej witryny:

```
<p class="textF2" id="text1_1">Nie jesteśmy nowicjuszami,</p>
<p class="textF3" id="text1_2">nie wzięliśmy się znikąd. Mamy wiedzę</p>
<p class="textF4" id="text1_3">i doświadczenie, bo pracowaliśmy</p>
<p class="textF5" id="text1_4">w znanych agencjach dla znanych marek.</p>
```

Boty internetowe systemu ISOWQ nie wykryły w powyższym kodzie źródłowym stref tekstowych, gdyż uznały, że w pojedynczym akapicie umieszczonym w znaczniku P jest niewystarczająca liczba znaków.

Jeśli powyższy fragment kodu zostałby zastąpiony:

```
<p>Nie jesteśmy nowicjuszami,<br>
nie wzięliśmy się znikąd. Mamy wiedzę<br>
i doświadczenie, bo pracowaliśmy<br>
w znanych agencjach dla znanych marek.</p>
```

algorytm wykryłby strefę tekstową i serwis otrzymałby za strukturę tekstu wyższą punktację, która wpłynęłaby dodatnio na końcową punktację ISOWQ Rank. Jednak na podstawie kodu źródłowego można przypuszczać, że zmiana ta wpłynęłaby negatywnie na formę prezentacji tekstu na stronie.

W tabeli 9. przedstawiono wyniki otrzymanych punktacji rankingowych ISOWQ Rank i MOZ.

Tabela 9. Wyniki ISOWQ Rank i MOZ DA uzyskane w przeprowadzonym badaniu, opracowanie własne

Lp.	Analizowany serwis	PM (0–20)	PK (0–20)	PT (0–20)	ISOWQ Rank (0–20)	MOZ DA (0–100)
1	alivia.org.pl	19,78	7,03	15,83	14,21	45
2	4wsk.pl	11,38	8,37	13,29	11,01	31
3	machineryzone.pl	12,03	6,87	12,12	10,34	30
4	argos.org.pl	9,57	9,57	9,57	9,57	31
5	melex.com.pl	8,49	7,75	12,47	8,73	31
6	link2europe.pl	11,15	5,10	7,88	8,04	30
7	palacporaj.pl	8,64	5,83	3,11	5,86	22
8	mlyny-rozdrabniacze.pl	6,95	9,37	0,59	5,64	7
9	wrobywatel.pl	5,86	4,03	1,58	3,82	14
10	brightmedia.pl	5,48	3,68	0,00	3,05	34

Do obliczenia korelacji pomiędzy wynikami uzyskanymi za pomocą algorytmów ISOWQ Rank i MOZ wykorzystano język programowania R oraz środowisko do obliczeń statystycznych i wizualizacji wyników RGui. Do obliczeń zastosowano bibliotekę programistyczną Kendall[38] dla języka R, a do zapisania wyników w plikach graficznych użyto oprogramowania Ghostscript[39].

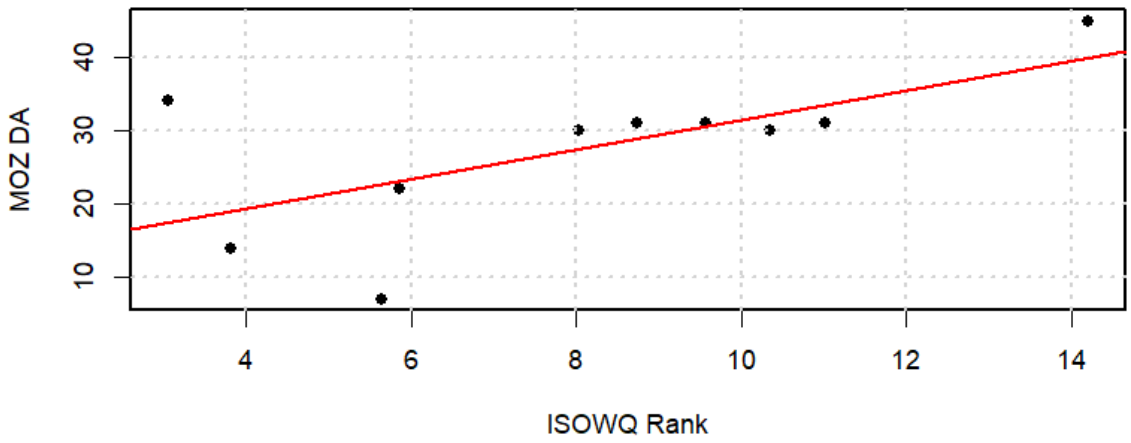
Kod źródłowy w języku R przedstawiono na listingu 11.

```
library("Kendall")

x <- c(71, 55, 52, 48, 44, 40, 29, 28, 19, 15) # ISOWQ Rank
y <- c(45 ,31, 30, 31, 31, 30, 22, 7, 14, 34)  # MOZ DA
t <- list("", "ISOWQ Rank", "MOZ DA")

k <- Kendall(x,y)

# drukuj wyniki
summary(k)

# zapisz wyniki w pliku graficznym
png(file = "kendall-isowq-mozda.png", width = 800, height = 400)
par(bg = rgb(1, 1, 1), cex = 1.5, ps = 12, lwd = 2)
plot(x, y, main = t[1], xlab = t[2], ylab = t[3], type="p", pch = 16)

grid()
abline(lm(y ~ x), col = "red")
dev.off()
```

Listing 11. Kod źródłowy w języku R do obliczenia korelacji pomiędzy ISOWQ Rank a MOZ DA

[38] Wersja 2.2.1 dla systemu Windows, wydana 20 marca 2022 r.
[39] Wersja 9.56.1 dla systemu Windows, wydana 4 kwietnia 2022 r.

Współczynnik τ-Kendalla wynosi 0,442, co oznacza, że na wstępnym etapie analizy danych algorytmy mają ze sobą umiarkowanie dodatni związek. Rysunek 29. przedstawia związek między rankingiem ISOWQ Rank a MOZ obliczony dla wszystkich stron WWW uwzględnionych w badaniu.



Rysunek 29. Związek między rankingiem ISOWQ Rank a MOZ, opracowanie własne

Porównując wyniki analiz serwisów WWW dla skrajnych punktacji ISOWQ Rank, można zauważyć, że podstawowe różnice dotyczą wykorzystania technologii społecznościowych, multimediów, szyfrowania SSL i systemu zarządzania treścią.

Zestawienie porównawcze analiz jest przedstawione w tabeli 10.

Tabela 10. Porównanie wyników analiz dla skrajnych punktacji

	brightmedia.pl	**alivia.org.pl**
ISOWQ Rank	3,05	14,21
MOZ DA	34	45
Alexa Rank (im punktacja niższa, tym lepiej)	1 095 670	496 768
Wykorzystanie wtyczek społecznościowych	Nie	Tak
Publikowanie filmów z witryny YouTube	Nie	Tak
Szyfrowanie SSL (HTTPS)	Nie	Tak
Wykorzystanie systemu CMS – WordPress	Nie	Tak
Optymalizacja kodu źródłowego	81%	91%
Wykorzystanie atrybutu TITLE w znaczniku A	91%	0%
Wykorzystanie atrybutu ALT w znaczniku IMG	100%	73%
Wykryte strefy tekstowe	Nie	Tak

Wstępna analiza danych wykazała, że istnieje dodatnia korelacja między rankingiem ISOWQ Rank a MOZ, czyli wzrost jednej punktacji powinien spowodować wzrost drugiej. Może to oznaczać, że aspekt techniczny serwisu WWW (wykorzystane technologie,

optymalizacja kodu, struktura tekstu) jest ważnym czynnikiem algorytmu MOZ, którego szczegółowa specyfikacja nie jest publicznie dostępna.

4.3. Korekta i końcowa analiza danych

Poprawne wykrycie i analiza stref tekstowych, zrozumiałych dla człowieka, jest zadaniem bardzo skomplikowanym, zwłaszcza tam, gdzie tekst odgrywa drugorzędną rolę. Pomimo że w kodzie serwisu brightmedia.pl znajdują się strefy tekstowe, choć umieszczone w niekonwencjonalny sposób, boty systemu ISOWQ ich nie wykryły, co ostatecznie spowodowało obniżenie końcowej punktacji rankingowej. Wpłynęło to również na wartość współczynnika korelacji τ-Kendalla służącego do oszacowania wielkości korelacji pomiędzy punktacjami uzyskanymi za pomocą algorytmów ISOWQ Rank i MOZ. W związku z tym, że wyniki dla strony brightmedia.pl nie pozwalają w pełni poprawnie wyliczyć współczynnika korelacji τ-Kendalla, podjęto decyzję o aktualizacji listy stron internetowych uwzględnionych w badaniu.

Utworzono dwie testowe grupy stron WWW: pierwsza – A, zawierała serwisy z listy pierwotnej z wyłączeniem strony brightmedia.pl, a w drugiej – B, w miejsce brightmedia.pl wstawiono losowo wybraną stronę WWW, która otrzymała zbliżoną liczbę punktów. Wybrano stronę naczterykopyta.pl, przeanalizowaną 5 kwietnia 2019 roku, która w rankingu ISOWQ Rank otrzymała 3,27 punktu[40] (1,95 punktu za wykorzystane technologie i pozycje rankingowe, 3,85 punktu za optymalizacje kodu źródłowego oraz 4,02 punktu za treść i strukturę tekstu), a w rankingu MOZ – 7 punktów za autorytet domeny (MOZ DA) i 18 punktów za autorytet strony (MOZ PA). Na rysunku 30. przedstawiono zrzut ekranu strony WWW naczterykopyta.pl wykonany 5 kwietnia 2019 roku.

Rysunek 30. Zrzut ekranu strony naczterykopyta.pl wykonany 5 kwietnia 2019 roku, opracowanie własne

[40] ISOWQ, https://www.isowq.org/website/naczterykopyta.pl/1441827, maj 2022 r.

Liczba podstron w serwisie naczterykopyta.pl jest zbliżona do liczby podstron na stronie brightmedia.pl. Również on nie stosuje szyfrowania SSL i przyjaznych adresów, a w kodzie nie wykryto wtyczek społecznościowych. Istotna różnica w wynikach analizy obu stron jest związana z wykryciem w kodzie HTML stref tekstowych, co ma istotny wpływ na punktację za treść i strukturę tekstu.

Lista stron internetowych w utworzonych grupach testowych A i B jest przedstawiona w tabeli 11. Grupa A obejmuje 9 stron WWW z wyłączeniem serwisu brightmedia.pl, a grupa B – 10 serwisów, przy czym stronę brightmedia.pl zamieniono na naczterykopyta.pl.

Tabela 11. Lista stron WWW w grupach testowych A i B

Lp.	Grupa A	Grupa B	ISOWQ Rank (0–20)	MOZ DA (0–100)
1	alivia.org.pl	alivia.org.pl	14,21	45
2	4wsk.pl	4wsk.pl	11,01	31
3	machineryzone.pl	machineryzone.pl	10,34	30
4	argos.org.pl	argos.org.pl	9,57	31
5	melex.com.pl	melex.com.pl	8,73	31
6	link2europe.pl	link2europe.pl	8,04	30
7	palacporaj.pl	palacporaj.pl	5,86	22
8	mlyny-rozdrabniacze.pl	mlyny-rozdrabniacze.pl	5,64	7
9	wrobywatel.pl	wrobywatel.pl	3,82	14
10	-	naczterykopyta.pl	3,27	7

Do obliczenia współczynnika korelacji τ-Kendalla wykorzystano język R z biblioteką Kendall oraz środowisko do obliczeń statystycznych i wizualizacji wyników RGui.

