dr hab. inż. Mikołaj Morzy, prof. PP

Instytut Informatyki Politechniki Poznańskiej

Mikołaj.Morzy@put.poznan.pl

Recenzja rozprawy doktorskiej

Tytuł rozprawy: **Prediction methods for networked bibliographic data**

Autor rozprawy: **Rajmund Klemiński**

Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza pracy) i czy zostało ono dostatecznie jasno sformułowane przez autora? Jaki charakter ma praca (teoretyczny, doświadczalny, inny)?

Przedstawiona do recenzji rozprawa doktorska jest poświęcona połączeniu dyscypliny naukometrii (ang. *scientometrics*) z nauką o sieciach (ang. *network science*) i jej metodami. Autor rozprawy pokazuje na różnych przykładach, jak przyjęcie perspektywy sieciowej otwiera nowe możliwości analizy danych bibliometrycznych. Przykładami takich nowych możliwości jest automatyczna ocena jakości dokumentów, automatyczna ekstrakcja popularnych tematów badawczych, czy predykcja sukcesu publikacyjnego poszczególnych prac. Przedstawione w rozprawie przykłady są ilustrowane obszernymi eksperymentami obliczeniowymi realizowanymi na bardzo dużych zbiorach danych. Praca ma charakter empiryczny i eksploracyjny. Wiele wyników przedstawionych w rozprawie jest obiecujących i mogą stanowić podstawę do dalszych badań o charakterze bardziej analitycznym.

Naukometria to ważne narzędzie do analizy badań naukowych oraz oceny wpływu i oddziaływania badań we współczesnej nauce. Analizując produkcję naukową poprzez naukometrię, można zidentyfikować trendy i wzorce w danej dziedzinie, ocenić wpływ badań i ocenić ich jakość. Naukometria może być również stosowana do pomiaru wpływu badań na społeczeństwo, do analizy geograficznego rozkładu badań oraz do identyfikacji najważniejszych tematów badawczych. Zastosowanie naukometrii jest niezbędne do zrozumienia obecnego stanu nauki i przewidywania przyszłych trendów. Jest to ważne narzędzie dla rozwoju współczesnej nauki, ponieważ pomaga naukowcom w identyfikacji najważniejszych tematów badawczych, ocenie jakości badań oraz ocenie wpływu nauki na społeczeństwo. Przedstawiona do recenzji rozprawa doktorska bardzo dobrze wpisuje się w ten trend, starając się dostarczyć nowatorskich narzędzi opisu i predykcji, łącząc tradycyjną naukometrię z perspektywą sieciową analizy danych.

W rozprawie poruszono szerokie spektrum zagadnień, od oceny jakości tekstów popularno-naukowych, poprzez automatyczną identyfikację interesujących trendów publikacyjnych, na predykcji przyszłego wpływu publikacji naukowych skończywszy. Wspólnym mianownikiem dla przedstawionych badań jest wykorzystanie danych bibliometrycznych analizowanych z perspektywy sieciowej,

tj. transformowanych do postaci grafów o określonych topologiach i semantyce wierzchołków i krawędzi. Recenzowane osiągnięcie nie jest zatem próbą rozwiązania jednego konkretnego i dobrze zdefiniowanego problemu. Raczej mamy do czynienia z szerokim empirycznym przeglądem możliwych analiz zahaczających o różnorakie aspekty naukometrii. Stanowi to jednocześnie główną zaletę, jak i wadę recenzowanej rozprawy. Niewątpliwie wieloaspektowe przyjrzenie się danym bibliometrycznym i próba ich analizy pod różnymi kątami jest zadaniem ambitnym i ciekawym. Z drugiej strony, brak precyzyjnie zdefiniowanego problemu badawczego powoduje, że rozprawa traci wewnętrzną spójność i sprawia wrażenie raczej złożenia kilku tematycznie pokrewnych publikacji (o czym świadczyć też może brak konsekwencji w stosowaniu oznaczeń matematycznych). Mimo to uważam, że poruszone w rozprawie tematy stanowią ważne cele badań naukowych i spełniają wymagania stawiane rozprawom doktorskim.

