

wpł. RDITT w dniu 12.06.2023
M. Skowron

dr hab. Agnieszka Nowak – Brzezińska, Prof. UŚ

Katowice, 11 maja 2023 r.

Uniwersytet Śląski
Wydział Nauk Ścisłych i Technicznych
Instytut Informatyki
e-mail: agnieszka.nowak-brzezinska@us.edu.pl

*RECENZJA ROZPRAWY DOKTORSKIEJ
DLA RADY DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA
POLITECHNIKI ŚLĄSKIEJ*

Tytuł rozprawy: Wyznaczanie rankingu stron WWW algorytmem ISOWQ Rank

Autor rozprawy: **mgr inż. Mariusz Duka**

Uwagi wstępne

Recenzja opracowana została na podstawie pisma Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej, Profesora dra hab. inż. Andrzeja Polańskiego, z dnia 6 marca 2023 r.

Łączna objętość przedłożonej do recenzji pracy wynosi 156 stron.

1. Problem naukowy (teza) rozprawy

Rozprawa poświęcona jest szukaniu rozwiązań pozwalających na ocenę jakości stron internetowych. Jakość witryny internetowej mówi nam na ile jest ona wydajna czy funkcjonalna, a także czy dostarcza cennych i dokładnych informacji. Na jakość strony internetowej składają się takie czynniki jak atrakcyjny wygląd, intuicyjna nawigacja, przejrzysty i czytelny interfejs, wysokiej jakości treść czy bezpieczeństwo strony. Użytkownik Internetu, który szuka stron poświęconych interesującej go tematyce, zawierających w swojej treści odpowiedź na pytanie zadane wyszukiwarce przez tegoż użytkownika, uzna, że dobra jakościowo strona internetowa to strona relewantna względem jego zapytań. Na ocenę jakości stron internetowych należy zatem spojrzeć z różnych punktów widzenia. Inne kryteria oceny jakości będą stosować użytkownicy Internetu, dla których dobra jakościowo strona internetowa to strona relewantna względem ich zapytań i łatwa w nawigacji. Inne kryteria będą jednak przyświecały twórcom stron internetowych, webmasterom, których zadanie polega na stworzeniu strony internetowej, która będzie wysoko pozycjonowana przez wyszukiwarki. By tak się stało, muszą oni spełnić wymagania nałożone na tworzone przez siebie strony w tym odpowiednio zoptymalizować kod HTML, uwzględnić certyfikaty bezpieczeństwa, obowiązujące technologie internetowe oraz dbać o zwiększenie zasięgu strony dzięki użyciu wtyczek mediów społecznościowych. Rynek każdej branży bazuje dziś na widzialności w Internecie. Dobra jakościowo strona WWW to taka strona internetowa, która zostanie uznana przez wyszukiwarki za wystarczająco dobrą by umieścić ją na liście wyników wyszukiwania odpowiedzi na pytania zadawane przez użytkowników.

Istnieje więc szereg czynników, które mogą być użyte do oceny jakości witryny internetowej. Uporządkowanie stron internetowych zgodnie z kryterium największej jakości

nazywamy rankingiem strony internetowych. Jeśli ranking stron internetowych wykonamy tylko w oparciu o jeden czy kilka takich czynników nie będzie on w pełni oddawał rzeczywistości. Wiarygodny ranking powinien zatem uwzględniać jak najwięcej, jeśli nie wszystkie możliwe czynniki. Wiele publikacji w Internecie wskazuje na istnienie około dwustu wskaźników branych pod uwagę przez wyszukiwarkę Google¹. Rynek wyszukiwarek jest w tym aspekcie wspierany przez wiele narzędzi (np. MOZ Analytics) pozwalających na wyznaczenie rankingu stron WWW. Są to zarówno narzędzia komercyjne jak i darmowe narzędzia dostępne dla każdego. Narzędzia te uwzględniają w większości przypadków te same czynniki, a więc poprawność kodu, szybkość ładowania się witryny, unikanie niedozwolonych technik wspomagających pozycjonowanie danej strony internetowej, autorytet strony itp.

Opracowanie konkurencyjnego systemu umożliwiającego bezpłatną analizę techniczną i ocenę jakości strony WWW w czasie rzeczywistym stało się motywacją do zaprojektowania algorytmu rankingowego ISOWQ Rank i systemu rankingowego ISOWQ. Doktorant postawił sobie za cel odkrycie czynników rankingowych wpływających na ranking MOZ, jako ten najbardziej znany i często stosowany, choć płatny i nie udostępniający publicznie szczegółów swojej zasady działania.