Obliczenia wykonano dla następujących par danych:

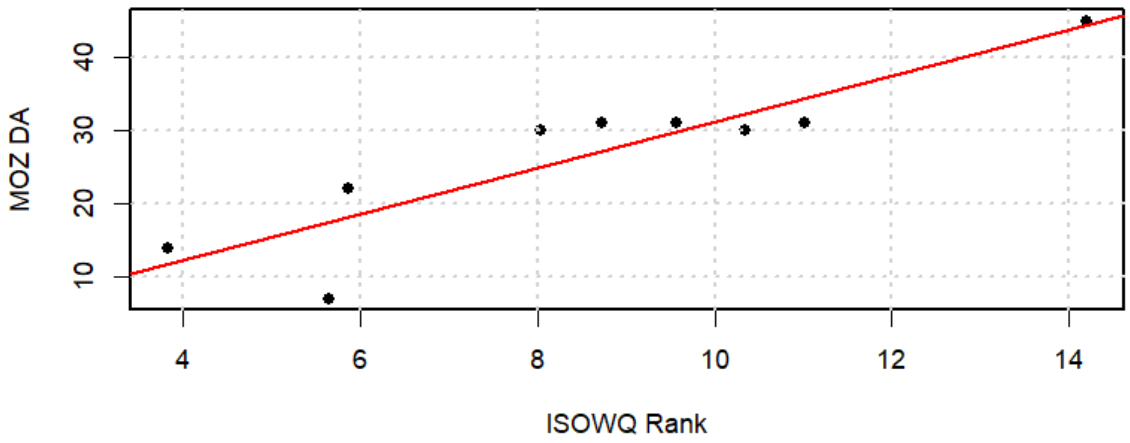
- ISOWQ Rank – MOZ DA, korelacja pomiędzy rankingami,
- $\overline{PK + PT}$ – MOZ DA, korelacja pomiędzy średnią arytmetyczną z PK i PT a punktacją MOZ DA,
- PM – MOZ DA, korelacja pomiędzy punktacją za wykorzystane technologie i pozycje rankingowe a punktacją MOZ DA,
- PK – MOZ DA, korelacja pomiędzy punktacją za optymalizację kodu źródłowego a punktacją MOZ DA,
- PT – MOZ DA, korelacja pomiędzy punktacją za treść i strukturę tekstu a punktacją MOZ DA.

Wyniki obliczeń współczynników korelacji τ-Kendalla i poziomów istotności *p* dla każdej pary danych w grupie A przedstawiono w tabeli 12.

Tabela 12. Związek czynników algorytmu ISOWQ Rank z MOZ DA – grupa A, opracowanie własne

	ISOWQ Rank	$\overline{PK + PT}$	PM	PK	PT
τ-Kendalla	0,766	0,825	0,530	0,295	0,884
Wartość *p*	0,007469	0,003863	0,068895	0,33552	0,001915
Ranking Kendalla	26	28	18	10	30

Usunięcie z pierwotnej listy danych strony brightmedia.pl, które uznano za anomalię, znacznie poprawiło wyniki obliczeń. Współczynnik τ-Kendalla wynosi 0,766, co przy zadeklarowanym poziomie istotności 0,05 oznacza, że algorytmy mają ze sobą dodatnią korelację. Na uwagę zasługuje wartość współczynnika τ-Kendalla pomiędzy punktacją za treść i strukturę tekstu – *PT*, a punktacją MOZ DA, z której wynika, że pomiędzy tą parą danych zachodzi silna dodatnia korelacja. Wyniki wskazują na słabą korelację pomiędzy parą PK a MOZ DA i korelację umiarkowaną pomiędzy parą PM a MOZ DA. Silna korelacja pomiędzy średnią punktacją za optymalizację kodu – *PK*, i za treść – *PT*, a MOZ DA może wskazywać, że czynniki wpływające na ranking MOZ są zbliżone do tych, które wpływają na wartość rankingową wyznaczaną przez algorytm ISOWQ Rank. Rysunek 31. przedstawia związek między punktacją ISOWQ Rank a MOZ DA obliczony dla stron WWW ujętych w grupie A.



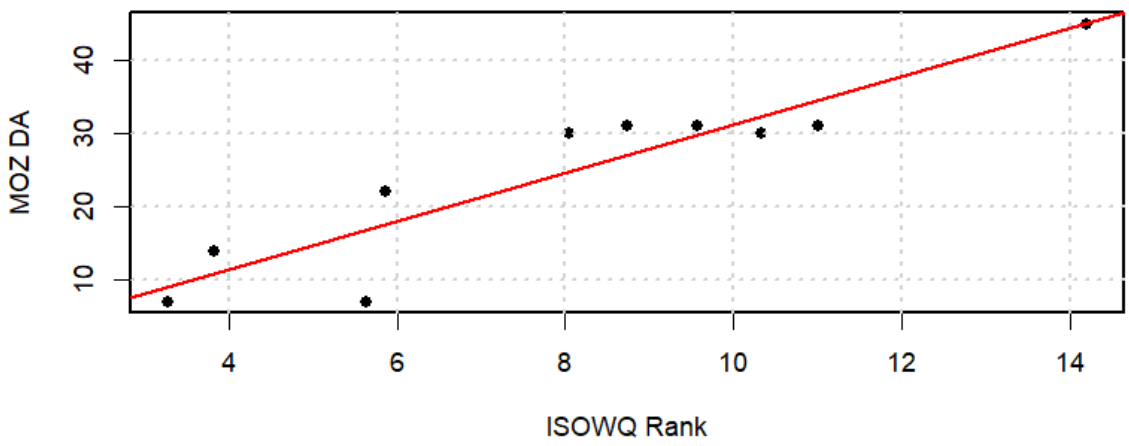
Rysunek 31. Związek między rankingiem ISOWQ Rank a MOZ w grupie A, opracowanie własne

Wyniki obliczeń współczynników korelacji τ-Kendalla i poziomów istotności *p* dla każdej pary danych w grupie B przedstawiono w tabeli 13.

Tabela 13. Związek czynników algorytmu ISOWQ Rank z MOZ DA – grupa B, opracowanie własne

	ISOWQ Rank	$\overline{PK + PT}$	PM	PK	PT
τ-Kendalla	0,801	0,801	0,613	0,424	0,801
Wartość *p*	0,0025204	0,0025204	0,022106	0,11966	0,0025204
Ranking Kendalla	34	34	26	18	34

Zamiana strony brightmedia.pl na stronę naczterykopyta.pl przyczyniła się, podobnie jak w grupie A, do uzyskania lepszych wyników. Współczynnik τ-Kendalla wynosi 0,801, co przy zadeklarowanym poziomie istotności 0,05 oznacza, że algorytmy mają ze sobą dodatnią korelację. Wyniki wskazują na silną dodatnią korelację par danych – PT i $\overline{PK + PT}$, z MOZ DA, co, analogicznie jak w przypadku wyników uzyskanych w grupie A, może wskazywać, że czynniki wpływające na ranking MOZ są podobne do czynników rankingu ISOWQ Rank. Rysunek 32. przedstawia związek między punktacją ISOWQ Rank a MOZ DA obliczony dla stron WWW ujętych w grupie B.



Rysunek 32. Związek między rankingiem ISOWQ Rank a MOZ w grupie B, opracowanie własne

Decyzja o usunięciu anomalii ze zbioru danych i utworzeniu grup testowych A i B okazała się optymalna dla poprawnego wykonania obliczeń współczynników korelacji τ--Kendalla. Wykazano, że istnieje dodatnia korelacja pomiędzy punktacją ISOWQ Rank a MOZ DA, co oznacza, że wzrost jednej powinien spowodować wzrost drugiej.

4.4. Podsumowanie

W niniejszym rozdziale zaprezentowano wyniki badań porównawczych algorytmów rankingowych ISOWQ Rank i MOZ. Do oceny ich wspólnej zależności wykorzystano

współczynnik korelacji τ-Kendalla przy zadeklarowanym poziomie istotności 0,05. Badanie wykazało, że pojawiające się anomalie w zbiorze danych mogą wpłynąć negatywnie na jakość obliczeń, a tym samym nie odzwierciedlać faktycznej zależności pomiędzy analizowanymi danymi.

Korekta danych, polegająca na utworzeniu dwóch testowych grup ze stronami WWW, wykazała, że w obu tych grupach wyniki jednoznacznie wskazują na dodatnią korelację pomiędzy punktacją ISOWQ Rank a punktacją MOZ. Ponadto przedstawiono kod źródłowy w języku R, za pomocą którego wykonano obliczenia na potrzeby przeprowadzonego badania.

5. Zakończenie

W niniejszej pracy opublikowano autorski algorytm rankingowy ISOWQ Rank, który nadaje stronom WWW określoną wartość, oznaczającą ich jakość. Skuteczność algorytmu zmierzono w trakcie badań porównawczych, w których wykazano dodatnią korelację pomiędzy punktacją uzyskaną dla strony WWW za pomocą algorytmu ISOWQ Rank a punktacją MOZ.

Algorytm MOZ do obliczenia punktacji wykorzystuje wiedzę na temat liczby wysokiej jakości hiperłączy przychodzących oraz analizuje kod źródłowy strony WWW pod kątem stosowanych technik optymalizacyjnych. Zgodnie z założeniem algorytmu im uzyskana punktacja jest wyższa, tym większe jest prawdopodobieństwo pojawienia się strony WWW na wyższych pozycjach w SERP. Biorąc pod uwagę autorytarność narzędzia MOZ Analytics wśród specjalistów SEO, stworzenie konkurencyjnego algorytmu rankingowego o zbliżonej skuteczności umożliwiło odkrycie wiedzy na temat czynników wpływających nie tylko na ranking MOZ, ale również na ranking w wyszukiwarkach.

W trakcie badań przestudiowano algorytmy rankingowe wykorzystujące do oceny jakości stron WWW strukturę hiperłączy, jak i analizę słów kluczowych, znaczników HTML czy liczbę odwiedzin. Zebrano aktualną wiedzę na temat technik optymalizacji w obrębie strony WWW i poza nią oraz omówiono narzędzia umożliwiające przeprowadzenie audytu technicznego witryn internetowych. W trakcie rozprawy przeanalizowano dotychczasowe badania związane z analizą wyników w SERP i próbami odkrycia czynników wpływających na pozycje rankingowe w wyszukiwarkach.

Niniejsza rozprawa jest podzielona na cztery rozdziały, z których pierwsze dwa obejmują część teoretyczną, a dwa pozostałe – część praktyczną. W rozdziale 1. dokonano przeglądu wyszukiwarek stron WWW, które od roku 1990, kiedy to udostępniono Archie, weszły do powszechnego użycia. Zaprezentowano rodzaje wyników w wyszukiwarkach z podziałem na płatne i organiczne, a także metody optymalizacji pod wyszukiwarki dokonywane bezpośrednio na stronie WWW, jak i poza nią. W rozdziale 2. omówiono fundamenty dzisiejszych algorytmów rankingowych i przedstawiono zasadę działania algorytmów PageRank i HITS, wykorzystując specjalnie w tym celu opracowane narzędzie programistyczne. Zamieszczone w pracy wyniki badań pozwalają jednoznacznie stwierdzić, że odkrycie wszystkich czynników wpływających na ranking, zwłaszcza w wyszukiwarce Google, jest w zasadzie niemożliwe.