Rozprawa doktorska jest napisana w języku angielskim i liczy 214 stron, wliczając bibliografię. Do rozprawy dołączono trzy załączniki w języku angielskim oraz streszczenie w języku polskim. Nieco dziwi umieszczenie spisu użytych oznaczeń matematycznych jako Załącznika A, spis oznaczeń byłby znacznie bardziej przydatny na początku rozprawy. Bibliografia liczy 226 pozycji, z których 4 są współautorstwa Doktoranta i jest on pierwszym autorem. Rozdział 1 stanowi wprowadzenie do rozprawy, określenie zakresu przeprowadzonych badań, motywację do podjęcia badań oraz cele stawiane przez Autora. Rozdział drugi stanowi bardzo obszerne omówienie aktualnego stanu wiedzy w obszarach rozważanych w rozprawie. Kolejne rozdziały przedstawiają wyniki badań przeprowadzonych nad: automatyczną oceną jakości artykułów popularno-naukowych (rozdział 3), detekcji obiecujących trendów i tematów badawczych w obszarze informatyki (rozdział 4), identyfikacji tematów przy użyciu różnych konstrukcji sieci cytowań (rozdział 5), oraz predykcji przyszłej liczby cytowań publikacji naukowych (rozdział 6). W rozdziale 7 Autor krótko podsumowuje zawartość rozprawy i wylicza uzyskane rezultaty.

Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle, świadczący o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący?

Autor przedstawia bardzo obszerny przegląd literatury w rozdziale 2. Na 40 stronach szczegółowo opisuje liczne prace dotyczące modelowania tematów, oceny jakości tekstu, czy predykcji cytowań. Biorąc pod uwagę to, jak wiele aspektów jest poruszonych w rozprawie, trudno oczekiwać, aby przegląd literatury w każdym z opisywanych obszarów był wyczerpujący. Większość omawianych prac pochodzi z ostatnich 5-10 lat i ich zestawienie stanowi rzetelny ogląd stanu wiedzy w danym obszarze. Może nieco dziwić niewielka reprezentacja badań korzystających z modeli neuronowych, szczególnie jeśli chodzi o modele predykcyjne czy analizę tematów. Być może jest to pochodną przyjętego aparatu eksperymentalnego, w ramach którego większość zadań analizy tekstu także jest

realizowane za pomocą tradycyjnych technik, takich jak LDA. Moja ogólna ocena doboru literatury jest pozytywna. Znów, główną trudność stanowi tu szeroki zakres omawianych tematów, uniemożliwiający faktycznie dogłębne przedstawienie stanu wiedzy. Poza tym, jak wspomniałem, w pracy nie stawia się precyzyjnej hipotezy badawczej, stąd nie sposób ocenić, na ile poszczególne tematy były we wcześniejszych pracach eksploatowane i jak dokładnie umiejscawia się rozprawa na tle wcześniejszych badań. Nie jest to zarzut wobec Autora, raczej obserwacja pewnych konsekwencji przyjętej ogólnej koncepcji rozprawy. Nie mam wątpliwości, że Autor jest bardzo dobrze zorientowany w omawianej tematyce, poszczególne pozycje nie są opisane pobieżnie ale w sposób znamionujący dobrą znajomość każdej pracy.

Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

Recenzowana praca ma charakter eksploracyjny, poszczególne rozdziały stanowią raczej prezentację potencjału kompilacji danych bibliometrycznych i metod analizy sieciowej niż propozycje rozwiązania konkretnych problemów badawczych. Przyjętą przez Autora metodą jest eksperyment obliczeniowy, żadne z poruszanych zagadnień nie jest "atakowane" w sposób analityczny. Stanowi to pewną wadę, ponieważ żaden z przedstawionych w pracy eksperymentów nie ma charakteru wyczerpującego i każdemu można postawić zarzut anegdotyczności. Z drugiej strony zdecydowanie należy docenić pracę włożoną w poszczególne eksperymenty, Autor każdorazowo stara się zgromadzić faktycznie bogaty i obszerny zbiór danych, na podstawie którego można pokusić się o próbę generalizacji uzyskanych wyników.