Głównym celem badań prowadzonych w ramach rozprawy stało się opracowanie i ocena 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 było praktyczne zastosowanie algorytmu ISOWQ Rank. W ramach prac zaprojektowano i wdrożono system informatyczny ISOWQ 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 rozprawy 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”*.

Uważam, że tak postawiona teza rozprawy została jasno sformułowana przez Doktoranta i potwierdzona badaniami wykonanymi przez Doktoranta.

2. Zawartość rozprawy

¹ <https://backlinko.com/google-ranking-factors>

Praca doktorska jest podzielona na cztery rozdziały, z których pierwsze dwa obejmują część teoretyczną, a dwa pozostałe – część praktyczną. W pierwszym rozdziale zawarto podstawową wiedzę pochodzącą z literatury na temat wyszukiwarek stron WWW oraz rodzajów wyników w wyszukiwarkach z podziałem na płatne i organiczne. Przedstawiono też metody optymalizacji stron WWW stosowane zarówno bezpośrednio na stronie WWW, jak i poza nią. W rozdziale drugim przedstawiono podstawową wiedzę o fundamentach dzisiejszych algorytmów rankingowych i badaniach naukowych związanych z próbami odkrycia czynników rankingowych wyszukiwarek. Omówiono bardzo krótko ideę takich algorytmów rankingowych jak popularny PageRank, HITS ale i MOZ czy Ahrefs. W mojej ocenie Doktorant mógł nieco głębiej zajrzeć w specyfikę każdego z nich i dokonać porównania wymienionych algorytmów pod kątem ich wad i zalet, ale przede wszystkim różnic pozwalających zrozumieć specyfikę każdego z nich i dokonać dobrego wyboru algorytmu rankingowego w zależności od tego jakie czynniki są dla nas bardziej istotne. Rozdział trzeci obejmuje opis algorytmu ISOWQ Rank, wraz z jego pseudokodem, metodą implementacji oraz architekturą systemu ISOWQ a także przykładowym raportem technicznym dla strony WWW. Wyniki badań, których celem było zbadanie korelacji między rankingiem tworzonym przez autorski algorytm ISOWQ Rank oraz algorytm MOZ umieszczono w rozdziale czwartym.

Można zatem stwierdzić, że wszystkie oczekiwane elementy zostały przez Doktoranta opisane w rozprawie. W moim odczuciu niedosyt stanowią rozdziały pierwszy i drugi, w których oczekiwałam szerszej analizy wskaźników oceny jakości stron internetowych a także tworzenia rankingów. Doktorant ograniczył się do najważniejszego algorytmu i rankingu jakim jest niewątpliwie obecnie MOZ. Jednak bardziej szczegółowa analiza wskazanych zagadnień pozwoliłaby lepiej ocenić unikalność zaproponowanego rozwiązania.

Co prawda rozdziały trzeci i czwarty prezentują zasadę działania algorytmu rankingowego ISOWQ Rank, jego założenia i metodykę przydzielania punktacji za wykorzystane technologie, pozycje rankingowe, optymalizację kodu źródłowego, treść i strukturę tekstu, zaś rozdział czwarty przedstawia wyniki badania korelacji między rankingiem wyznaczonym dla danej strony przez algorytm ISOWQ Rank oraz MOZ, jednak nie wyjaśniono tego dlaczego poszczególne czynniki, wzięte pod uwagę w ocenie, przyjmują takie rzędy wielkości, dlaczego wyniki rankingu są wartościami w przedziale od 0 do 20. Autor mógł wyjaśnić czy zastosowane wartości wynikają z wcześniejszych badań, które wykazały, że te zakresy będą optymalne? To czego także brakuje w rozprawie to po pierwsze umiejętność oceny przez Doktoranta zalet i wad zaproponowanego algorytmu a także umiejętność oceny możliwych kierunków rozwoju badań.

3. Aktualność i ważność tematyki rozprawy

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.

Pomimo postępu technologicznego i rosnącej biegłości technicznej użytkowników, strony internetowe wciąż potrzebują narzędzi oceny, które mogą poprawić zarówno ich wydajność, jak i wygodę użytkownika. Jest to najbardziej widoczne, gdy strony te należą do sektora, którego treść, funkcje i usługi charakteryzują się zestawem szczególnych wymagań.