W części praktycznej, w rozdziale 3., przedstawiono zasadę działania algorytmu rankingowego ISOWQ Rank, omówiono jego założenia i metodykę przydzielania punktacji za wykorzystane technologie, pozycje rankingowe, optymalizację kodu źródłowego, treść i strukturę tekstu. Zaprezentowano pseudokod algorytmu oraz sposób jego implementacji. Omówiono architekturę systemu ISOWQ, jego budowę i strukturę baz danych. Przedstawiono analizę zgromadzonych danych począwszy od 2011 roku i omówiono przykładowy raport techniczny dla strony WWW, z podziałem na informacje zbiorcze dla całego serwisu i jego strony głównej. W rozdziale 4. zaprezentowano wyniki badań porównawczych algorytmów rankingowych ISOWQ Rank i MOZ. Do oceny ich wspólnej zależności wykorzystano współczynnik korelacji τ-Kendalla przy zadeklarowanym poziomie istotności 0,05. Wykazano, że istnieje dodatnia korelacja pomiędzy punktacją ISOWQ Rank a MOZ, co oznacza, że wzrostowi jednej powinien towarzyszyć wzrost drugiej.

Najważniejsze konkluzje zebrano poniżej:

- Już od czasu pierwszych wyszukiwarek z lat 90. XX wieku istnieje problem z ustaleniem optymalnej listy wyników. Pierwsze metody opierały się na analizie słów kluczowych i znaczników HTML, jednak metody te nie przynosiły zadowalających wyników, głównie ze względu na niską wiarygodność i częste nadużycia ze strony projektantów stron WWW. Problemy z ustaleniem, które zasoby w sieci można uznać za istotne, doprowadziły do opracowania w 1998 roku algorytmu HITS oraz algorytmu PageRank, które stały się fundamentem większości dzisiejszych algorytmów rankingowych.
- Sergey Brin i Larry Page, twórcy algorytmu PageRank, podobnie jak Jon Kleinberg, który stworzył HITS, założyli, że sieć połączonych ze sobą hiperłączami stron WWW przypomina graf. Przyjęli również założenie, że o wadze publikacji świadczy liczba odwołań z innych publikacji, czyli waga strony WWW może być mierzona liczbą hiperłączy wskazujących tę stronę z innych stron. Algorytm PageRank, w odróżnieniu od zwykłego zliczania hiperłączy, stosuje ważenie ich wartości, co powoduje, że strona WWW może uzyskać wysoką pozycję rankingową, jeśli jest linkowana ze stron o wysokim rankingu.
- Obecnie pierwsze miejsce wśród wyszukiwarek internetowych zajmuje Google, która, rozwijana od 25 lat, osiągnęła udział 91,90%. Pozostałe miejsca zajmują: Bing (2,88%), Yahoo! (1,51%), Yandex (1,28%), Baidu (1,14%) i pozostałe (1,28%). Przez lata SERP zmieniał swój wygląd i formę, głównie za sprawą wyszukiwarki Google,

która wytycza trendy w formach prezentowania wyników wyszukiwań. Google oferuje też najwięcej rodzajów wyników wyszukiwania.

- SERP obejmuje bezpłatne wyniki organiczne, czyli naturalne wyniki wyszukiwania, o których kolejności decyduje algorytm wyszukiwarki, oraz płatne, w formie reklam tekstowych i produktowych. O pozycji rankingowej w wynikach organicznych decyduje algorytm wyszukiwarki, natomiast w przypadku reklam płatnych decydującym czynnikiem jest mechanizm aukcyjny platformy. Wyszukiwarki starają się poprawiać dokładność odpowiedzi za pomocą wyszukiwania semantycznego, polegającego na próbie zrozumienia języka naturalnego, co biorąc pod uwagę liczbę języków, jest zadaniem bardzo trudnym.
- Optymalizacja w obrębie strony WWW jest istotnym elementem działań zmierzających do poprawy pozycji rankingowych w wyszukiwarkach. Działania związane z optymalizacją struktury witryny internetowej, występujących na niej treści i słów kluczowych powinny być wykonane na początkowym etapie budowy witryny internetowej. Proces optymalizacji powinien być realizowany cyklicznie, wraz ze zmieniającymi się wytycznymi algorytmów wyszukiwarek. Najważniejsze elementy optymalizacji w obrębie strony WWW to znaczniki TITLE i META DESCRIPTION, nagłówki H1–H6, nasycenie słów kluczowych w treści, atrybut ALT znacznika IMG, atrybut TITLE znacznika A, linkowanie wewnętrze, składnia hiperłączy, zgodność ze standardami W3C, dostosowanie do urządzeń mobilnych i szybkość wyświetlania się w przeglądarkach internetowych.
- Optymalizacja poza stroną WWW obejmuje działania związane z pozyskiwaniem wysokiej jakości hiperłączy z innych serwisów internetowych, forów, blogów, optymalizację wizytówki Google czy rekomendacje z mediów społecznościowych. Pozyskiwane hiperłącza są traktowane przez wyszukiwarki jako polecenia, dlatego mają znaczący wpływ na pozycje rankingowe w SERP.
- Wyniki badań przedstawione w rozdziale 2., odnoszące się do analizy SERP w wyszukiwarce Google, wskazują, że w praktyce nie jest możliwe dokładne wyliczenie pozycji rankingowej, głównie ze względu na brak dostępu do informacji na temat konstrukcji algorytmu rankingowego. Do dziś nie opracowano skutecznej metody umożliwiającej odkrycie wszystkich czynników, które wpływają na ranking w wyszukiwarce Google. W przedstawionych badaniach dokonano szczegółowej analizy

treści i kodu stron WWW, wykorzystano wartość PageRank, a pomimo to nie udało się odkryć czynników, które znacząco decydują o pozycjach w SERP.

- Algorytm ISOWQ Rank umożliwia ocenę jakości strony WWW na podstawie analizy wykorzystanych technologii, pozycji rankingowych i parametrów technicznych serwera WWW – *PM*, optymalizacji kodu źródłowego – *PK*, oraz treści i struktury tekstu – *PT*. Podczas obliczania rankingu przyjęto założenie, że statystyczna strona WWW prezentująca podstawowe informacje biznesowe to struktura połączonych hiperłączy, która obejmuje stronę główną i co najmniej cztery podstrony. Taka struktura powinna zawierać niezbędne dane o firmie, informacje na temat bieżącej działalności, aktualną ofertę i formularz kontaktowy. Ranking jest obliczany dla pierwszych 30 adresów URL odnalezionych w kodzie strony WWW, które zostaną poprawnie pobrane z serwera, czyli wtedy, kiedy serwer zwróci kod HTTP 2xx. Przyjęto założenie, że najważniejsze informacje na stronie WWW powinny się znajdować na pierwszych dwóch poziomach linkowania. Przyjęto też zasadę, że punktacja zostanie proporcjonalnie obniżona z wykorzystaniem współczynnika *LR*, jeśli w analizowanej stronie WWW jest mniej podstron niż wymagane cztery, a także w przypadku, kiedy występują odnośniki wewnętrzne do nieistniejących stron, czyli wtedy, gdy serwer zwraca kod błędu HTTP 4xx lub 5xx. Punktacja za wykorzystane technologie i pozycje rankingowe – *PM*, jest obliczana na podstawie informacji o wartościach rankingowych MOZ i Alexa Rank, zastosowaniu wtyczek społecznościowych i szyfrowania SSL oraz liczbie hiperłączy przychodzących. Punktacja uwzględnia również informacje o fizycznej lokalizacji serwera WWW, występowaniu na stronie jawnych adresów e-mail i zarejestrowaniu adresu IP serwera hostującego w bazach DNSbl. Z kolei punktacja za optymalizację kodu źródłowego – *PK*, jest obliczana na podstawie informacji o niepowtarzalności tytułów i opisów stron zawartych w znacznikach META, zastosowaniu danych strukturalnych i poprawnym stosowaniu znaczników HTML. Ponadto analizowana jest optymalizacja wielkości kodu, relacja rozmiaru tekstu do rozmiaru kodu źródłowego oraz użycie stylów kaskadowych i kodu JavaScript w zewnętrznych plikach. Na punktację za treść i strukturę tekstu – *PT*, wpływają zaś rozmiar i formatowanie widocznego tekstu na stronie oraz wyniki testów czytelności.

Jedną z najważniejszych konkluzji tej pracy jest stwierdzenie, że treść i struktura tekstu na stronie WWW są kluczowe dla osiągnięcia wysokich pozycji rankingowych w wyszukiwarkach. Witryny internetowe, na których umieszczone są treści wysokiej jakości,

będą wyświetlane znacznie wyżej w SERP niż te, w których treści nie są merytoryczne lub są przesycone słowami kluczowymi. Odpowiednie i naturalne umieszczenie słów kluczowych w treści pozwala przyciągnąć uwagę użytkowników, a wyszukiwarkom umożliwia odpowiednią kategoryzację poruszanej tematyki. Istotne jest to, aby tekst na stronie WWW zawierał informacje przeznaczone dla realnego obiorcy, a nie dla botów wyszukiwarek. Powiązanie treści na stronie WWW przez wewnętrzne linkowanie przyśpiesza proces indeksowania rozbudowanej witryny.

Technicznym wynikiem badań jest zaprojektowanie i wdrożenie zautomatyzowanego systemu analitycznego ISOWQ, dostępnego pod adresem www.isowq.org, w którym zaimplementowano algorytm rankingowy ISOWQ Rank. System gromadzi i udostępnia dane w formie okresowych raportów, co umożliwia monitorowanie zmian zachodzących w stosowanych technikach budowy i optymalizacji stron WWW.

Efekty kształcenia osiągnięto przez umiejętne wyszukiwanie źródeł, poszukiwanie nowych informacji, a także przez nabycie nowych umiejętności i rozwinięcie własnego warsztatu w pracy badawczej. Pogłębienie wiedzy w zakresie metod analizy danych i eksploracji zasobów sieci internet, a także zdobycie dodatkowych umiejętności wykorzystywania narzędzi programistycznych oraz projektowania i konstruowania systemów informatycznych pozwoliło szerzej poznać i lepiej ocenić zagadnienia poruszane w niniejszej pracy.

6. Bibliografia

[1] F. Ali and S. Khusro, “Content and link-structure perspective of ranking webpages: A review,” *Comput. Sci. Rev.*, vol. 40, p. 100397, May 2021, doi: 10.1016/j.cosrev.2021.100397.

[2] J. Marszałkowski, J.M. Marszałkowski, and M. Drozdowski, “Empirical study of load time factor in search engine ranking,” *J. Web Eng.*, vol. 13, no. 2, pp. 114–128, 2014.

[3] C.-J. Luh, S.-A. Yang, and T.-L.D. Huang, “Estimating Google’s search engine ranking function from a search engine optimization perspective,” *Online Inf. Rev.*, vol. 40, no. 2, pp. 239–255, Apr. 2016, doi: 10.1108/OIR-04-2015-0112.

[4] D. Sharma, R. Shukla, A.K. Giri, and S. Kumar, “A Brief Review on Search Engine Optimization,” in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2019, vol. 2, no. 4, pp. 687–692, doi: 10.1109/CONFLUENCE.2019.8776976.

