

Recenzja rozprawy doktorskiej Mariusza Duki: *Wyznaczanie rankingu stron WWW algorytmem ISOWQ Rank*

dr hab. Przemysław Stpiczyński, prof. UMCS*

7 maja 2023

1 Informacje ogólne

Przedmiotem recenzji jest rozprawa doktorska mgr. inż. Mariusza Duki *Wyznaczanie rankingu stron WWW algorytmem ISOWQ Rank*. Recenzja została wykonana na mocy Uchwały nr 15/23 z dnia 14 lutego 2023 Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Śląskiej. Promotorem w postępowaniu o nadanie stopnia doktora jest dr hab. Artur Strzelecki, prof. Uniwersytetu Ekonomicznego w Katowicach.

2 Sylwetka doktoranta

Mgr inż. Mariusz Duka, od roku 2018 student niestacjonarnych studiów III stopnia na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej, uzyskał tytuł zawodowy licencjata w roku 2003 na Wydziale Zarządzania i Marketingu Wyższej Szkoły Biznesu w Dąbrowie Górniczej. Tytuł zawodowy magistra uzyskał po ukończeniu odbywanych w latach 2016-2018 studiów niestacjonarnych na Wydziale Informatyki i Zarządzania Wyższej Szkoły Informatyki i Umiejętności w Łodzi. Tematem pracy magisterskiej była *Kryptografia krzywych eliptycznych (ECC) i algorytm Argon2 w języku PHP 7.2 z wykorzystaniem biblioteki OpenSSL i Sodium*.

Doktorant jest autorem bądź współautorem trzech publikacji naukowych opublikowanych w czasopiśmie z wykazu MEiN z dnia 21 grudnia 2021: *Sustainability* (jedna w roku 2023) oraz *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska* (dwie w roku 2020). Jest również autorem czterech monografii opublikowanych w wydawnictwie Helion w latach 2020-2023. Jego kariera zawodowa trwa od roku 1994. W latach 1994–1998 pracował jako kierownik placówki Vobis Microcomputer w Chorzowie. W latach 1995–2004 prowadził własną działalność gospodarczą pod

*Instytut Informatyki, Uniwersytet Marii-Curie Skłodowskiej, Akademicka 9/519, 20-032 Lublin, email:przemyslaw.stpiczynski@umcs.pl

nazwą *Interface*. Od roku 2001 do chwili obecnej pełni funkcję prezesa zarządu oraz kierownika działu IT w Interface Poland Sp. z o.o. w Chorzowie. W latach 2010–2020 pełnił analogiczne funkcje w Interface Latvia SIA w Rydze (Łotwa).

3 Opis rozprawy

Przedstawiona do recenzji rozprawa składa się z nienumerowanego wprowadzenia, pięciu rozdziałów oraz bibliografii liczącej 227 pozycji. Treść uzupełniają dodatkowo wykaz oznaczeń używanych w pracy, spis ilustracji oraz spis listingów i pseudokodów. Całość zajmuje 156 stron. Struktura pracy jest właściwa i sprzyja bezproblemowemu poznawaniu przedstawianych treści. Autor często odwołuje się do obszernej bibliografii zawierającej również najnowsze pozycje literatury naukowej z lat 2021–2022.

Pierwszy rozdział (Wprowadzenie, str. 13–17) zawiera krótkie wprowadzenie do problematyki tworzenia algorytmów rankingowych dla stron internetowych stosowanych w wyszukiwarkach, motywację badań i uzasadnienie wyboru tematu rozprawy doktorskiej, cele i tezę pracy oraz opis jej struktury. Autor stawia sobie za cel opracowanie oraz ocenę skuteczności algorytmu ISOWQ Rank, który wartościuje strony WWW pod względem ich jakości oraz wykryciu ewentualnej korelacji pomiędzy rezultatami działania tego algorytmu i algorytmu wykorzystywanego w komercyjnym systemie MOZ oraz zbadanie jaki wpływ na tę korelację mają poszczególne czynniki rankingowe, w tym treść i struktura tekstu na stronie WWW. Drugi cel pracy stanowi wykazanie praktycznej przydatności opracowanego algorytmu poprzez jego implementację w systemie służącym do analizy jakości stron internetowych. Autor formułuje tezę rozprawy następująco: *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.*