Internet stał się dziś kluczowym kanałem komunikacji większości firm i organizacji, a także zwykłych użytkowników. Nowo powstające strony WWW wypełniają Internet treściami niepewnej jakości. 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 treści o niskiej wartości merytorycznej lub treści powielone z innych serwisów. W wynikach wyszukiwania znajdują się też nie rzadko 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.

Ważnym elementem marketingu w wyszukiwarkach SEM (ang. Search Engine Marketing) jest proces pozycjonowania strony, aby przez odpowiednią jej optymalizację uzyskać możliwie wysoką pozycję w wynikach organicznych wyszukiwarek internetowych, które według badań generują ponad 85% ruchu na stronach internetowych. 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, dziś funkcjonalność ta jest wbudowana w prawie każdej profesjonalnej wyszukiwarce. Badacze na całym świecie nieustannie podejmują próby odkrycia najważniejszych czynników wpływających na ranking w czołowych wyszukiwarkach, takich jak Google, Bing, Yahoo!, Yandex czy Baidu. 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 miar² i metod, z rankingiem MOZ, Ahrefs czy Majestic SEO na czele, ułatwiających wyznaczanie wskaźników jakości strony WWW, na podstawie których można przewidzieć wzrost lub spadek pozycji rankingowej strony WWW w wyszukiwarkach.

Wiele publikacji naukowych poświęconych tej tematyce^{3,4} zwraca uwagę na to, że wśród różnych czynników mogących decydować o jakości strony WWW należy zawsze uwzględnić odpowiednią domenę, obecność mapy strony, odpowiednio zoptymalizowane znaczniki META, wartościowe linki, zoptymalizowane parametry techniczne stron internetowych (np. opisane: pliki graficzne, filmy, adresy URL, linki wewnętrzne), słowa i frazy kluczowe, optymalizacja treści strony, aktualność strony. Jednym z najbardziej popularnych algorytmów rankingowych jest MOZ mierzący autorytet domeny i strony internetowej i na tej podstawie budujący ranking stron. Narzędzie to jest jednak narzędziem płatnym a szczegóły samego algorytmu rankingowego nie są publicznie dostępne. To stało się bezpośrednim motywatorem dla Doktoranta by opracować konkurencyjny system umożliwiający bezpłatną analizę techniczną i ocenę jakości strony WWW w czasie rzeczywistym.

² Domain Authority dla MOZ, Trust Score dla SEMrush, Domain Rating dla Ahrefs czy Trust Flow dla Majestic SEO.

³ „The Impact of Website Quality on Information Quality, Value, and Loyalty Intentions in Apparel Retailing”, Hyejeong Kim, Linda S. Niehm, *Journal of Interactive Marketing*, Volume 23, Issue 3, 2009, Pages 221-233, ISSN 1094-9968, <https://doi.org/10.1016/j.intmar.2009.04.009>.

⁴ „A Study on Website Quality Models”, R.Anusha, *International Journal of Scientific and Research Publications*, Volume 4, Issue 12, December 2014 1, ISSN 2250-3153

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.

Doktorant zaproponował w rozprawie autorski algorytm ISOWQ Rank wyznaczający numeryczną wartość jakości danej strony WWW uwzględniając większość z tych czynników. Algorytm zakłada, że jakość strony WWW będzie wyznaczona w oparciu o trzy grupy: czynniki oceniające wykorzystane technologie i pozycje rankingowe (PM), czynniki oceniające kod źródłowy (PK) oraz czynniki oceniające treść i strukturę tekstu (PT). Zaproponowany przez Doktoranta autorski algorytm ISOWQ Rank wyznacza wartość jakości strony WWW jako średnią z sumy punktów uzyskanych dla każdego z tych trzech wskaźników, odpowiednio PM, PK oraz PT. Każdy wskaźnik wyznaczany jest w 2 etapach: w pierwszym etapie tylko dla strony głównej, a w drugim dla każdej podstrony. Wynik tak uzyskany jest następnie normalizowany do przedziału 0 .. 20. Punktacja końcowa to suma punktów uzyskanych w każdym etapie z uwzględnieniem korekcji o tzw. współczynnik LR, pełniący rolę współczynnika korygującego ranking gdy na danej stronie nie ma wystarczającej liczby podstron bądź gdy zostaną wykryte hiperłącza do nieistniejących podstron.