[5] M. Duka, “Ranking of websites created with the use of ISOWQ Rank algorithm,” *Inform. Autom. Pomiary w Gospod. i Ochr. Środowiska*, vol. 10, no. 2, pp. 16–19, Jun. 2020, doi: 10.35784/iapgos.898.

[6] V. Jindal, S. Bawa, and S. Batra, “A review of ranking approaches for semantic search on Web,” *Inf. Process. Manag.*, vol. 50, no. 2, pp. 416–425, Mar. 2014, doi: 10.1016/j.ipm.2013.10.004.

[7] G.O. Strawn, “Leadership in Science and Technology: A Reference Handbook,” SAGE Publications, Inc., 2012, doi: 10.4135/9781412994231.

[8] D.G. Perry, S.H. Blumenthal, and R.M. Hinden, “The ARPANET and the DARPA Internet,” *Libr. Hi Tech*, vol. 6, no. 2, pp. 51–62, Feb. 1988, doi: 10.1108/eb047726.

[9] G.O’Regan, “Giants of Computing”. London: Springer London, 2013, doi: 10.1007/978-1-4471-5340-5.

[10] T. Berners-Lee and J.-F. Groff, “WWW,” *ACM SIGBIO Newsl.*, vol. 12, no. 3, pp. 37–40, Sep. 1992, doi: 10.1145/147126.147133.

[11] D. Tjondronegoro and A. Spink, “Web search engine multimedia functionality,” *Inf. Process. Manag.*, vol. 44, no. 1, pp. 340–357, Jan. 2008, doi: 10.1016/j.ipm.2007.03.004.

[12] T. Seymour, D. Frantsvog, and S. Kumar, “History Of Search Engines,” *Int. J. Manag. Inf. Syst.*, vol. 15, no. 4, p. 47, Sep. 2011, doi: 10.19030/ijmis.v15i4.5799.

[13] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian, “The Connectivity Server: fast access to linkage information on the Web,” *Comput. Networks ISDN Syst.*, vol. 30, no. 1–7, pp. 469–477, Apr. 1998, doi: 10.1016/S0169-7552(98)80047-0.

[14] M. Beg, “A subjective measure of web search quality,” *Inf. Sci. (Ny).*, vol. 169, no. 3–4, pp. 365–381, Feb. 2005, doi: 10.1016/j.ins.2004.07.003.

[15] C. Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis, and A. Tsakalidis, “A web page usage prediction scheme using sequence indexing and clustering techniques,” *Data Knowl. Eng.*, vol. 69, no. 4, pp. 371–382, Apr. 2010, doi: 10.1016/j.datak.2009.04.010.

[16] W.A. Aiello, “Algorithms and Models for the Web-Graph”, vol. 4936. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, doi: 10.1007/978-3-540-78808-9.

[17] Y. Du, C. Li, Q. Hu, X. Li, and X. Chen, “Ranking webpages using a path trust knowledge graph,” *Neurocomputing*, vol. 269, pp. 58–72, Dec. 2017, doi: 10.1016/j.neucom.2016.08.142.

[18] A. Makkar and N. Kumar, “User behavior analysis-based smart energy management for webpage ranking: Learning automata-based solution,” *Sustain. Comput. Informatics Syst.*, vol. 20, pp. 174–191, Dec. 2018, doi: 10.1016/j.suscom.2018.02.003.

[19] J. Cho, H. Garcia-Molina, and L. Page, “Efficient crawling through URL ordering,” *Comput. Networks ISDN Syst.*, vol. 30, no. 1–7, pp. 161–172, Apr. 1998, doi: 10.1016/S0169-7552(98)00108-1.

[20] A.K. Sharma, V. Shrivastava, and H. Singh, “Experimental performance analysis of web crawlers using single and Multi-Threaded web crawling and indexing algorithm for the application of smart web contents,” *Mater. Today Proc.*, vol. 37,

no. Part 2, pp. 1403–1408, 2021, doi: 10.1016/j.matpr.2020.06.596.

[21] G. Suchacka, A. Cabri, S. Rovetta, and F. Masulli, “Efficient on-the-fly Web bot detection,” *Knowledge-Based Syst.*, vol. 223, p. 107074, Jul. 2021, doi: 10.1016/j.knosys.2021.107074.

[22] M. Kumar, A. Bindal, R. Gautam, and R. Bhatia, “Keyword query based focused Web crawler,” *Procedia Comput. Sci.*, vol. 125, pp. 584–590, 2018, doi: 10.1016/j.procs.2017.12.075.

[23] M.D. Dikaiakos, A. Stassopoulou, and L. Papageorgiou, “An investigation of web crawler behavior: characterization and metrics,” *Comput. Commun.*, vol. 28, no. 8, pp. 880–897, May 2005, doi: 10.1016/j.comcom.2005.01.003.

[24] Q. Bai, G. Xiong, Y. Zhao, and L. He, “Analysis and Detection of Bogus Behavior in Web Crawler Measurement,” *Procedia Comput. Sci.*, vol. 31, pp. 1084–1091, 2014, doi: 10.1016/j.procs.2014.05.363.

[25] A. Batzios, C. Dimou, A. Symeonidis, and P. Mitkas, “BioCrawler: An intelligent crawler for the semantic web,” *Expert Syst. Appl.*, vol. 35, no. 1–2, pp. 524–530, Jul. 2008, doi: 10.1016/j.eswa.2007.07.054.

[26] D. Stevanovic, A. An, and N. Vlajic, “Feature evaluation for web crawler detection with data mining techniques,” *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8707–8717, Aug. 2012, doi: 10.1016/j.eswa.2012.01.210.

[27] T. Tanaka, H. Niibori, S. Li, S. Nomura, H. Kawashima, and K. Tsuda, “Bot Detection Model using User Agent and User Behavior for Web Log Analysis,” *Procedia Comput. Sci.*, vol. 176, pp. 1621–1625, 2020, doi: 10.1016/j.procs .2020.09.185.

[28] V. Shrivastava, H. Singh, and A.K. Sharma, “Meta-heuristic approach to enhance the performance of web crawler for web page clustering and link priority evaluation,” *Mater. Today Proc.*, Oct. 2020, doi: 10.1016/j.matpr.2020.09.342.

[29] R.-C. Chen and C.-H. Hsieh, “Web page classification based on a support vector machine using a weighted vote schema,” *Expert Syst. Appl.*, vol. 31, no. 2, pp. 427–435, Aug. 2006, doi: 10.1016/j.eswa.2005.09.079.

[30] E. Buber and B. Diri, “Web Page Classification Using RNN,” *Procedia Comput.*

Sci., vol. 154, pp. 62–72, 2019, doi: 10.1016/j.procs.2019.06.011.

[31] C. Nigam and A.K. Sharma, "Experimental performance analysis of web recommendation model in web usage mining using KNN page ranking classification approach," *Mater. Today Proc.*, Oct. 2020, doi: 10.1016/j.matpr.2020.09.364.

[32] S.A. Özel, "A Web page classification system based on a genetic algorithm using tagged-terms as features," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3407–3415, Apr. 2011, doi: 10.1016/j.eswa.2010.08.126.

[33] D. Wilkinson and M. Thelwall, "Search markets and search results: The case of Bing," *Libr. Inf. Sci. Res.*, vol. 35, no. 4, pp. 318–325, Oct. 2013, doi: 10.1016/j.lisr.2013.04.006.

[34] E. Van Couvering, "The History of the Internet Search Engine: Navigational Media and the Traffic Commodity," 2008, pp. 177–206, doi: 10.1007/978-3-540-75829-7_11.

[35] M. Jiang and K. Okamoto, "National Identity, Ideological Apparatus, or Panopticon? A Case Study of the Chinese National Search Engine Jike," *Policy & Internet*, vol. 6, no. 1, pp. 89–107, Mar. 2014, doi: 10.1002/1944-2866.POI353.

[36] R. Gao and C. Shah, "Toward creating a fairer ranking in search engine results," *Inf. Process. Manag.*, vol. 57, no. 1, p. 102138, Jan. 2020, doi: 10.1016/j.ipm.2019.102138.

[37] A. Strzelecki, "Eye-Tracking Studies of Web Search Engines: A Systematic Literature Review," *Information*, vol. 11, no. 6, p. 300, Jun. 2020, doi: 10.3390/info11060300.

[38] D. Lewandowski and Y. Kammerer, "Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research," *Behav. Inf. Technol.*, pp. 1–31, May 2020, doi: 10.1080/0144929X.2020.1761450.

[39] M. Lee, W. Kim, and S. Park, "Searching and ranking method of relevant resources by user intention on the Semantic Web," *Expert Syst. Appl.*, vol. 39, no. 4, pp. 4111–4121, Mar. 2012, doi: 10.1016/j.eswa.2011.09.127.

[40] D. Sánchez, L. Martínez-Sanahuja, and M. Batet, "Survey and evaluation of web

search engine hit counts as research tools in computational linguistics," *Inf. Syst.*, vol. 73, pp. 50–60, Mar. 2018, doi: 10.1016/j.is.2017.12.007.

[41] F.Z. Fagroud, L. Ajallouda, E.H. Ben Lahmar, H. Toumi, K. Achtaich, and S. El Filali, "IOT Search Engines: Exploratory Data Analysis," *Procedia Comput. Sci.*, vol. 175, pp. 572–577, 2020, doi: 10.1016/j.procs.2020.07.082.

[42] V. Derhami, E. Khodadadian, M. Ghasemzadeh, and A.M. Zareh Bidoki, "Applying reinforcement learning for web pages ranking algorithms," *Appl. Soft Comput.*, vol. 13, no. 4, pp. 1686–1692, Apr. 2013, doi: 10.1016/j.asoc .2012.12.023.

[43] H. Li, "Internet Tourism Resource Retrieval Using PageRank Search Ranking Algorithm," *Complexity*, vol. 2021, pp. 1–11, May 2021, doi: 10.1155/2021/ 5114802.

[44] M. Jiang, "Search Concentration, Bias, and Parochialism: A Comparative Study of Google, Baidu, and Jike's Search Results From China," *J. Commun.*, vol. 64, no. 6, pp. 1088–1110, Dec. 2014, doi: 10.1111/jcom.12126.

[45] N. Dwivedi, L. Joshi, and N. Gupta, "Statistical Analysis of Search Engines (Google, Yahoo and Altavista) for Their Search Result," *Int. J. Comput. Theory Eng.*, pp. 298–301, 2013, doi: 10.7763/IJCTE.2013.V5.697.

[46] C. Behnert and D. Lewandowski, "Ranking Search Results in Library Information Systems — Considering Ranking Approaches Adapted From Web Search Engines," *J. Acad. Librariansh.*, vol. 41, no. 6, pp. 725–735, Nov. 2015, doi: 10.1016/j.acalib.2015.07.010.

[47] A. Strzelecki, "Google Web and Image Search Visibility Data for Online Store," *Data*, vol. 4, no. 3, p. 125, Aug. 2019, doi: 10.3390/data4030125.

[48] N. Höchstötter and D. Lewandowski, "What users see – Structures in search engine results pages," *Inf. Sci. (Ny).*, vol. 179, no. 12, pp. 1796–1812, May 2009, doi: 10.1016/j.ins.2009.01.028.