Pierwszy numerowany rozdział *SEO jako metoda budowania ruchu z wyników organicznych* (str. 17–36) prezentuje najważniejsze informacje związane z działaniem wyszukiwarek internetowych stron WWW i opisuje dość szczegółowo techniki, które powinny być użyte w trakcie tworzenia kodu strony WWW oraz wykonanie konkretnych działań poza nią tak, aby możliwe było osiągnięcie przez stronę dobrej pozycji w danej wyszukiwarce. Omówione są również dostępne narzędzia (Google Search Console, Ahrefs, Majestic, Semrush) umożliwiające przeprowadzenie audytu strony pod kątem jej przygotowania do współpracy z wyszukiwarkami, występowania ewentualnych błędów w kodzie źródłowym oraz identyfikowania brakujących elementów w zawartości.

Drugi rozdział (str. 37–47) poświęcony jest przedstawieniu ogólnych zasad działania wyszukiwarek internetowych w oparciu o algorytmy rankingowe dla stron WWW. Autor skupia się na omówieniu dwóch najważniejszych algorytmów wykorzystywanych w pozycjonowaniu stron WWW: PageRank i HITS oraz wspomina o innych, których szczegóły są objęte tajemnicą handlową (np. MOZ Rank). Na potrzeby pre-

zentacji algorytmów PageRank i HITS autor opracował narzędzie umożliwiające obliczenie wartości rankingowych dla dowolnej struktury linkujących się wzajemnie stron WWW.

Następne rozdziały stanowią zasadniczą część pracy zawierającą własne wyniki autora. Rozdział trzeci (str. 50–110) zawiera szczegółowe omówienie sposobu działania algorytmu rankingowego ISOWQ Rank oraz prezentację wyników działania systemu rankingowego ISOWQ. W odróżnieniu od omówionych wcześniej algorytmów PageRank i HITS, które wymagają od wyszukiwarek zebrania jak największej ilości danych o powiązaniach pomiędzy stronami, co z uwagi potencjalnie bardzo dużą liczbę hiperłączy dostępnych w sieci może być bardzo kosztowne, algorytm ISOWQ wyznacza rangę strony WWW w oparciu o ich jakość. Oznacza to, że potencjalnie będzie on wymagał znacznie mniejszych zasobów i jego użycie nie będzie się wiązało z koniecznością przeszukiwania bardzo rozległych zasobów sieci. Algorytm określa jakość strony jako średnią arytmetyczną punktów uzyskanych przez stronę w trzech kategoriach: wykorzystane technologie i pozycje rankingowe, optymalizacja kodu źródłowego, treść i struktura tekstu. Punkty w ramach każdej kategorii naliczone są za spełnienie przez stronę konkretnych warunków (poprawność i jakość kodu źródłowego, użycie konkretnych technologii, wtyczki społecznościowe itp), bądź uzyskanie przez nią określonych wartości rankingowych MOZ i Alexa Rank. Finalnie wartość ISOWQ Rank jest liczona w skali od 0 do 20 punktów. Autor kolejno prezentuje poszczególne czynniki uwzględniane przez algorytm i opisuje sposób przyznawania konkretnych wartości punktowych i podaje jego pseudokod. Skrótowo omawia również jego implementację w języku PHP. W dalszej części rozdziału omawiana jest architektura systemu rankingowego ISOWQ, będącego narzędziem do technicznej analizy stron internetowych. Przedstawiona została architektura systemu oraz wnioski płynące z jego funkcjonowania przez okres dziesięciu lat, w trakcie których w bazie danych systemu odnotowano szczegółową analizę ponad 1.3 milionów stron WWW i ponad 26 mln adresów URL. Warto podkreślić, że system umożliwia gromadzenie i udostępnianie danych w formie okresowych raportów, dzięki czemu możliwe jest monitorowanie zmian zachodzących w stosowanych technikach tworzenia stron internetowych.