Warto podkreślić, że zaproponowany przez Doktoranta, autorski algorytm ISOWQ Rank jest zasztyty w ramach zaprojektowanego i utworzonego przez Doktoranta systemu ISOWQ. System udostępnia rankingi stron WWW w wielu możliwych zakresach. Pozwala porównać domeny narodowe pod kątem pozycji rankingowej ISOWQ Rank, MOZ DA, MOZ PA, social media, wtyczki społecznościowe, wykorzystane usługi Google, liczba linków, struktura kodu HTML, formaty multimedialne czy testy czytelności. Raporty takie można przeglądać dla serwisów wykorzystujących wtyczki społecznościowe, pliki archiwum, pliki biurowe, muzyczne, graficzne, wideo czy serwis Youtube. System bada strony internetowe za pomocą botów, czyli robotów internetowych, które są specjalnie zaprojektowane na potrzeby systemu. Można w takim raporcie podejrzeć 5, 10 lub 20 wyników. Dużą wartością systemu jest możliwość sprawdzenia i porównania dowolnych domen narodowych z całego świata w wybranym przedziale czasowym. System ten na dzień dzisiejszy przebadał 243 domeny narodowe, ponad 13,3 mln. wykonanych audytów i prawie 27 mln. przebadanych stron WWW.

4. Oryginalny dorobek Doktoranta, jego znaczenie poznawcze oraz przydatność praktyczna dla nauki i techniki

Efektom 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.

Oryginalny dorobek rozprawy zawarty jest w rozdziałach 3-4 rozprawy. Do dorobku tego należy zaliczyć:

- Projekt i implementacja autorskiego algorytmu rankingującego strony WWW o nazwie ISOWQ Rank. Doktorant zaproponował w rozprawie autorski algorytm ISOWQ Rank wyznaczający numeryczną wartość jakości danej strony WWW uwzględniając większość z czynników uznanych za istotne przy budowaniu rankingów stron WWW. Algorytm zakłada, że jakość strony WWW będzie wyznaczona w oparciu o trzy grupy czynników: wykorzystane technologie i pozycje rankingowe (PM), kod źródłowy (PK) oraz treść i strukturę tekstu (PT).
- Projekt i implementacja autorskiego systemu rankingowego ISOWQ pozwalającego analizować strony WWW w wybranym okresie czasu. System pozwala porównać domeny narodowe pod kątem pozycji rankingowej ISOWQ Rank, MOZ DA, MOZ PA, social media, wtyczki społecznościowe, wykorzystane usługi Google, liczba linków, struktura kodu HTML, formaty multimedialne czy testy czytelności. Raporty takie można przeglądać dla serwisów wykorzystujących wtyczki społecznościowe, pliki archiwum, pliki biurowe, muzyczne, graficzne, wideo czy serwis Youtube. System bada strony internetowe za pomocą botów, czyli robotów internetowych, które są specjalnie zaprojektowane na potrzeby systemu. Można w takim raporcie podejrzeć 5, 10 lub 20 wyników. Dużą wartością systemu jest możliwość sprawdzenia i porównania dowolnych domen narodowych z całego świata w wybranym przedziale czasowym, co pozwala ocenić czy na przestrzeni czasu ranking danej strony uległ poprawie. System ten na dzień dzisiejszy przebadał 243 domeny narodowe, ponad 13,3 mln. wykonanych audytów i prawie 27 mln. przebadanych stron www.
- Zaprojektowano i przeprowadzono eksperymenty potwierdzające tezę pracy postawioną na początku. Doktorant tak zaprojektował, a następnie przeprowadził badania, aby możliwe stało się udowodnienie głównej tezy rozprawy. Eksperymenty wykonane na potrzeby rozprawy doktorskiej zrealizowano z użyciem systemu rankingowego ISOWQ, który powstawał w latach 2010-2011 a więc ma na swoim koncie ponad 10 lat funkcjonowania i przez to potencjał do oceny swojej funkcjonalności. Analiza wyników umieszczona w rozdziale 3.3. pokazuje wykorzystanie multimedii, wtyczek społecznościowych czy wersji języka znaczników HTML w konkretnych grupach domen ccTLD (tabela 5 z 2016 roku, tabela 6 dla 2021 roku). Można zauważyć, postęp dążący do poprawy jakości kodu stron WWW poszczególnych domen. Analiza wykazała dodatnią korelację między rankingiem ISOWQ Rank a MOZ. Wartościowe jest badanie korelacji par algorytmu ISOWQ Rank w całości oraz poszczególnych jego składowych przedstawione na stronie 119, 120 i 121. Dodatkowo wyniki przedstawiają procentowy udział niezakodowanych adresów e-mail w kodzie strony oraz procent rejestracji adresów IP serwerów hostujących jako nośników spamu, w bazach DNSbl – odpowiednio dla lat 2016 i 2021 (tabele 7 i 8). Potwierdzono, że opracowany algorytm ISOWQ Rank

wyznacza ranking stron WWW, który istotnie koreluje ze znanym w literaturze, i powszechnie stosowanym rankingiem MOZ.