[49] A. Strzelecki and P. Rutecka, "The Snippets Taxonomy in Web Search Engines," 2019, pp. 177–188, doi: 10.1007/978-3-030-31143-8_13.

[50] A. Strzelecki and P. Rutecka, "Direct Answers in Google Search Results," *IEEE*

Access, vol. 8, pp. 103642–103654, 2020, doi: 10.1109/ACCESS.2020.2999160.

[51] A. Singhal, "Official Google Blog: Introducing the Knowledge Graph: things, not strings," *Official Google Blog*, 2012.

[52] H. Lee, "Viewing Local Organization Data from Google My Business," in *Hands On With Google® Data Studio*, Wiley, 2020, pp. 173–219, doi: 10.1002/9781119616238.ch7.

[53] M. Durica and L. Svabova, "Improvement of Company Marketing Strategy Based on Google Search Results Analysis," *Procedia Econ. Financ.*, vol. 26, pp. 454–460, 2015, doi: 10.1016/S2212-5671(15)00873-4.

[54] B. Ciepłuch, R. Jacob, P. Mooney, and A.C. Winstanley, "Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps," *Proc. Ninth Int. Symp. Spat. Accuracy Assess. Nat. Resources Enviromental Sci.*, p. 337, Jul. 2010.

[55] A. Strzelecki, M. Wałach, and K. Deja, „Wyszukiwanie sponsorowane jako wsparcie procesu rekrutacji na studia na wydziale Informatyki i Komunikacji", w: *Zarządzanie i Informatyka – Dylematy i kierunki rozwoju*, 2010, str. 637–650.

[56] M. Bronicki and K. Sołoducha, „Realizacja strategii inbound marketingu przy wykorzystaniu wyszukiwarek internetowych", *Nowocz. Syst. Zarządzania*, vol. 12, nr 1, str. 183–199, marzec 2017, doi: 10.37055/nsz/129460.

[57] B.J. Jansen and T. Mullen, "Sponsored search: an overview of the concept, history, and technology," *Int. J. Electron. Bus.*, vol. 6, no. 2, pp. 114–171, 2008, doi: 10.1504/IJEB.2008.018068.

[58] A. Strzelecki, „Zarządzanie reputacją w wyszukiwarkach internetowych", *Pr. Nauk. Uniw. Ekon. we Wrocławiu. Inform. Ekon.*, t. 18, nr 119, *Systemy informacyjne w zarządzaniu. Zastosowania praktyczne*, str. 303–310, 2010.

[59] F. Etro, "Advertising and search engines. A model of leadership in search advertising," *Res. Econ.*, vol. 67, no. 1, pp. 25–38, Mar. 2013, doi: 10.1016/j.rie.2012.10.001.

[60] G. Bleoju, A. Capatina, E. Rancati, and N. Lesca, "Exploring organizational propensity toward inbound–outbound marketing techniques adoption: The case of

pure players and click and mortar companies," *J. Bus. Res.*, vol. 69, no. 11, pp. 5524–5528, Nov. 2016, doi: 10.1016/j.jbusres.2016.04.165.

[61] D. Lewandowski, S. Sünkler, and N. Yagci, "The influence of search engine optimization on Google's results," in *13th ACM Web Science Conference 2021*, Jun. 2021, pp. 12–20, doi: 10.1145/3447535.3462479.

[62] M. Nagpal and J.A. Petersen, "Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance?," *J. Retail.*, vol. 97, no. 4, pp. 746–763, Dec. 2021, doi: 10.1016/j.jretai.2020.12.002.

[63] M. Vállez and A. Ventura, "Analysis of the SEO visibility of university libraries and how they impact the web visibility of their universities," *J. Acad. Librariansh.*, vol. 46, no. 4, Jul. 2020, doi: 10.1016/j.acalib.2020.102171.

[64] A. Hora, "Characterizing top ranked code examples in Google," *J. Syst. Softw.*, vol. 178, p. 110971, Aug. 2021, doi: 10.1016/j.jss.2021.110971.

[65] R. Ferraz, "Exploring Web Attributes Related to Image Accessibility and their Impact on Search Engine Indexing," *Procedia Comput. Sci.*, vol. 67, pp. 171–184, 2015, doi: 10.1016/j.procs.2015.09.261.

[66] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," *Tour. Manag.*, vol. 31, no. 2, pp. 179–188, Apr. 2010, doi: 10.1016/j.tourman.2009.02.016.

[67] T. Mavridis and A.L. Symeonidis, "Semantic analysis of web documents for the generation of optimal content," *Eng. Appl. Artif. Intell.*, vol. 35, pp. 114–130, Oct. 2014, doi: 10.1016/j.engappai.2014.06.008.

[68] T.G. Shipley and A. Bowker, "Investigating Websites and Webpages," in *Investigating Internet Crimes*, Elsevier, 2014, pp. 293–314, doi: 10.1016/B978-0-12-407817-8.00013-8.

[69] M. Abdou, S. AbdelGaber, and M. Farhan, "A semi-automated framework for semantically annotating web content," *Futur. Gener. Comput. Syst.*, vol. 81, pp. 94–102, Apr. 2018, doi: 10.1016/j.future.2017.11.008.

[70] G. Egri and C. Bayrak, "The Role of Search Engine Optimization on Keeping the User on the Site," *Procedia Comput. Sci.*, vol. 36, no. C, pp. 335–342, 2014, doi:

10.1016/j.procs.2014.09.102.

[71] F. Fahimnia and M. Eltemasi, “Comparative analysis of Iranian medical academic libraries websites the base Google SEO component,” *J. Acad. Librariansh.*, vol. 47, no. 4, p. 102354, Jul. 2021, doi: 10.1016/j.acalib.2021.102354.

[72] V. Luque Centeno, C. Delgado Kloos, J. Arias Fisteus, and L. Álvarez Álvarez, “Web Accessibility Evaluation Tools: A Survey and Some Improvements,” *Electron. Notes Theor. Comput. Sci.*, vol. 157, no. 2, pp. 87–100, May 2006, doi: 10.1016/j.entcs.2005.12.048.

[73] S. Zhang and N. Cabage, “Search Engine Optimization: Comparison of Link Building and Social Sharing,” *J. Comput. Inf. Syst.*, vol. 57, no. 2, pp. 148–159, Apr. 2017, doi: 10.1080/08874417.2016.1183447.

[74] Veglis and Giomelakis, “Search Engine Optimization,” *Futur. Internet*, vol. 12, no. 1, p. 6, Dec. 2019, doi: 10.3390/fi12010006.

[75] A. Noruzi, “A Study of HTML Title Tag Creation Behavior of Academic Web Sites,” *J. Acad. Librariansh.*, vol. 33, no. 4, pp. 501–506, Jul. 2007, doi: 10.1016/j.acalib.2007.03.008.

[76] T.C. Craven, “Variations in use of meta tag descriptions by Web pages in different languages,” *Inf. Process. Manag.*, vol. 40, no. 3, pp. 479–493, May 2004, doi: 10.1016/S0306-4573(02)00121-8.

[77] M. Pérez-Montoro and L. Codina, “The Essentials of Search Engine Optimization,” in *Navigation Design and SEO for Content-Intensive Websites*, Elsevier, 2017, pp. 109–124, doi: 10.1016/B978-0-08-100676-4.00005-5.

[78] M. Pérez-Montoro and L. Codina, “SEO for Content-Intensive Sites,” in *Navigation Design and SEO for Content-Intensive Websites*, Elsevier, 2017, pp. 125–137, doi: 10.1016/b978-0-08-100676-4.00006-7.

[79] N. Yalçın and U. Köse, “What is search engine optimization: SEO?,” *Procedia – Soc. Behav. Sci.*, vol. 9, pp. 487–493, 2010, doi: 10.1016/j.sbspro.2010.12.185.

[80] K. Choudhari and V.K. Bhalla, “Video Search Engine Optimization Using Keyword and Feature Analysis,” *Procedia Comput. Sci.*, vol. 58, pp. 691–697, 2015, doi: 10.1016/j.procs.2015.08.089.

[81] J.B. Killoran, “How to Use Search Engine Optimization Techniques to Increase Website Visibility,” *IEEE Trans. Prof. Commun.*, vol. 56, no. 1, pp. 50–66, Mar. 2013, doi: 10.1109/TPC.2012.2237255.

[82] M. Pérez-Montoro and L. Codina, “Mobile Web and SEO,” in *Navigation Design and SEO for Content-Intensive Websites*, Elsevier, 2017, pp. 139–151, doi: 10.1016/B978-0-08-100676-4.00007-9.

[83] A. Ismail and K.S. Kuppusamy, “Web accessibility investigation and identification of major issues of higher education websites with statistical measures: A case study of college websites,” *J. King Saud Univ. – Comput. Inf. Sci.*, Apr. 2019, doi: 10.1016/j.jksuci.2019.03.011.

[84] A. Ganapathy, “Friendly URLs in the CMS and Power of Global Ranking with Crawlers with Added Security,” *Eng. Int.*, vol. 5, no. 2, pp. 87–96, 2017, doi: 10.18034/ei.v5i2.541.

[85] A. Ismail and F. Abdallah, “A Survey on Search Engine Optimization (SEO),” *Int. J. Comput. Commun. Instrum. Eng.*, vol. 4, no. 2, Oct. 2017, doi: 10.15242/IJCCIE.AE0417136.

[86] A. Erdmann and J.M. Ponzoa, “Digital inbound marketing: Measuring the economic performance of grocery e-commerce in Europe and the USA,” *Technol. Forecast. Soc. Change*, vol. 162, p. 120373, Jan. 2021, doi: 10.1016/j.techfore.2020.120373.

[87] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, “Approximating PageRank from In-Degree,” in *Algorithms and Models for the Web-Graph*, vol. 4936 LNCS, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 59–71, doi: 10.1007/978-3-540-78808-9_6.

[88] O.-R. Jeong, J. Oh, D.-J. Kim, H. Lyu, and W. Kim, “Determining the titles of Web pages using anchor text and link analysis,” *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4322–4329, Jul. 2014, doi: 10.1016/j.eswa.2013.12.033.

[89] C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, “Important Factors for Improving Google Search Rank,” *Futur. Internet*, vol. 11, no. 2, p. 32, Jan. 2019, doi: 10.3390/fi11020032.

[90] D.W. Kim, P. Yan, and J. Zhang, “Detecting fake anti-virus software distribution webpages,” *Comput. Secur.*, vol. 49, pp. 95–106, Mar. 2015, doi: 10.1016/j.cose .2014.11.008.

[91] R. Wang, Y. Zhu, J. Tan, and B. Zhou, “Detection of malicious web pages based on hybrid analysis,” *J. Inf. Secur. Appl.*, vol. 35, pp. 68–74, Aug. 2017, doi: 10.1016/j.jisa.2017.05.008.

[92] C. Wright, “Auditing Web-Based Applications,” in *The IT Regulatory and Standards Compliance Handbook*, Elsevier, 2008, pp. 515–560, doi: 10.1016/B978-1-59749-266-9.00018-7.

[93] R.-W. Bello and F.N. Otobo, “Conversion of Website Users to Customers-The Black Hat SEO Technique,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 8, no. 6, p. 29, Jun. 2018, doi: 10.23956/ijarcsse.v8i6.714.