Kolejny rozdział pracy (str. 111–122) zawiera badanie porównawcze algorytmów rankingowych ISOWQ Rank i MOZ. Autor wykorzystał statystykę τ Kendalla przy zadeklarowanym poziomie istotności 0.05 do oszacowania relacji pomiędzy punktacjami uzyskanymi za pomocą tych dwóch algorytmów. W tym celu wybrał dziesięć witryn internetowych o różnej tematyce charakteryzujących się określoną złożonością. Rozdział dość szczegółowo opisuje przeprowadzone badanie statystyczne i poprawnie identyfikuje czynniki wpływające na uzyskane wyniki, które potwierdzają dodatnią korelację pomiędzy punktacją ISOWQ Rank a punktacją MOZ. W ostatnim rozdziale (str. 123–127) autor podsumowuje uzyskane w pracy wyniki oraz przedstawia wnioski wskazujące jakie elementy stron internetowych oraz sposoby ich umieszczania są kluczowe dla osiągnięcia wysokich pozycji rankingowych w wyszukiwarkach.

4 Dyskusja

Autor podjął się realizacji ciekawego i ważnego tematu badawczego. Oczywistym jest fakt, że twórcom komercyjnych serwisów internetowych zależy dostosowaniu strony technicznej oraz formy prezentowanych treści do tego, aby ich strony znalazły się wysoko w rankingu. Zatem projektując serwis internetowy warto zawczasu zadbać o to, by spełniał określone własności. Wyszukiwarki internetowe opierają swoje działania na niejawnych modyfikacjach algorytmów PageRank i HITS, które do oceny jakości strony WWW analizują strukturę hiperłączy mierząc liczbę odwołań pochodzących z innych stron. Implementacja takich algorytmów wymaga bardzo dużych zasobów. Stąd podejście autora, by oceniać jakość strony biorąc pod uwagę treść i strukturę tekstu oraz optymalizację kodu źródłowego, w tym użyte technologie wraz z detekcją problemów technicznych występujących na stronie i serwerze WWW, daje szansę na poprawę kodu strony, która może się przełożyć na wysoką pozycję rankingową w wyszukiwarkach internetowych. Oczywiście otwarty problem stanowi to, co dokładnie należy wziąć pod uwagę i jakie wagi nadać poszczególnym elementom. Istotne jest również opracowanie metodologii tworzenia kodu źródłowego stron pozwalającej między innymi na łatwą ekstrakcję treści, a przez to osiągnięcie wyższej pozycji rankingowej.

Autorowi rozprawy udało się osiągnąć zamierzone cele i rozwiązać oryginalny problem, w wyniku czego powstało rozwiązanie, na które składa się algorytm rankingowy ISOWQ Rank oraz system do technicznej analizy stron internetowych demonstrujący zastosowanie wyników własnych badań naukowych w sferze gospodarczej. Autor zastosował również metody statystyczne do wykazania dodatniej korelacji pomiędzy algorytmem ISOWQ Rank i algorytmem wykorzystywanym przez komercyjny system MOZ. Przeanalizował również wpływ wybranych czynników rankingowych na korelację wyników działania tych algorytmów. Trzeba również podkreślić, że rozprawa dość dobrze prezentuje ogólną wiedzę teoretyczną autora oraz jego umiejętność samodzielnego prowadzenia pracy naukowej skutkującej uzyskaniem ciekawych wyników, które można zastosować w praktyce.

Reasumując, do najważniejszych oryginalnych wyników, które zostały przedstawione w rozprawie, zaliczyć należy następujące:

1. Opracowanie własnego złożonego algorytmu rankingowego ISOWQ Rank, który umożliwia wartościowanie stron WWW na podstawie technicznych własności strony oraz wartości rankingowych MOZ DA, MOZ PA i Alexa Rank.
2. Sformułowanie szeregu wniosków odnośnie sposobu tworzenia stron internetowych tak, by zwiększyć szansę na ich wysoką pozycję w rankingach.
3. Opracowanie architektury i wdrożenie systemu rankingowego ISOWQ stanowiącego narzędzie do technicznej analizy stron internetowych.
4. Przeprowadzenie badań statystycznych potwierdzających dodatnią korelację pomiędzy pozycjami rankingowymi wyznaczanymi za pomocą algorytmu ISOWQ Rank a punktacją MOZ.