- Pewnym globalnym miernikiem uznania istotności prac Autora mogą być indeksy bibliometryczne, tj. liczba punktów przyznanych publikacji, wskaźnik Impact Factor czy licznik cytowań. W dorobku Doktoranta odnotowujemy 3 publikacje. Wśród tych publikacji znajdziemy m.in. artykuł w czasopiśmie *Sustainability* (Core A, 100 pkt, IF 3.889) z liczbą cytowań równą 4, przy czym warto podkreślić, że publikacja ukazała się dopiero w ostatnich tygodniach. Dwie pozostałe publikacje ukazały się w renomowanym czasopiśmie Politechniki Lubelskiej o nazwie *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, z punktacją ministerialną na poziomie 20 pkt. To sprawia, że dorobek publikacyjny Doktoranta stanowią 3 publikacje na łączną liczbę punktów równą 140, z łączną liczbą cytowań równą 5.
- Doktorant może się poszczycić znaczącym doświadczeniem praktycznym w zakresie znajomości technologii internetowych. Jest autorem wielu podręczników z zakresu programowania w językach PHP czy SQL a także z zakresu Internetu rzeczy. Aktualnie wydawnictwo Helion udostępnia w sprzedaży 4 podręczniki autorstwa Doktoranta⁵.
- Oceniając całościowo dorobek rozprawy stwierdzam, że Autor w sposób systematyczny zrealizował zaplanowane badania oraz ocenił ich wyniki, dobrze je dokumentując. Doktorant wykazał się dużą starannością oraz pracowitością przygotowując rozprawę.

5. Wiedza Autora oraz znajomość współczesnej literatury z dyscypliny naukowej, której dotyczy rozprawa

Autor rozprawy wykazał się dobrą znajomością dorobku literaturowego dotyczącego zagadnień, którym poświęcona jest rozprawa. Przegląd tych zagadnień został wykonany w rozdziale pierwszym oraz drugim rozprawy. Podstawą oceny literatury światowej jest spis liczący 227 pozycji bibliografii. Doktorant uwzględnił tu zarówno prace odnoszące do najnowszych badań w temacie rozprawy jak i prace stanowiące genezę analizy regresji czy metodologii analizy predykcyjnej. Tak ujęty dobór literatury oceniam pozytywnie. Świadczy on o dostatecznej wiedzy Doktoranta w danej dyscyplinie naukowej.

6. Wady i słabe strony rozprawy, uwagi dyskusyjne

Rozprawa jako całość nie ma istotnych wad. Treść rozprawy, użyty język, sposób prezentacji wyników, a także sformułowane wnioski, które cechuje ceniona zwięzłość i przejrzystość, świadczą o wysokich umiejętnościach Autora do poprawnego i przekonującego przedstawienia uzyskanych wyników.

Wśród słabszych stron bądź uchybień można wymienić następujące:

- a) Nie uwzględniono złożoności obliczeniowej, nie przedstawiono wad algorytmów, brak definicji jakości strony, brak definicji rankingu strony
- b) Autor pisze, że „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”

⁵ https://helion.pl/autorzy/mariusz-duka?gclid=Cj0KCQjwmtGjBhDhARIsAEqfDEF_NpRrk95FKzAt8RTXieC1mKafPgE834oHuolCyaWeUZMzxc5n5LEaAkAEEALw_wcB

jednak w rozprawie nie podzielił się tą analizą i wnioskami z niej wyciągniętymi wprost. Trudno zatem ocenić, które to były prace badawcze i jakie wnioski Doktorant wyciągnął z ich lektury, jaką wiedzę wykorzystał bezpośrednio w implementowanym algorytmie ISOWQ Rank oraz systemie rankingowym o tej samej nazwie.