Fundamentalnym założeniem poczynionym przez Autora jest przyjęcie, że dane bibliometryczne stanowią rzetelne odzwierciedlenie procesu naukowego. Jest zupełnie zrozumiałe, dlaczego takie założenie zostało poczynione, natomiast osobiście mam pewien problem z jego akceptacją. Z góry zaznaczam, że reszta bieżącego paragrafu ma charakter bardzo subiektywnej oceny i stanowi raczej zaproszenie Doktoranta do dyskusji niż sformułowanie jakiegokolwiek zarzutu. Jestem przekonany, że nauka jest konstruktem społecznym, odzwierciedlającym wartości i uprzedzenia społeczeństwa, w którym jest uprawiana. Sieci cytowań, które stanowią podstawę wiedzy naukowej, są kształtowane przez politykę, zachęty i normy kulturowe, które określają, które badania są uważane za ważne. Co więcej, wiele cytowań może być przypadkowych i faktycznie nie związanych z prowadzonymi badaniami. Może to również obejmować odniesienia z książek, czasopism, magazynów lub innych źródeł, które są tylko mgliście związane z pracą. Ta praktyka quasi-losowych odniesień jest często spotykana w pracach naukowych w celu wypełnienia wymaganej liczby odniesień. Na to nakładają się agregatory publikacji naukowych które mogą z łatwością dostarczać list artykułów do powierzchownego cytowania (żeby nie wspomnieć o modelach generatywnych pokroju Galactiki które mogą całkowicie automatycznie generować cytowania w pracach). Znamy liczne przypadki nieuczciwego zachowania się

naukowców i edytorów, którzy sztucznie konstruowali wokół siebie sieci cytowań w celu maksymalizacji różnych współczynników bibliometrycznych. Biorąc pod uwagę, że kraje rozwinięte wydają 2%-3% swojego produktu narodowego brutto na badania naukowe i rozwój, nie trudno zrozumieć, skąd biorą się zachęty do tego typu działań. Podsumowując, w mojej opinii jest całkiem prawdopodobne, że niewielka część cytowań reprezentuje prawdziwy wpływ, a znakomita większość cytowań jest efektem "wędkowania po cytowania" przy użyciu agregatorów (Google Scholar, PubMed, Semantic Scholar), kopiowania cytowań z wcześniej cytowanych prac, itp. Gdyby tak faktycznie było, to wiele wniosków sformułowanych przez Autora straciłoby na sile.

Jeszcze raz pragnę zaznaczyć, że powyższe stanowi mój prywatny i krytyczny osąd rzeczywistości publikacyjnej. W pełni rozumiem, dlaczego Autor przyjął takie założenia, jakie przyjął i uważam je za prawidłowe. Chciałbym jednak zachęcić Autora do refleksji i podzielenia się swoimi przemyśleniami na temat tego, na ile zaprezentowana w rozprawie wizja świata nauki jest realistyczna.

Czy autor wykazał umiejętność poprawnego i przekonującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Poziom redakcyjny rozprawy jest bardzo dobry. Praca jest napisana w języku angielskim w sposób przejrzysty i zrozumiały. Poza pojedynczymi literówkami nie zauważyłem żadnych istotnych błędów gramatycznych czy stylistycznych. O dziwo, więcej błędów pojawia się w streszczeniu napisanym w języku polskim. W rzeczywistości nie jest to streszczenie, to raczej dosłowne tłumaczenie rozdziału pierwszego pracy. Nie zawiera więc opisów przeprowadzonych eksperymentów czy podsumowania uzyskanych wyników. W streszczeniu rzucają się w oczy błędy, które mogłyby zostać z łatwością wyeliminowane przy nieco staranniejszej edycji tekstu:

- w języku polskim nie występuje słowo "scientometria", w przeciwieństwie do umieszczonej w SJP "naukometrii",
- kapitalizacja słów w tytułach podrozdziałów, akceptowalna w języku angielskim, w języku polskim stanowi błąd,
- w tekście występują literówki, np. "cieci współautorstwa", "rozwiązaniem jest analizy", "na wykorzystaniu zbiory",
- na etapie edycji nie włączono biblioteki umożliwiającej poprawne dzielenie wyrazów w języku polskim, czego efektem są podziały takie jak mi-ary, przeprowad-zone, nauc-zony, itd.
- angielski termin "clustering" w piśmiennictwie polskim jest tłumaczony jako "analiza skupień" lub "grupowanie", a nie jako "gronowanie".

W całej pracy przeszkadzała mi też nieco niestaranność w użyciu trybu matematycznego w języku LaTeX, która skutkowała rzucającymi się w oczy błędami w składzie czcionek. Przykładowo, funkcja masy prawdopodobieństwa jest zapisywana jako $pmf_V(l)$ zamiast $pmf_V(l)$, EMD zamiast EMD, $Neigh$ zamiast

Neigh, itp. To jest oczywiście drobiazg, ale trochę żal nie korzystać z możliwości edycji i składu oferowanych przez LaTeX.

Jakie są słabe strony rozprawy i jej główne wady?

Swoje uwagi postanowiłem pogrupować zgodnie z kolejnymi rozdziałami pracy. Niektóre uwagi wskazują na pomyłki, część uwag dotyczy fragmentów rozprawy które są dla mnie niejasne, a część uwag ma charakter polemiczny i stanowi zaproszenie do dyskusji.

Rozdział 3

- w dodatku A i definicji 1 używa się symbolu v do oznaczenia wierzchołka, podczas gdy w sekcji 3.1 wierzchołki są oznaczone jako V_i, V_j
- w definicji 7 mylnie użyto operatora $\in$ zamiast operatora zawierania się zbiorów $\subset$
- w definicji 8 pojawia się niezdefiniowane wcześniej pojęcie $C(\mathcal{G})$ i czytelnik musi się domyślać, że chodzi o zbiór klik w grafie
- w pierwszym zdaniu po definicji 8 pojawia się niezdefiniowany wcześniej symbol v
- w równaniu 3.3 pojawiają się niezdefiniowane wcześniej ani później v i v_c'
- w równaniu 3.4 pojawia się L, czy to jest to samo co $\mathcal{L}$?
- w równaniu 3.4 występuje funkcja $\phi(.)$ która zgodnie z def. 9 jest funkcją zdefiniowaną dla zbioru wierzchołków, a nie pary wartości skalarnych, poza tym pojawia się funkcja Ψ która została zdefiniowana dopiero w równaniu 3.5
- pomysł, żeby nagle zamienić funkcję na zbiór funkcji, i to funkcji o różnych sygnaturach, nie pomaga w śledzeniu myśli przewodniej Autora
- głównym uzasadnieniem wyboru algorytmu LBP było założenie, że te same osoby pracują nad powiązanymi artykułami, co uprawdopodabnia możliwość transmisji jakości między artykułami, czy w takim wypadku nie byłoby lepszym rozwiązaniem skonstruowanie grafu na podstawie historii edycji a nie odnośników między stronami?
- w rozdziale 3.2.2 Autor pomija 4 wymiary oceny i koncentruje się jedynie na wiarygodności, zakładając, że "[...] similarity should be apparent in each of the dimensions. As such, analyzing one of these qualities should provide sufficient insight about other dimensions" - jest to bardzo mocne i niepoparte żadnym dowodem założenie, a wystarczyłoby zaraportować wzajemne korelacje ocen w poszczególnych cechach żeby uprawdopodobnić to założenie
- Autor dość swobodnie wprowadza kluczowy parametr progu klasyfikacji binarnej 0.74 i pisze "[...] additional analysis was conducted aimed at determining a threshold value suitable as a binary similar/dissimilar qualifier", biorąc pod uwagę istotność tego parametru warto byłoby doprecyzować, na czym owa analiza polegała
- mam wrażenie, że wprowadzony w rozdziale 3.2.5 benchmark dla ocen 1-4 ma na celu jedynie ratowanie wyników eksperymentu, które nie są