[94] R. Aswani, A.K. Kar, P.V. Ilavarasan, and Y.K. Dwivedi, “Search engine marketing is not all gold: Insights from Twitter and SEOClerks,” *Int. J. Inf. Manage.*, vol. 38, no. 1, pp. 107–116, Feb. 2018, doi: 10.1016/j.ijinfomgt .2017.07.005.

[95] S. Schultheiß and D. Lewandowski, “‘Outside the industry, nobody knows what we do’ SEO as seen by search engine optimizers and content providers,” *J. Doc.*, vol. 77, no. 2, pp. 542–557, Dec. 2020, doi: 10.1108/JD-07-2020-0127.

[96] Y. Sun, Z. Zhuang, I.G. Councill, and C.L. Giles, “Determining Bias to Search Engines from Robots.txt,” in *IEEE/WIC/ACM International Conference on Web Intelligence (WI’07)*, Nov. 2007, pp. 149–155, doi: 10.1109/WI.2007.98.

[97] A. Tomič and M. Šupín, “Increasing website traffic of woodworking company using digital marketing methods,” 2019.

[98] D. Wijaya, B. Daniawan, and Y. Gunawan, “Search Engine Optimization (SEO) As A Promotional Media On Google Search,” *Bit-Tech*, vol. 4, no. 1, 2021, doi: https://doi.org/10.32877/bt.v4i1.237.

[99] M. Nen, V. Popa, and A. Scurtu, “The Computer Management – SEO Audit,” vol. 18, no. 3, p. 297, 2017.

[100] D. Gek, V. Kukartsev, V. Tynchenko, A. Bondarev, M. Pokushko, and N.

Dalisova, “The problem of SEO promotion for the organization’s web representation,” *SHS Web Conf.*, vol. 69, p. 00122, Oct. 2019, doi: 10.1051/shsconf/20196900122.

[101] L. Moreno and P. Martinez, “Overlapping factors in search engine optimization and web accessibility,” *Online Inf. Rev.*, vol. 37, no. 4, pp. 564–580, Aug. 2013, doi: 10.1108/OIR-04-2012-0063.

[102] A. Shenoy and A. Prabhu, “SEO Hub: Utilities and Toolsets,” in *Introducing SEO*, Berkeley, CA: Apress, 2016, pp. 103–117, doi: 10.1007/978-1-4842-1854-9_10.

[103] S. Katumba and S. Coetzee, “Employing Search Engine Optimization (SEO) Techniques for Improving the Discovery of Geospatial Resources on the Web,” *ISPRS Int. J. Geo-Information*, vol. 6, no. 9, p. 284, Sep. 2017, doi: 10.3390/ijgi6090284.

[104] K. Król, „Stopień optymalizacji witryn internetowych obiektów turystyki wiejskiej dla wyszukiwarek internetowych”, *Rocz. Nauk. Ekon. Rol. i Rozw. Obsz. Wiej.*, vol. 105, no. 2, str. 110–121, grudzień 2018, doi: 10.22630/RNR.2018.105.2.20.

[105] I. Gregurec and P. Grd, “Search Engine Optimization (SEO): Website analysis of selected faculties in Croatia,” In *Proceedings of Central European Conference on Information and Intelligent Systems*, pp. 211–218, Varaždin, Croatia, 2012 [online]. Available: https://www.researchgate.net/publication/267404649.

[106] N. Vankov, “Significance, Role and Principles of Website SEO for Digital Entrepreneurship – the Battle for the Top Google Rankings,” Jun. 2017.

[107] M.P. Evans, “Analysing Google rankings through search engine optimization data,” *Internet Res.*, vol. 17, no. 1, pp. 21–37, Feb. 2007, doi: 10.1108/ 10662240710730470.

[108] C. Weiqing, H. Yangyang, Y. Qiaofeng, and C. Jiajia, “Measuring web page complexity by analyzing TCP flows and HTTP headers,” *J. China Univ. Posts Telecommun.*, vol. 24, no. 6, pp. 1–13, Dec. 2017, doi: 10.1016/S1005-8885(17)60237-1.

[109] H. Lee, “Using Google Search Console for Audience Insights,” in *Hands On With Google® Data Studio*, Wiley, 2020, pp. 135–171, doi: 10.1002/

9781119616238.ch6.

[110] A. Strzelecki, “Website removal from search engines due to copyright violation,” *Aslib J. Inf. Manag.*, vol. 71, no. 1, pp. 54–71, Jan. 2019, doi: 10.1108/AJIM-05-2018-0108.

[111] M. Molodchik, S. Paklina, and P. Parshakov, “Digital relational capital of a company,” *Meditari Account. Res.*, vol. 26, no. 3, pp. 443–462, Sep. 2018, doi: 10.1108/MEDAR-08-2017-0186.

[112] M. Mirkovic, “Examining the web presence of Croatian wine festivals,” *Proc. Int. Sci. Conf. Juraj Dobrila Univ. Pula, Dep. Econ. Tour.*, pp. 203–223, 2019.

[113] K.K. Nambiar, “Theory of search engines,” *Comput. Math. with Appl.*, vol. 42, no. 12, pp. 1523–1526, Dec. 2001, doi: 10.1016/S0898-1221(01)00259-0.

[114] J. Bar-Ilan, “Comparing rankings of search results on the Web,” *Inf. Process. Manag.*, vol. 41, no. 6, pp. 1511–1519, Dec. 2005, doi: 10.1016/j.ipm .2005.03.008.

[115] J.M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999, doi: 10.1145/324133.324140.

[116] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” *Comput. Networks ISDN Syst.*, vol. 30, no. 1–7, pp. 107–117, Apr. 1998, doi: 10.1016/S0169-7552(98)00110-X.

[117] D. Rafiei and A.O. Mendelzon, “What is this page known for? Computing Web page reputations,” *Comput. Networks*, vol. 33, no. 1–6, pp. 823–835, Jun. 2000, doi: 10.1016/S1389-1286(00)00078-5.

[118] N. Duhan, A.K. Sharma, and K.K. Bhatia, “Page Ranking Algorithms: A Survey,” in *2009 IEEE International Advance Computing Conference*, Mar. 2009, pp. 1530–1537, doi: 10.1109/IADCC.2009.4809246.

[119] A. Strzelecki, „Autorytatywne i eksperckie strony źródłem rzetelnych wyników w wyszukiwarkach internetowych”, w: *Informatyka dla przyszłości*, 2008, str. 193–201.

[120] S. Lakshminarayana, “Categorization of web pages – Performance enhancement

to search engine," *Knowledge-Based Syst.*, vol. 22, no. 1, pp. 100–104, Jan. 2009, doi: 10.1016/j.knosys.2008.07.006.

[121] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1998.

[122] A. Strzelecki, „Zastosowania algorytmu PageRank w wyszukiwaniu," *Pr. Nauk. Ekon. w Katowicach*, str. 347–353, 2008.

[123] M. Sree Vani, R. Bhramaramba, D. Vasumati, and O. Yaswanth Babu, "Classifying Web Spam using Block-based TrustRank," vol. 2, no. 4, 2012 [online]. Available: https://www.researchgate.net/publication/290899468.

[124] V. Krishnan and R. Raj, "Web Spam Detection with Anti-Trust Rank," pp. 37–44, 2006.

[125] M. Chatterjee and A.S. Namin, "A fuzzy Dempster–Shafer classifier for detecting Web spams," *J. Inf. Secur. Appl.*, vol. 59, p. 102793, Jun. 2021, doi: 10.1016/j.jisa.2021.102793.

[126] X. Zhuang, Y. Zhu, Q. Peng, and F. Khurshid, "Using deep belief network to demote web spam," *Futur. Gener. Comput. Syst.*, vol. 118, pp. 94–106, May 2021, doi: 10.1016/j.future.2020.12.023.

[127] K. Goldsmith, "If It Doesn't Exist on the Internet, It Doesn't Exist," Aug. 27, 2005 [online]. Available: http://writing.upenn.edu/epc/authors/goldsmith/if_it_doesnt_exist.html [accessed Feb. 09, 2022].

[128] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Comput. Networks*, vol. 50, no. 10, pp. 1448–1463, Jul. 2006, doi: 10.1016/j.comnet.2005.10.020.

[129] A.K. Sedigh and M. Roudaki, "Identification of the dynamics of the Google's ranking algorithm," 2003.

[130] A. Bifet, C. Castillo, P.A. Chirita, and I. Weber, "An Analysis of Factors Used in Search Engine Ranking," 2005.

[131] A. Strzelecki, "Google Medical Update: Why Is the Search Engine Decreasing Visibility of Health and Medical Information Websites?," *Int. J. Environ. Res.*

Public Health, vol. 17, no. 4, p. 1160, Feb. 2020, doi: 10.3390/ijerph17041160.

[132] M. Ismail, I. Jamil, and R. Jamil, "Using SEO techniques Google Panda to Improve the Website Ranking," *International Journal of Engineering Works*, vol. 1, pp. 6–9, 2014, doi: https://doi.org/10.5281/zenodo.15742.

[133] D. Schubert, "Influence of Mobile-friendly Design to Search Results on Google Search," *Procedia – Soc. Behav. Sci.*, vol. 220, pp. 424–433, May 2016, doi: 10.1016/j.sbspro.2016.05.517.

[134] A. Patil, J. Pamnani, and D. Pawade, "Comparative Study Of Google Search Engine Optimization Algorithms: Panda, Penguin and Hummingbird," in *2021 6th International Conference for Convergence in Technology (I2CT)*, Apr. 2021, pp. 1–5, doi: 10.1109/I2CT51068.2021.9418074.

[135] Z. Chen, Q. Chen, J. Li, Z. Li, and L. Chen, "A probabilistic ranking framework for web-based relational data imputation," *Inf. Sci. (Ny).*, vol. 355–356, pp. 152–168, Aug. 2016, doi: 10.1016/j.ins.2016.03.036.

[136] D. Kumar Sharma and A.K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms," *IJCSE) Int. J. Comput. Sci. Eng.*, vol. 02, no. 08, pp. 2670–2676, 2010.

[137] N. Tyagi and S. Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page," *Int. J. Soft Comput. Eng.*, no. 2, pp. 2231–2307, 2012.

[138] R. Lempel and S. Moran, "SALSA: The Stochastic Approach for Link-Structure Analysis," *ACM Trans. Inf. Syst.*, vol. 19, no. 2, pp. 131–160, 2001.

[139] D. Fuentes-Lorenzo, N. Fernández, J.A. Fisteus, and L. Sánchez, "Improving large-scale search engines with semantic annotations," *Expert Syst. Appl.*, vol. 40, no. 6, pp. 2287–2296, May 2013, doi: 10.1016/j.eswa.2012.10.042.

[140] H. Wang, Y. Li, and K. Guo, "Countering Web Spam of Link-based Ranking Based on Link Analysis," *Procedia Eng.*, vol. 23, pp. 310–315, 2011, doi: 10.1016/j.proeng.2011.11.2507.

[141] M. Kale and P.S. Thilagam, "DYNA-RANK: Efficient Calculation and Updation of PageRank," in *2008 International Conference on Computer Science and*

Information Technology, Aug. 2008, pp. 808–812, doi: 10.1109/ICCSIT.2008.118.

[142] P.V. Vidya, P.C.R. Raj, and V. Jayan, "Web Page Ranking Using Multilingual Information Search Algorithm – A Novel Approach," *Procedia Technol.*, vol. 24, pp. 1240–1247, 2016, doi: 10.1016/j.protcy.2016.05.102.