- c) Doktorant pisze na stronie 15 „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.”, jednak w rozdziale 4.3. gdzie przedstawia wyniki badań nie tłumaczy tej korelacji, ograniczając się jedynie do stwierdzenia, że jest to silna dodatnia korelacja.
- d) Doktorant pisze na stronie 16 rozprawy, że „w rozdziale 3 omówiono architekturę systemu ISOWQ oraz jego budowę i strukturę baz danych, a także zaprezentowano przykładowy raport techniczny dla strony WWW. W mojej ocenie Systemowi temu nie poświęcono wystarczająco dużo uwagi. Prezentując strukturę baz danych doktorant przedstawił schemat bazy danych jednak nie zinterpretował i nie opisał relacji zachodzących między poszczególnymi tabelami zawierającymi gromadzone przez system dane. Z kolei pisząc o architekturze ograniczono się jedynie do bardzo podglądowego rysunku nr 11 przedstawiającego schemat systemu ale ponownie bez interpretacji tych składowych i wyjaśnienia ich roli w systemie i zadań realizowanych przez te składowe np. automatyczny system typowania serwisów do analizy. Nie dowiadujemy się w pracy jak działa taki system.
- e) Autor pisze na stronie 23, że „Przyszłością wyszukiwarek będzie wykorzystanie sztucznej inteligencji”. Tak przecież jednak już się dzieje od jakiegoś czasu (Google czy Microsoft używają sztucznej inteligencji i przetwarzania języka naturalnego do lepszego rozumienia pytań zadanych przez użytkowników), stąd niezgoda moja na to by mówić o tym jako o zadaniach na przyszłość. Inna sprawa, że Doktorant nie pochylił się nad tematem wykorzystania algorytmów sztucznej inteligencji i uczenia maszynowego w wyszukiwarkach, a z pewnością prowadzone przez niego badania i sama rozprawa doktorska na pewno zyskałyby na wartości.
- f) Cenny jest rozdział pierwszy w którym Doktorant przedstawia popularne technologie i metodyki tworzenia rankingów. Szkoda, że nie ujęto np. tabeli porównującej omówione technologie. Dobrze byłoby dokonać syntezy z której wynikałoby które czynniki są np. najczęściej analizowane przez dostępne narzędzia tworzące rankingi a które najważniejsze? Taka synteza byłaby już prostą drogą do określenia tych wskaźników, które koniecznie powinny znaleźć się w systemie rankingowym.
- g) Rozdział 2.3 ma omawiać metody nadawania rankingów. Oczekiwałamby więc, że doktorant rozpocznie od definicji rankingów a następnie pokaże możliwe metody nadawania rankingów znane z literatury. Doktorant ograniczył się do dwóch najbardziej znanych podejść tj. PageRank oraz HITS, przedstawił ogólnie ideę każdego z nich ale użył do ich prezentacji zupełnie innych przykładów. W mojej ocenie o wiele bardziej cenne byłoby pokazanie różnych podejść w odniesieniu do tego samego przykładu, co pozwoliłoby wychwycić ewentualnie różnice.
- h) W założeniach algorytmu ISOWQ czytamy, że jakość strony WWW jest wyznaczona jako średnia z sumy punktów dla trzech składowych: wykorzystanie technologii, optymalizacja kody źródłowego oraz treść i struktura tekstu. Moja uwaga dotyczy tu rozkładu wag dla każdego z tych 3 składników. W mojej ocenie System byłby bardziej atrakcyjny pozwalając użytkownikowi systemu ISOWQ decydować o rozkładzie wag