imponujące, bo już sam przyjęty algorytm referencyjny (uśrednienie ocen) jest bardzo słaby, a w praktyce interesowałoby nas przewidywanie właśnie najwyższych ocen jakości artykułów

rozdział 4

- we wstępie pojawia się zdanie, które chyba jest ucięte przed zakończeniem ("broaden their horizons by")
- w równaniu 4.3 definiowana jest funkcja rankingu, więc chyba $\mathbb{R} \rightarrow \mathbb{Z}$
- w eksperymentach wykorzystuje się tradycyjne modele LDA które, jak trafnie zauważa Autor, nie nadają się do odkrywania tematów w relatywnie krótkich tekstach, tym bardziej szkoda, że Autor nie rozważał wykorzystania metod odkrywania tematów bazujących na gęstej reprezentacji wektorowej tekstu, takich jak np. TopicBERT.
- w sekcji 4.3.2 Autor definiuje n-gram corpus jako efekt konkatenacji dokumentów związanych z tematem t_i, jest jednak zupełnie niejasne, skąd ten temat się wziął, na schemacie 4.1 krok n-gram corpus nie jest poprzedzony krokiem, w którym następowałaby ekstrakcja tematów z tytułów (chyba, że jest to schowane wewnątrz kroku "processed data", choć nie wynika to z opisu)
- w sekcji 4.3.3.2 na przestrzeni kilku zdań występuje trzykrotnie zmienna n (lub N), przy czym za każdym razem oznacza coś zupełnie innego (liczba słów, liczba tematów), nie ułatwia to śledzenia myśli Autora
- architektura eksperymentu z sekcji 4.4 jest dziwna, nie jest dla mnie jasne, dlaczego zrezygnowano z metod bazujących na predykcji szeregów czasowych (skoro właśnie o przewidywanie przyszłych wartości w szeregu czasowym chodzi), zamiast tego agregując dane z kilku poprzedzających lat; dodatkowo, wyniki eksperymentu nie dają się zinterpretować w oparciu o dostarczony opis, nie wiadomo np. jaka jest zmienność cech w czasie, jeśli jest niewielka to zadanie jest trywialne i wysoka precyzja uzyskana dla cech *DAF*, *DFT* czy *Pop* nie jest niczym dziwnym. Autor nie wspomina też nigdzie, jaka jest łączna liczba tematów w opisywanym eksperymencie.

rozdział 5

- nie zrozumiałem, jaki jest związek między procedurą ekstrakcji n-gramów a różnymi topologiami sieci cytowań
- na stronie 106 prawdopodobnie chodzi o indeksy dolne H_0, H_1
- na stronie 106 pojawia się, występujący w dalszych równaniach, czynnik $O(G_t)$ opisany jako "an observation of G_t's connectivity", ale ten czynnik nigdzie nie zostaje formalnie zdefiniowany
- z opisu p_C na stronie 107 wydaje się, że to jest procent izolowanych wierzchołków w G_t^X a nie parametr
- Autor bez żadnego uzasadnienia czy wyprowadzenia prezentuje wzory definiujące $\log P(O(G_t^r)|H_i)$ dla $i = \{0, 1\}$, to są kluczowe wzory dla zrozumienia całego algorytmu prezentowanego w rozdziale 5 a brak ich szczegółowego opisu lub objaśnienia jest istotnym uchybieniem
- sformułowanie "questions stated in the paper" mogło było zostać zmo-

