prof. dr hab. Dominik Ślęzak, prof. UW
Wydział MIM, Uniwersytet Warszawski
ul. Banacha 2, 02-097 Warszawa
e-mail: slezak@mimuw.edu.pl

2022-03-19

Recenzja rozprawy doktorskiej

mgr Pavla Savova

pod tytułem

Measuring the Novelty of Scientific Papers

1. Problem badawczy i jego znaczenie

Rozprawa dotyczy rozwoju skutecznych metod wykrywania tych spośród publikacji naukowych, które są najbardziej przełomowe, najbardziej przyczyniają się do rozwoju nauki. Parametryzacja prac naukowych jest aspektem niezwykle ważnym dla środowiska akademickiego, ale nie tylko – wiąże się ona właśnie także ze zrozumieniem dynamiki rozwoju dziedzin, powstawania przełomowych odkryć. Tymczasem obecne techniki oceny publikacji – tak zwane indeksy scjentometryczne – wciąż bazują bardziej na danych bibliometrycznych (takich jak np. cytowania), a mniej na zrozumieniu istoty wyników przedstawianych w publikacjach (a więc także np. na zrozumieniu kontekstu cytowań). Właściwy pomiar innowacyjności publikacji naukowych pozostaje wciąż problemem otwartym.

Pan Pavel Savov podchodzi do idei zaprojektowania miary innowacyjności w interesujący sposób. Otóż proponuje on wyznaczać ją bazując na nowatorskiej analizie, na ile zawartość analizowanych prac koreluje się z tematyką prac z przeszłości i przyszłości. Korelacje – czy też podobieństwa – są tu postrzegane w terminach błędów modeli szacujących lata publikacji analizowanych artykułów, które to modele zostały wyuczone na zebranych korpusach artykułów historycznych. Doktorant pokazuje na wielu przykładach, że opracowana miara daje wartościowe wyniki i że jest ona istotnie odmiennym indeksem oceny niż standardowo pojmowane miary bazujące na liczbach cytowań artykułów.

W mojej ocenie badania nad tak pojmowanymi miarami innowacyjności są ważne nie tylko dla ludzi nauki, czy też dla instytucji oceniających ludzi nauki. Możliwość skutecznego identyfikowania prac przełomowych może być również bardzo istotna w edukacji, czy też np. dla współpracy pomiędzy środowiskami akademickimi i przemysłem pragnącym wdrażać nowatorskie rozwiązania.

2. Wkład autora w dziedzinę

Rozprawa bazuje na trzech pracach opublikowanych w materiałach międzynarodowych, z czego jedna to artykuł w uznanym czasopiśmie naukowym Information Processing and Management (rok 2020), zaś dwie to prace na konferencji International Conference on Computational Science (2020 i 2021):

[A] Pavel Savov, Adam Jatowt, Radosław Nielek: Identifying breakthrough scientific papers. Inf. Process. Manag. 57(2): 102168 (2020)

[B] Pavel Savov, Adam Jatowt, Radosław Nielek: Innovativeness Analysis of Scholarly Publications by Age Prediction Using Ordinal Regression. ICCS (2) 2020: 646-660

[C] Pavel Savov, Adam Jatowt, Radosław Nielek: Predicting the Age of Scientific Papers. ICCS (1) 2021: 728-735

Jak widać, we wszystkich tych pracach doktorant jest pierwszym autorem. Co więcej, zgodnie z informacjami z Google Scholar (na dzień 19 marca 2022) artykuł [A] ma już 28 cytowań. Ranga czasopisma Information Processing and Management, jak i serii konferencji ICCS, a także istniejące już cytowania wskazują, że wkład Pana Pavla Savova w dziedzinę, którą się zajmuje, jest już znaczny, ogólnie zauważalny, z pewnością wystarczający na tym etapie kariery naukowej.