[143] I. Hernández, C.R. Rivero, D. Ruiz, and R. Corchuelo, "CALA: An unsupervised URL-based web page classification system," *Knowledge-Based Syst.*, vol. 57, pp. 168–180, Feb. 2014, doi: 10.1016/j.knosys.2013.12.019.

[144] I. Rasekh, "A New Competitive Intelligence-based Strategy for Web Page Search," *Procedia Comput. Sci.*, vol. 62, pp. 450–456, 2015, doi: 10.1016/j.procs.2015.08.505.

[145] N. Dai and B.D. Davison, "Freshness matters," in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval – SIGIR '10*, 2010, p. 114, doi: 10.1145/1835449.1835471.

[146] G. Feng, T.-Y. Liu, X.-D. Zhang, T. Qin, B. Gao, and W.-Y. Ma, "Level-Based Link Analysis," in *Lecture Notes in Computer Science*, vol. 3399, Springer Verlag, 2005, pp. 183–194, doi: 10.1007/978-3-540-31849-1_19.

[147] A. Dong *et al.*, "Towards recency ranking in web search," in *Proceedings of the third ACM international conference on Web search and data mining – WSDM '10*, 2010, p. 11, doi: 10.1145/1718487.1718490.

[148] M. Zhukovskiy, D. Vinogradov, G. Gusev, P. Serdyukov, and A. Raigorodskii, "Recency-sensitive model of web page authority," in *Proceedings of the 21st ACM international conference on Information and knowledge management – CIKM '12*, 2012, p. 2627, doi: 10.1145/2396761.2398708.

[149] B. Manaskasemsak, A. Rungsawang, and H. Yamana, "Time-weighted web authoritative ranking," *Inf. Retr. Boston.*, vol. 14, no. 2, pp. 133–157, Apr. 2011, doi: 10.1007/s10791-010-9138-4.

[150] M. Richardson, A. Prakash, and E. Brill, "Beyond PageRank," in *Proceedings of the 15th international conference on World Wide Web – WWW '06*, 2006, p. 707, doi: 10.1145/1135777.1135881.

[151] K. Berberich, M. Vazirgiannis, and G. Weikum, "T-Rank: Time-Aware Authority

Ranking," in *LNCS*, vol. 3243, 2004, pp. 131–142, doi: 10.1007/978-3-540-30216-2_11.

[152] S. Hariharan, S. Dhanasekar, and K. Desikan, "Reachability Based Web Page Ranking Using Wavelets," *Procedia Comput. Sci.*, vol. 50, pp. 157–162, 2015, doi: 10.1016/j.procs.2015.04.078.

[153] A.M. Zareh Bidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages," *Inf. Process. Manag.*, vol. 44, no. 2, pp. 877–892, Mar. 2008, doi: 10.1016/j.ipm.2007.06.004.

[154] A. Kritikopoulos, M. Sideri, and I. Varlamis, "Wordrank: A Method for Ranking Web Pages Based on Content Similarity," in *24th British National Conference on Databases (BNCOD'07)*, 2007, pp. 92–100, doi: 10.1109/BNCOD.2007.24.

[155] P. O'Brien, T. Abou-Assaleh, T. Das, W. Gao, Y. Miao, and Z. Zhen, "A link-based ranking scheme for focused search," in *Proceedings of the 16th international conference on World Wide Web – WWW '07*, 2007, p. 1125, doi: 10.1145/1242572.1242727.

[156] A.M. Zareh Bidoki, P. Ghodsnia, N. Yazdani, and F. Oroumchian, "A3CRank: An adaptive ranking method based on connectivity, content and click-through data," *Inf. Process. Manag.*, vol. 46, no. 2, pp. 159–169, Mar. 2010, doi: 10.1016/j.ipm.2009.12.005.

[157] M. Almulla, H. Yahyaoui, and K. Al-Matori, "A new fuzzy hybrid technique for ranking real world Web services," *Knowledge-Based Syst.*, vol. 77, pp. 1–15, Mar. 2015, doi: 10.1016/j.knosys.2014.12.021.

[158] V.X. Tran, H. Tsuji, and R. Masuda, "A new QoS ontology and its QoS-based ranking algorithm for Web services," *Simul. Model. Pract. Theory*, vol. 17, no. 8, pp. 1378–1398, Sep. 2009, doi: 10.1016/j.simpat.2009.06.010.

[159] Y. Du and Y. Hai, "Semantic ranking of web pages based on formal concept analysis," *J. Syst. Softw.*, vol. 86, no. 1, pp. 187–197, Jan. 2013, doi: 10.1016/j.jss.2012.07.040.

[160] A.K. Singh and R. Kumar, "A Comparative Study of Page Ranking Algorithms for Information Retrieval," 2009.

[161] Z. Gyongyi, H. Garciamolina, and J. Pedersen, “Combating Web Spam with TrustRank,” in *Proceedings 2004 VLDB Conference*, Elsevier, 2004, pp. 576–587, doi: 10.1016/B978-012088469-8/50052-8.

[162] S. Giannoulakis and N. Tsapatsoulis, “Filtering Instagram Hashtags Through Crowdtagging and the HITS Algorithm,” *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 3, pp. 592–603, Jun. 2019, doi: 10.1109/TCSS.2019.2914080.

[163] B.Q. Hung, M. Otsubo, Y. Hijikata, and S. Nishida, “HITS algorithm improvement using semantic text portion,” *Web Intell. Agent Syst. An Int. J.*, vol. 8, no. 2, pp. 149–164, 2010, doi: 10.3233/WIA-2010-0184.

[164] M. Agosti and L. Pretto, “A Theoretical Study of a Generalized Version of Kleinberg’s HITS Algorithm,” *Inf. Retr. Boston.*, vol. 8, no. 2, pp. 219–243, Apr. 2005, doi: 10.1007/s10791-005-5660-1.

[165] S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida, “Analysis and improvement of HITS algorithm for detecting Web communities,” *Syst. Comput. Japan*, vol. 35, no. 13, pp. 32–42, Nov. 2004, doi: 10.1002/scj.10425.

[166] A. Farahat, T. LoFaro, J.C. Miller, G. Rae, and L.A. Ward, “Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization,” *SIAM J. Sci. Comput.*, vol. 27, no. 4, pp. 1181–1201, Jan. 2006, doi: 10.1137/S1064827502412875.

[167] X. Zhong, Y. Zhang, D. Yan, Q. Wu, Y.T. Yan, and W. Li, “Recommendations for Mobile Apps Based on the HITS Algorithm Combined With Association Rules,” *IEEE Access*, vol. 7, pp. 105572–105582, 2019, doi: 10.1109/ACCESS .2019.2931756.

[168] T. Mavridis and A.L. Symeonidis, “Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms,” *Eng. Appl. Artif. Intell.*, vol. 41, pp. 75–91, May 2015, doi: 10.1016/j.engappai .2015.02.002.

[169] R. Aswani, S.P. Ghrera, S. Chandra, and A.K. Kar, “Outlier Detection Among Influencer Blogs Based on off-Site Web Analytics Data,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10595 LNCS, Springer Verlag, 2017, pp.

251–260, doi: 10.1007/978-3-319-68557-1_23.

[170] A. Brahma and R. Dutta, “Role of Social Media and E-Commerce for Business Entrepreneurship,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 01–18, Nov. 2020, doi: 10.32628/CSEIT206559.

[171] M. Bugliesi, S. Calzavara, and R. Focardi, “Formal methods for web security,” *J. Log. Algebr. Methods Program.*, vol. 87, pp. 110–126, Feb. 2017, doi: 10.1016/j.jlamp.2016.08.006.

[172] R. Ramakrishnan and A. Kaur, “An empirical comparison of predictive models for web page performance,” *Inf. Softw. Technol.*, vol. 123, p. 106307, Jul. 2020, doi: 10.1016/j.infsof.2020.106307.

[173] G. Palaniappan, S.S, B. Rajendran, Sanjay, S. Goyal, and B.S. Bindhumadhava, “Malicious Domain Detection Using Machine Learning On Domain Name Features, Host-Based Features and Web-Based Features,” *Procedia Comput. Sci.*, vol. 171, pp. 654–661, 2020, doi: 10.1016/j.procs.2020.04.071.

[174] T. Eggendorfer and J. Keller, “Preventing spam by dynamically obfuscating email-addresses,” *Informatikzentrum*, 2005.

[175] A. Szewczyk, “Internet marketing in social media,” *Zesz. Nauk. Uniw. Szczecińskiego. Stud. Inform.*, vol. 36, pp. 119–133, 2015, doi: 10.18276/si .2015.36-09.

[176] B. Oberer and A. Erkollar, “Social Media Integration in Higher Education. Cross-Course Google Plus Integration Shown in the Example of a Master’s Degree Course in Management,” *Procedia – Soc. Behav. Sci.*, vol. 47, pp. 1888–1893, 2012, doi: 10.1016/j.sbspro.2012.06.918.

[177] J.K. Nurminen, A.J.R. Meyn, E. Jalonen, Y. Raivio, and R. Garcıa Marrero, “P2P media streaming with HTML5 and WebRTC,” in *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2013, pp. 63–64, doi: 10.1109/INFCOMW.2013.6970739.

[178] R. Meusel, P. Petrovski, and C. Bizer. “The webdatacommons microdata, rdfa and microformat dataset series,” *International Semantic Web Conference*, Springer, Cham, 2014, p. 277–292, doi: 10.1007/978-3-319-11964-9_18.

[179] S. Chen, D. Hong and V.Y. Shen, “An Experimental Study on Validation Problems with Existing HTML Webpages,” in *Proceedings of the 2005 International Conference on Internet Computing*, ICOMP'05, Las Vegas, 2005 [online]. Available: https://www.researchgate.net/publication/220968242.

[180] W. Xu, F. Zhang, and S. Zhu, “The power of obfuscation techniques in malicious JavaScript code: A measurement study,” in *2012 7th International Conference on Malicious and Unwanted Software*, Oct. 2012, pp. 9–16, doi: 10.1109/MALWARE.2012.6461002.

[181] C. Wang, J. Lu, and G. Zhang, “Mining key information of web pages: A method and its application,” *Expert Syst. Appl.*, vol. 33, no. 2, pp. 425–433, Aug. 2007, doi: 10.1016/j.eswa.2006.05.017.

[182] J.C. Roldán, P. Jiménez, and R. Corchuelo, “On extracting data from tables that are encoded using HTML,” *Knowledge-Based Syst.*, vol. 190, p. 105157, Feb. 2020, doi: 10.1016/j.knosys.2019.105157.

[183] S. Rane and W. Sun, “Privacy preserving string comparisons based on Levenshtein distance,” in *2010 IEEE International Workshop on Information Forensics and Security*, Dec. 2010, pp. 1–6, doi: 10.1109/WIFS.2010.5711449.

[184] A. Barbar and A. Ismail, “Search Engine Optimization (SEO) for Websites,” in *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, Apr. 2019, vol. Part F148262, pp. 51–55, doi: 10.1145/3323933.3324072.

[185] A. Hussien, “Factors Affect Search Engine Optimization”, *Int. J. Comput. Sci. Netw. Secur.*, 2014, vol. 14, no. 9, pp. 28–33 [online]. Available: http://paper.ijcsns.org/07_book/201409/20140904.pdf.