dyfikowane podczas adaptowania treści artykułu do postaci rozdziału w rozprawie

rozdział 6

- myślę, że struktura rozdziału mogła być przemyślana w taki sposób, aby uniknąć odnośników do pojęć definiowanych w dalszych częściach tekstu, np. def. 12 nie można zrozumieć bez przeczytania def. 13 i def. 14
- w całym rozdziale korzysta się z embeddingów wierzchołków w różnych projekcjach sieci bibliometrycznej, ale nie mogłem doszukać się w tekście rozprawy, jaka metoda jest wykorzystywana do wyznaczenia tych embeddingów
- sekcja 6.4.2: nigdy nie spotkałem się z podejściem, w którym ważność jednego z d wymiarów reprezentacji wektorowej była traktowana jako ważność całego atrybutu i mam bardzo poważne wątpliwości co do metodologicznej poprawności takiego rozumowania. Gdyby embeddingiem była prosta reprezentacja one-hot, to użycie jednego z wymiarów co najwyżej wskazywałoby na istotność konkretnej wartości danego atrybutu, a nie atrybutu jako całości. W przypadku gęstej reprezentacji wektorowej (w której najprawdopodobniej $d \geq 100$), uogólnienie ważności jednego z tych wymiarów na cały atrybut jest, moim zdaniem, nieuzasadnione
- sekcja 6.4.3: eksperyment przedstawiony w tej części rozprawy jest bardzo przyjemnym eksperymentem myślowym, natomiast metodologicznie jest niestety zupełnie niepoprawny. Brak tu miejsca żeby szczegółowo się rozwodzić, problem polega na tym, że Autor próbuje z poziomu danych (czyli asocjacji) przejść do poziomu tzw. counter-factuals (wyobrażonych "światów" w których publikacja została zgłoszona do innego czasopisma czy konferencji) z pominięciem pośredniej warstwy interwencji. Istnieje cała bogata literatura poświęcona wnioskowaniu przyczynowemu (ang. *causality learning*) do której można się odwołać. Reasumując, przedstawione wyniki można traktować jako interesującą anegdotę, ale nie można ich traktować jako faktycznego modelu predykcyjnego.

Jaka jest przydatność rozprawy dla nauk technicznych?

Przedstawiona do recenzji rozprawa doktorska dowodzi dużej użyteczności warsztatu naukowego rozwijanego w ramach nauki o sieciach dla prowadzenia zaawansowanych badań w obszarze naukometrii. Do głównych kontrybucji rozprawy zaliczam:

- adaptację algorytmu LBP do wielokryterialnej oceny jakości dokumentów,
- przeprowadzenie eksperymentalnej ewaluacji użyteczności różnych projekcji sieci bibliometrycznej dla zadania odkrywania tematów w tytułach prac naukowych,
- metodę projekcji sieci bibliometrycznej na homogeniczne podsieci reprezentujące relacje między autor(k)ami, miejscami publikacji i tematami.

Lektura rozprawy przekonuje mnie o wysokich kompetencjach naukowych Autora.

Mimo, że prezentowane wyniki nie stanowią, w mojej ocenie, ostatecznego rozwiązania żadnego konkretnego problemu, są obiecującą wskazówką do kontynuacji badań nad użyciem metod sieciowych w naukometrii. Praktyczne konsekwencje prowadzonych badań mogą być bardzo istotne ze względu na ogromne nakłady inwestowane przez państwa rozwinięte w obszar B+R. Cząstkowe wyniki prezentowane w rozprawie były prezentowane w czasopismach naukowych (Journal of Information Science, Journal of Infometrics) i materiałach konferencji naukowych (ENIC, ACIIDS).

Rozprawa doktorska pana mgr. Rajmunda Klemińskiego pt. "*Prediction methods for networked bibliographic data*" spełnia warunki określone w art. 13 ustawy Prawo z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2017 poz. 1789). Recenzowana rozprawa stanowi oryginalne rozwiązanie w zakresie zastosowania wyników własnych badań naukowych w sferze społecznej. Rozprawę oceniam pozytywnie i wnoszę o jej dopuszczenie do publicznej obrony.

Poznań, 11 lipca 2022 r.