- dla każdego z tych składników. Moje pytanie jest zatem takie, czy Doktorant rozważał możliwość przydzielania wag wskaźnikom PM, PT i PK?
- i) Nie jest znany algorytm przydziału wartości liczbowych poszczególnym wskaźnikom. Jednym z założeń jest to, że dobrze przygotowana strona ma minimum 4 podstrony i jest to kryterium, za którego spełnienie strona WWW otrzyma tyle samo punktów co za treść i strukturę tekstu. W mojej ocenie Doktorant powinien bardziej przejrzysto wyjaśnić metodykę zastosowaną do ustalenia zakresów wartości poszczególnym wskaźnikom. Na stronie 53 czytamy, że czynniki wpływające na obliczenie trzech głównych parametrów, tj. PM, PK i PT przez lata się zmieniały. Ale Doktorant nie tłumaczy w jakim zakresie się zmieniały, jaka była tego przyczyna i jakie wnioski. Znajomość odpowiedzi na takie pytanie pozwoliłaby na pewno lepiej zrozumieć i na pewno docenić wkład pracy Autora.
 - j) Na stronie 54 autor tłumaczy, że uzyskana z wyliczeń wartość rankingowa dla danej strony podlega procesowi przekształcania do przedziału 0..20. Jeśli oznacza to po prostu użycie normalizacji wyniku, metodą MinMax⁶, to warto to jasno określić.
 - k) Rozdział 3.1.7. stanowi podsumowanie. Doktorant pisze w nim, że „Algorytm ISOWQ Rank jest wykorzystywany przez system rankingowy ISOWQ do nadawania rankingów stronom WWW. System ten grupuje strony WWW w przedziałach co 5 punktów, odpowiednio wyróżniając je dodatkowym oznaczeniem graficznym. Szkoda, że Doktorant nie przypisał etykiety typu wzorowa, bardzo dobra, dobra, dostateczna, niedostateczna. Ułatwiałyby one ocenę uzyskanej wartości liczbowej rankingów.
 - l) Doktorant często ogranicza rozważania do stron biznesowych. Str. 24: "rozszerzeniem wyników organicznych, najczęściej przy zapytaniach o konkretną markę są (...) do podstron". Warto byłoby zatem już na początku uwzględnić założenie, że praca skupia się na dążeniach do uzyskania wysokiej pozycji rankingowej przez strony internetowe firmowe.
 - m) W punkcie 4.2. dokonano wstępnej analizy danych jednak bazując na treści tego podrozdziału należy wywnioskować, że analiza objęła wybranych 10 stron WWW pod kątem takich czynników jak HTML5, media społecznościowe, usługi Google, framework JS, multimedia czy szyfrowanie. Powstaje pytanie, czy uwzględniono treść strony w tej analizie?.
 - n) Najsłabiej w mojej ocenie opisano część oceny jakości strony internetowej pod kątem oceny jakości treści i struktury tekstu. Wiadomo, że na jakość tę wpływa rozmiar i formatowanie widocznego tekstu na stronie. Brane są też pod uwagę testy czytelności mierzące stopień trudności rozumienia tego tekstu. W pracy możemy jedynie wyczytać, że w tym celu algorytmu ISOWQ Rank obliczy całkowitą liczbę liter i słów w tekście na stronie (optymalnie by było to 100 słów), sprawdzi liczbę znaczników H1..H6, i dobrze, jeśli strona zawiera co najmniej jeden znacznik H1 i co najmniej jeden znacznik H2. Jeśli zawiera dodatkowe znaczniki wyróżnienia tekstu na stronie (Znaczniki I, LI czy inne) to zdobywa więcej punktów za jakość struktury tekstu. Jeśli treść umieszczona na stronie zwraca wystarczającą wartość indeksów czytelności tekstu to za każdy taki badany test czytelności można uzyskać zwiększenie punktacji jakości. Doktorant nie wspomina czy algorytm ISOWQ Rank bierze pod uwagę spójność nagłówków z treścią. Wiadomo, że bardzo ważnym czynnikiem, który każda wyszukiwarka powinna brać pod uwagę jest unikalność tekstu umieszczonego na

⁶ <https://www.oreilly.com/library/view/hands-on-machine-learning/9781788393485/fd5b8a44-e9d3-4c19-bebb-c2fa5a5ebfee.xhtml>

stronie. Umieszczając unikalne, nasycone słowami kluczowymi teksty możemy liczyć na „przychyłość” silnika wyszukiwarki, dzięki czemu mamy większe szanse na widoczność w sieci. Za kopiowanie treści z innych stron naszą witrynę może spotkać kara, w postaci obniżenia pozycji w indeksie wyszukiwarki i odczuwalnego spadku popularności serwisu. W mojej ocenie część zaproponowanego algorytmu ISOWQ Rank poświęcona ocenie treści i struktury tekstu mogła zostać rozbudowana chociażby o ten element badania podobieństwa treści umieszczonych w nagłówkach z tekstem połączonym z danym nagłówkiem. To pozwoliłoby sprawdzić na ile spójna jest cała strona i treść ją tworząca.