Jak już wspomniałem w sekcji 1, doktorant podchodzi do postawionego sobie problemu badawczego w ciekawy sposób, który przyczynia się moim zdaniem nie tylko do ulepszenia działania algorytmów w omawianej dziedzinie indeksowania scjentometrycznego, ale też stanowi wskazówkę, jak można łączyć różne metody, w różnych warstwach rozwiązań informatycznych i analitycznych. Mamy tutaj bowiem do czynienia z pomysłem niejako „dwupoziomowym", gdzie z jednej strony szacowany jest inteligentnymi metodami rok publikacji danego analizowanego artykułu, zaś z drugiej strony – obserwowany błąd tego szacowania jest wykorzystywany do oceny innowacyjności artykułu.

Jeśli model szacujący rok publikacji oznacza artykuł jako późniejszy niż był on faktycznie wydany, to prowadzi to do konkluzji, że artykuł ten w jakimś sensie „wyprzedził swój czas" i że w późniejszym okresie pojawiło się znacząco więcej prac na tematy omawiane w tym artykule. Ten późniejszy wzrost nasilenia zajmowania się daną tematyką przez międzynarodową społeczność naukową wpływa na błąd modelu wyuczonego poprzez analizę zawartości korpusu prac, a z drugiej strony wskazuje, że skoro obserwowane nasilenie nastąpiło, to dany analizowany artykuł mógł ten trend zapoczątkować.

Choć powyższy pomysł wydaje się intuicyjny i „oczywisty", (a warto przy okazji dodać, że najlepsze rozwiązania cechują się zazwyczaj właśnie intuicyjnością i „oczywistością",) doktorant musiał włożyć dużo pracy w jego doszlifowanie. Praca ta, której przebieg dokumentują publikacje [A,B,C], zawiera wiele elementów innowacyjnych, autorskich. (Można zatem zaryzykować stwierdzenie, że jeżeli za kilka lat ktoś zechciałby określić omawianymi metodami stopień innowacyjności artykułów [A,B,C], to powinien otrzymać wysoki wynik, nie tylko w terminach wspomnianej liczby cytowań.)

Po pierwsze, samo zadanie predykcji roku publikacji, na podstawie dostępnych informacji o tychże publikacjach, (ale z wyłączeniem danych, które czyniłyby rozwiązanie tego zadania w jakimś sensie bezużytecznym,) nie jest proste. Pan Pavel Savov posłużył się w tym przypadku modelowaniem tematycznym, nadzorowanymi modelami predykcyjnymi (np. maszynami wektorów podpierających) i modelami osadzania słów (bardzo popularne podejście BERT), trenowanymi na diachronicznych korpusach artykułów naukowych pochodzących z różnych źródeł i z ponad 20 lat.

Po drugie, niejako w kolejnym kroku, doktorant opracował miarę wyrażającą innowacyjność w terminach konkretnych liczb, odwołującą się do podobieństwa zawartości analizowanych artykułów do zawartości publikacji z przyszłości oraz przeszłości. Miara ta nie jest jednak prostym błędem pomiędzy faktyczną datą publikacji a rokiem przewidzianym przez modele z kroku pierwszego. Autor słusznie zauważa, że modele predykcyjne mogą przecież zwracać rozkłady prawdopodobieństw na poszczególne lata i całe te rozkłady (a nie tylko daty najprawdopodobniejsze) powinny być użyte jako wejście do miary odzwierciedlającej innowacyjność. Ponadto, Pan Savov umiejętnie wziął pod uwagę to, że liczba publikowanych na świecie artykułów stale rośnie, i w odpowiedni sposób przeskalował oszacowania innowacyjności. A więc mamy tu do czynienia nie tylko z autorskim zaadaptowaniem samych technik analizy i uczenia się z danych, ale i z nietrywialnym modelowaniem matematycznym, wymagającym gruntownego zrozumienia dziedziny praktycznej, której dotyczy rozprawa.

Konkludując uważam, iż wkład mgr Savova w dziedzinę jest w pełni wystarczający dla spełnienia wymagań stawianych przez ustawę w odniesieniu do prac doktorskich w dyscyplinie informatyki.