Przedstawiona do recenzji rozprawa ma również swoje słabe strony i pewne niedociągnięcia. Poniżej przedstawiam zagadnienia warte przedyskutowania.

1. Trafność działania algorytmu rankingowego ISOWQ Rank zależy od punktów uzyskanych za poszczególne elementy brane pod uwagę w trakcie jej analizy. Autor nie wyjaśnia dokładnie w jaki sposób doszedł do tego, że zaprezentowane w rozprawie zakresy punktowe stanowią najlepsze rozwiązanie. Na stronie 54 możemy jedynie znaleźć informację, że *„ich wartości zmieniały się w czasie w zależności od ówczesnie 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.”* Jak była metodologia prowadzenia takich badań empirycznych?
2. Kontynuując ten wątek warto zapytać o to, jak zmieni się punktacja poszczególnych elementów algorytmu ISOWQ Rank, gdy pewne elementy staną się nieaktualne. Wydaje się, że taka sytuacja ma obecnie miejsce: w grudniu 2021 poinformowano o zamknięciu serwisu 1.05.2022 roku oraz o zamknięciu API do 15 grudnia tego samego roku.
3. Autor z przesadną swobodą używa terminów „optymalizacja” oraz „optymalny”. Przykładowo na stronie 101 pisze *„Decyzja o wykorzystaniu formatu SQLite okazała się optymalna”*. Optymalna, czyli najlepsza. Jakie przyjęto zatem kryterium optymalności oraz jak upewniono się, że nie ma innych rozwiązań dających lepsze wyniki? Również formułując tezę pracy autor stwierdza, że *„algorytm ISOWQ Rank w sposób optymalny wyznacza ranking stron WWW przez nadanie im określonej wartości, oznaczającej ich jakość.”* Kryterium optymalności nie zostało tutaj jasno określone.
4. Prezentacja implementacji algorytmu pozostawia pewien niedosyt. Co prawda autor podaje dużo odwołań do literatury, ale pewne zagadnienia mogłyby zostać przedstawione w sposób bardziej szczegółowy, na przykład algorytm ekstrakcji tekstu ze strony WWW przez przekształcenie kodu HTML w obiekt drzewa DOM (str. 98).
5. Przedstawione na stronie 99 wymagania sprzętowe dla działania systemu ISOWQ wydają się stosunkowo niewielkie. Czy zwiększenie zasobów sprzętowych przyspieszyłoby działanie systemu, bądź też umożliwiło analizę większej liczby serwisów internetowych?
6. Autor nie ustrzegł się pewnych błędów redakcyjnych. Przykładowo na stronie 51 słowo „Internet” pisane jest małą literą. Rozdział 1 nie zaczyna się na początku nowej strony tak, jak pozostałe. We wzorze (1) na stronie 42 występują symbole n oraz N , a z kontekstu wynika, że chodzi o tę samą wartość. Ponadto w rozdziale 3 niektóre rysunki mogłyby być lepszej rozdzielczości (np. rys. 17). Rozdział trzeci wydaje się zdecydowanie za długi (to niemal połowa tekstu rozprawy). Jego treść mogłaby zapełnić dwa oddzielne rozdziały. Czytelność

algorytmu ISOWQ Rank byłaby z pewnością lepsza, gdyby autor przedstawił pogrupowane elementy oceny i odpowiadające im wartości punktowe na przykład w postaci tabeli.

5 Konkluzja

Podsumowując uwagi zawarte w recenzji stwierdzam, że rozprawa doktorska prezentuje ogólną wiedzę teoretyczną doktoranta i potwierdza, że posiada on umiejętność samodzielnego prowadzenia pracy naukowej oraz stanowi oryginalne rozwiązanie problemu naukowego, również w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej, a więc spełnia wymogi określone w ustawie z dnia 20 lipca 2018 roku *Prawo o szkolnictwie wyższym i nauce* (Dz. U. 2018, poz. 1668 z późn. zm.). Wnoszę zatem o dopuszczenie mgr. inż. Mariusza Duki do publicznej obrony jego rozprawy doktorskiej oraz dalszych etapów postępowania o nadanie stopnia doktora w dyscyplinie *Informatyka techniczna i telekomunikacja*.