- o) Na stronie 88 Doktorant wymienia indeksy czytelności stosowane do ceny czytelności treści strony, jak indeks czytelności Flescha, indeks czytelności Gunninga-Foga i dalej kolejne na stronie 89 ale nie definiuje ich bliżej czytelnikowi. Nie precyzuje też dokładnie, które z nich użyto w autorskim algorytmie i systemie rankingowym ISOWQ. Pseudokod umieszczony na stronie 90 zawiera skróty, które następnie czytający pracę musi sam dopasować do odpowiadających im testów czytelności. Doktorant nie wyjaśnia w pracy dlaczego dodatkowe punkty, które strona internetowa może otrzymać za spełnienie warunków dyktowanych przez dany test czytelności, są różne tj. 0,25, 0,5 i 0,75.
- p) Jako, że praca dotyczy dyscypliny informatyka spodziewałam się nieco większego wkładu Autora w zagadnienia związane zarówno z analizą proponowanego algorytmu w zakresie np. złożoności obliczeniowej (jako kryterium pozwalającego ocenić wagę danego algorytmu), jak i implementacji wskazanego algorytmu (np. szczegółów dotyczących konkretnego środowiska obliczeniowego).
- q) Korekta rozprawy jest dość dokładna i w całej pracy nie znalazłam błędów ortograficznych, czy znaczących interpunkcyjnych, co w tak liczonym opracowaniu jest praktycznie nie do uniknięcia. Wymienię kilka błędów literowych:
- Str. 9 - słowniczek - wszystkie oznaczenia są podane z wielkiej litery (ang.) a dla ccTLD nie zastosowano tej reguły.
 - Str. 17, rozdział 1, wszystkie rozdziały powinny być od nowej strony.
 - Str. 24, "Wyszukiwarka Google...Żadna inna wyszukiwarka nie ma tak wielu różnych typów" - ale "typów czego" ?
 - Str. 41, "należą do nich są algorytmy" -> "należą do nich algorytmy".
 - Str. 44, niefortunne określenie strony WWW jako "publikacji" w podrozdziale 2.2. Doktorant powinien wcześniej zaznaczyć jeśli synonimem strony internetowej, czy strony WWW jest słowo publikacja.
 - We wszystkich pseudokodach zamiast symbolu € użyto symbol € przeznaczony do reprezentacji walut.
 - W pracy w wielu miejscach Doktorant posługuje się nietypowym dla języka polskiego stylem. Dla przykładu, na stronie 87: "wyróżnić tekst na stronie WWW w formie pochylenia pozwala znacznik". Znacznie lepiej brzmiałoby "Znacznik... pozwala wyróżnić tekst na stronie..."
 - Str. 80,83,85 itd. „0,00 punktów” -> „0 punktów”. Nie ma potrzeby zapisu wartości z dwoma miejscami po przecinku gdy wiemy, że tą minimalną możliwą wartością jest 0.

- Str. 102, w tabeli 5 jest błąd w opisie pod tabelą. Umieszczony tam opis tj. „serwisy w domenie pl. - tylko 22%” powinno być „prawie 40%” skoro mowa o języku HTML.
- Doktorant nie pilnuje konsekwencji w nagłówkach tabel. Skoro przedstawiają te same dane, dla dwóch różnych okresów czasu tj. roku 2016 i 2021 to powinny mieć te same nagłówki.
- Str. 111, literówka „Maps:" - > "Maps,".
- Str. 149, Zdanie o użyciu języka R i bibliotek Kendall jest powtórzeniem ze strony 116 pod tabelą 9 a przed kodem źródłowym 11.

i szereg innych, podobnych drobnych usterek, które nie obniżają w sposób znaczący wartości pracy.

7. Wniosek końcowy

W podsumowaniu stwierdzam, że mimo drobnych uchybień przedstawionych powyżej, przedłożona do recenzji praca doktorska wykonana przez Pana mgra inż. Mariusza Dukę spełnia w mojej opinii wymagania stawiane kandydatom do stopnia naukowego doktora, określone w art. 14 ust. 2 pkt 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz stopniach i tytule w zakresie sztuki (Dz. U. z 2017 r. poz. 1789), § 6 ust. 1 rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 19 stycznia 2018 r. w sprawie szczegółowego trybu i warunków przeprowadzania czynności w przewodzie doktorskim, w postępowaniu habilitacyjnym oraz w postępowaniu o nadanie tytułu profesora (Dz. U. z 2018 r. poz. 261) w związku z art. 179 ust. 1 ustawy z dnia 3 lipca 2018 r. – Przepisy wprowadzające ustawę-Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2018 r. poz. 1669, z późn. zm.). W mojej ocenie przedłożona rozprawa zawiera oryginalne rozwiązanie problemu. Doktorant osiągnął stawiany cel, wykazując się niezbędną wiedzą i umiejętnościami do samodzielnego rozwiązywania problemów naukowo-technicznych z wykorzystaniem metod informatycznych.

Wnoszę zatem o dopuszczenie Pana *mgra inż. Mariusza Duki* do następnej fazy przewodu doktorskiego.

Robert - Brzezinski