3. Poprawność i redakcja

Rozprawa ma w mojej ocenie bardzo elegancką, przejrzystą strukturę. Wspomniane wcześniej trzy artykuły naukowe współautorstwa Pana Pavla Savova są wkomponowane w rozprawę jako rozdziały nr 3, 4 oraz 5 (odpowiadające kolejno pracom [A], [B] oraz [C] wymienionym w sekcji 2 niniejszej recenzji), z jasnym przedstawieniem głównych elementów innowacyjnych w każdym z nich.

Co więcej, na zawartość rozprawy składa się wstęp (rozdział nr 1) z rzeczowym przedstawieniem problematyki i proponowanego podejścia, przegląd literatury (rozdział nr 2), odnoszący się zarówno do obecnie stosowanych metod scjentometrycznych, jak i technik analizy danych, których doktorant używa do poprawy tych metod, a na koniec podsumowanie i pewne pomysły na dalsze badania w tej dziedzinie (rozdział nr 6). Rozprawa jest opatrzona w szereg tabel, wykresów, zarówno idee, metody, podstawy matematyczne, jak i wyniki eksperymentalne są przedstawione poprawnie.

Warto podkreślić, że ta rozprawa pod względem redakcyjnym stanowi jedność, nie jest zaś zwykłym sklejeniem wcześniejszych publikacji. Świadczy o tym chociażby uspójniona bibliografia zawarta na końcu rozprawy, która zastąpiła analogiczne sekcje w oryginalnych publikacjach. To bardzo pomaga w lekturze. Natomiast drobne błędy redakcyjne (np. czasami brak spacji przed odwołaniem do pozycji bibliograficznej w tekście) nie mają żadnego znaczenia, jeśli chodzi o zrozumiałość rozprawy.

4. Wiedza kandydata

Tak jak pisałem pod koniec sekcji 2, doktorant musiał – w celu sprostania problemowi badawczemu, na którym skoncentrował się w rozprawie – połączyć wiedzę oraz doświadczenie z wielu dziedzin, poczynając od analizy danych, w tym analizy dużych korpusów tekstowych, poprzez metody uczenia modeli predykcyjnych (nadmienione już maszyny wektorów podpierających, jak również np. metody regresji), modelowanie matematyczne (niezbędne na poziomie konstrukcji matematycznych miar innowacyjności, które faktycznie oddawałaby naturę zagadnienia), a wreszcie zrozumienie specyfiki danego zadania praktycznego, jakim jest rozwój narzędzi do oceny publikacji naukowych.

Jeżeli złożymy to w całość z umiejętnością przygotowania wysokiej jakości artykułów (takich jak [A,B,C], ale nie tylko – są też inne) oraz przestawiania wyników szerszej publiczności (miałem m.in. przyjemność wysłuchania referatu Pana Savova na seminarium badawczym na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego, gdzie odpowiedział on na wiele pytań dotyczących otrzymanych rezultatów, jak i ogólnych aspektów dziedziny rozprawy), otrzymamy obraz osoby kompetentnej, wszechstronnej, umiejącej formułować i rozwiązywać problemy naukowe.

5. Podsumowanie i ocena końcowa

W świetle moich opinii zawartych w poprzednich sekcjach, biorąc pod uwagę ustawowe wymagania stawiane doktoratom w obszarze informatyki, oceniam rozprawę pozytywnie. Uważam, że rozprawa doktorska mgr Pavla Savova może być dopuszczona do publicznej obrony.

Co więcej uważam, że byłoby to z dużą wartością dla środowiska akademickiego, gdyby Pan Savov kontynuował swoje badania naukowe. Wskazują na to obserwacje zawarte w rozdziale nr 6 rozprawy. Ze swojej strony dodałbym, że stosowanie takich metod mogłoby być przydatne także w nieco innych dziedzinach zastosowań, na przykład w analizie zgłoszeń patentowych.

Dominik Ślęzak