[186] P.S. Vadivu, P. Sumathy, and A. Vadivel, “Image Retrieval from WWW using Attributes in HTML TAGs,” *Procedia Technol.*, vol. 6, pp. 509–516, 2012, doi: 10.1016/j.protcy.2012.10.061.

[187] G. Matošević, “Measuring the Utilization of On-Page Search Engine Optimization in Selected Domain”, J. inf. organ. sci., vol. 39, no. 2, Dec. 2015 [online]. Available: https://jios.foi.hr/index.php/jios/article/view/974.

[188] M. Shema, “HTML Injection & Cross-Site Scripting (XSS),” in *Hacking Web Apps*, Elsevier, 2012, pp. 23–78, doi: 10.1016/B978-1-59-749951-4.00002-3.

[189] Y. Xue *et al.*, “Web page title extraction and its application,” *Inf. Process. Manag.*, vol. 43, no. 5, pp. 1332–1347, Sep. 2007, doi: 10.1016/j.ipm.2006.11.007.

[190] A. Gandour and A. Regolini, “Web site search engine optimization: a case study of Fragfornet,” vol. 28, no. 6, 2011, doi: 10.1108/07419051111173874ï.

[191] N. Solihin, “Search engine optimization: a survey of current best practices,” 2013, *Technical Library*, p. 151 [online]. Available: https://scholarworks.gvsu.edu/cistechlib/151.

[192] R. Tabarés, “HTML5 and the evolution of HTML; tracing the origins of digital platforms,” *Technol. Soc.*, vol. 65, p. 101529, May 2021, doi: 10.1016/j.techsoc.2021.101529.

[193] A.S. Bozkir and E. Akcapinar Sezer, “Layout-based computation of web page similarity ranks,” *Int. J. Hum. Comput. Stud.*, vol. 110, pp. 95–114, Feb. 2018, doi: 10.1016/j.ijhcs.2017.10.008.

[194] N. Batalas, V.-J. Khan, and P. Markopoulos, “Executable HTML,” *SoftwareX*, vol. 14, p. 100691, Jun. 2021, doi: 10.1016/j.softx.2021.100691.

[195] W. Yuyin and C. Yuhang, “Influence of virtual imaging technology based on html5 technology on digital painting,” *Microprocess. Microsyst.*, vol. 82, p. 103855, Apr. 2021, doi: 10.1016/j.micpro.2021.103855.

[196] A.I.C. Mohideen, M. Rajiullah, R. Secchi, G. Fairhurst, A. Brunstrom, and F. Weinrank, “Evaluating the impact of transport mechanisms on web performance for effective web access,” *J. Netw. Comput. Appl.*, vol. 137, pp. 25–34, Jul. 2019, doi: 10.1016/j.jnca.2019.04.006.

[197] D. Raggett, “Clean up your Web pages with HP’s HTML Tidy,” *Comput. Networks ISDN Syst.*, vol. 30, no. 1–7, pp. 730–732, Apr. 1998, doi: 10.1016/S0169-7552(98)00122-6.

[198] C. Fu, “Exploration of Web front-end development technology and optimization direction,” 2016, doi: 10.2991/icence-16.2016.35.

[199] L.F.A. Aristizabal and N. Dario Duque Mendez, “SEO (Search Engine Optimization) schema application for websites with an emphasis on optimizing pages developed in flash,” in *2012 7th Colombian Computing Congress (CCC)*, Oct. 2012, pp. 1–6, doi: 10.1109/ColombianCC.2012.6398011.

[200] R.T. Gutiérrez, “Understanding the role of digital commons in the web; The making of HTML5,” *Telemat. Informatics*, vol. 35, no. 5, pp. 1438–1449, Aug. 2018, doi: 10.1016/j.tele.2018.03.013.

[201] H. Artail and K. Fawaz, “A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations,” *Data Knowl. Eng.*, vol. 66, no. 2, pp. 326–337, Aug. 2008, doi: 10.1016/j.datak.2008.04.003.

[202] W. Kanwal, “Exploring Search Engine Optimization (SEO) Techniques for Dynamic Websites,” Department of Computer Science and Engineering Chalmers University of Technology, June 2011 [online]. Available: http://www.diva-portal .org/smash/get/diva2:832232/FULLTEXT01.pdf.

[203] Dr. Birajkumar V. Patel, Dr. Raina D. Gaharwar, “Search Engine Optimization (SEO) using HTML Meta-Tags,” *International Journal of Scientific Research in Science and Technology* (IJSRST), vol. 4, Is. 9, pp. 298–302, July-August 2018. Print ISSN: 2395-6011, online ISSN: 2395-602X. Available: https://ijsrst.com/ IJSRST184976.

[204] N. Kumar et al., “Search engine optimization: is your website optimized with correct SEO techniques?,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 2, no. 6, pp. 2321–9653, 2014.

[205] T.S. Dronova and Y.Y. Trygub, “Increasing the travel agency’s leading positions by optimizing its website,” *Eur. J. Manag. Issues*, vol. 28, no. 3, pp. 81–91, Sep. 2020, doi: 10.15421/192008.

[206] G. Kumar and R.K. Paul, “Literature Review on On-Page & Off-Page SEO for Ranking Purpose,” *United Int. J. Res. Technol.*, vol. 1, no. 6, pp. 30–34, Apr. 2020 [online]. Available: https://uijrt.com/articles/v1/i6/UIJRTV1I60005.pdf.

[207] V.N. Gudivada, D. Rao, and J. Paris, “Understanding Search-Engine Optimization,” *Computer (Long. Beach. Calif).*, vol. 48, no. 10, pp. 43–52, Oct. 2015, doi: 10.1109/MC.2015.297.

[208] D. Giomelakis and A. Veglis, “Employing Search Engine Optimization Techniques in Online News Articles,” *Stud. Media Commun.*, vol. 3, no. 1, Mar. 2015, doi: 10.11114/smc.v3i1.683.

[209] O. Modiri, D. Guha, N.M. Alotaibi, G.M. Ibrahim, N. Lipsman, and A. Fallah, “Readability and quality of wikipedia pages on neurosurgical topics,” *Clin. Neurol. Neurosurg.*, vol. 166, pp. 66–70, Mar. 2018, doi: 10.1016/j.clineuro.2018.01.021.

[210] R. Rai, A. Landsberg, A. Nguyen, and S.M. Wiseman, “Online educational materials for appendectomy patients have good quality but poor readability,” *Am. J. Surg.*, vol. 221, no. 6, pp. 1203–1210, Jun. 2021, doi: 10.1016/j.amjsurg. 2021.02.022.

[211] R.I. Zraick, M. Azios, M.M. Handley, M.L. Bellon-Harn, and V. Manchaiah, “Quality and readability of internet information about stuttering,” *J. Fluency Disord.*, vol. 67, p. 105824, Mar. 2021, doi: 10.1016/j.jfludis.2020.105824.

[212] K.R. Shetty, K. Wong, S. Hashemi, A. Shetty, and J.R. Levi, “Transoral robotic surgery: Differences between online information and academic literature,” *Am. J. Otolaryngol.*, vol. 41, no. 4, p. 102395, Jul. 2020, doi: 10.1016/j.amjoto .2020.102395.

[213] P.K. Ojha, A. Ismail, and K. Kundumani Srinivasan, “Perusal of readability with focus on web content understandability,” *J. King Saud Univ. – Comput. Inf. Sci.*, vol. 33, no. 1, pp. 1–10, Jan. 2021, doi: 10.1016/j.jksuci.2018.03.007.

[214] M. Duka, “Elliptic-curve cryptography (ECC) and Argon2 algorithm in PHP using OpenSSL and Sodium libraries,” *Inform. Autom. Pomiary w Gospod. i Ochr. Środowiska*, vol. 10, no. 3, Sep. 2020, doi: 10.35784/iapgos.897.

[215] T.C. Du, F. Li, and I. King, “Managing knowledge on the Web – Extracting ontology from HTML Web,” *Decis. Support Syst.*, vol. 47, no. 4, pp. 319–331, Nov. 2009, doi: 10.1016/j.dss.2009.02.011.

[216] Y.-C. Wu, “Language independent web news extraction system based on text detection framework,” *Inf. Sci. (Ny).*, vol. 342, pp. 132–149, May 2016, doi: 10.1016/j.ins.2015.12.025.

[217] I. Lima, J. Cândido, and M. d’Amorim, “Practical detection of CMS plugin

conflicts in large plugin sets," *Inf. Softw. Technol.*, vol. 118, p. 106212, Feb. 2020, doi: 10.1016/j.infsof.2019.106212.

[218] N. Gali, R. Mariescu-Istodor, and P. Fränti, "Using linguistic features to automatically extract web page title," *Expert Syst. Appl.*, vol. 79, pp. 296–312, Aug. 2017, doi: 10.1016/j.eswa.2017.02.045.

[219] I.A. Ahmad Sabri, M. Man, W.A. W. Abu Bakar, and A.N. Mohd Rose, "Web Data Extraction Approach for Deep Web using WEIDJ," *Procedia Comput. Sci.*, vol. 163, pp. 417–426, 2019, doi: 10.1016/j.procs.2019.12.124.

[220] Y.-S. Kim and K.-H. Lee, "Extracting logical structures from HTML tables," *Comput. Stand. Interfaces*, vol. 30, no. 5, pp. 296–308, Jul. 2008, doi: 10.1016/j.csi.2007.08.006.

[221] M. Alpuente and D. Romero, "A Visual Technique for Web Pages Comparison," *Electron. Notes Theor. Comput. Sci.*, vol. 235, no. C, pp. 3–18, Apr. 2009, doi: 10.1016/j.entcs.2009.03.002.

[222] B. Christos, K. Giorgos, and M. Ioannis, "A web content manipulation technique based on page Fragmentation," *J. Netw. Comput. Appl.*, vol. 30, no. 2, pp. 563–585, Apr. 2007, doi: 10.1016/j.jnca.2006.01.005.

[223] Y. Zheng, X. Cheng, and K. Chen, "Filtering noise in Web pages based on parsing tree," *J. China Univ. Posts Telecommun.*, vol. 15, no. SUPPL., pp. 46–50, Sep. 2008, doi: 10.1016/S1005-8885(08)60153-3.

[224] D. Shen, Q. Yang, and Z. Chen, "Noise reduction through summarization for Web-page classification," *Inf. Process. Manag.*, vol. 43, no. 6, pp. 1735–1747, Nov. 2007, doi: 10.1016/j.ipm.2007.01.013.

[225] O.-W. Kwon and J.-H. Lee, "Text categorization based on k-nearest neighbor approach for Web site classification," *Inf. Process. Manag.*, vol. 39, no. 1, pp. 25–44, Jan. 2003, doi: 10.1016/S0306-4573(02)00022-5.

[226] G. Della Penna, D. Magazzeni, and S. Orefice, "Visual extraction of information from web pages," *J. Vis. Lang. Comput.*, vol. 21, no. 1, pp. 23–32, Feb. 2010, doi: 10.1016/j.jvlc.2009.06.001.

[227] M. Walesiak, „Zagadnienie oceny podobieństwa zbioru obiektów w czasie

w syntetycznych badaniach porównawczych", *Przegląd Stat.*, vol. 40, no. 1, str. 95–102, 1993.

7. Spis ilustracji

8. Listingi

8.1. Lista pseudokodów

8.2. Lista kodów źródłowych