dr hab. inż. Jakub Nalepa, prof. PŚ
Katedra Algorytmiki i Oprogramowania
Politechnika Śląska
Jakub.Nalepa@polsl.pl

Gliwice, 8 marca 2024 r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy: Modele rekomendacyjne wspólnej filtracji w serwisach ogłoszeniowych

Autor: Mgr Robert Kwieciński

Promotor: Dr hab. Tomasz Górecki, prof. UAM

Promotor pomocniczy: Dr hab. Agata Jolanta Filipowska, prof. UEP

1 Tematyka rozprawy

Recenzowana rozprawa doktorska Pana mgr. Roberta Kwiecińskiego dotyczy zagadnień związanych z projektowaniem, analizą i wdrażaniem technik uczenia maszynowego w systemach rekomendacyjnych systemów ogłoszeniowych, ze szczególnym uwzględnieniem ogłoszeń w kategorii „praca". Prace podjęte przez Doktoranta skupiły się na stworzeniu rzeczywistego zbioru danych, który został wykorzystany do oceny jakości opracowywanych modeli oraz na zaproponowaniu szeregu modeli rekomendacji. Autor szczególną uwagę zwrócił na aspekty wdrożeniowe zaproponowanych technik – praca była realizowana w ramach III edycji programu „Doktorat wdrożeniowy", we współpracy z Grupą OLX sp. z o.o. Uważam, że tematyka rozprawy jest aktualna, zgodna z obecnym nurtem badań i dotyczy istotnych aspektów praktycznego wykorzystywania technik i modeli uczenia maszynowego oraz zaawansowanej analizy danych, zwłaszcza w kontekście ogromnej ilości danych, które są obecnie generowane w serwisach ogłoszeniowych.

2 Ocena strony merytorycznej rozprawy doktorskiej

Rozprawa doktorska jest napisana w języku polskim i składa się z ośmiu rozdziałów oraz bibliografii. **Rozdział pierwszy** jest wprowadzeniem do tematyki rozprawy, w którym Autor krótko omawia istotność podjętych badań i ich praktyczny potencjał, a także przedstawia wyzwania związane z opracowywaniem systemów rekomendacyjnych w serwisach ogłoszeniowych. W rozdziale tym przedstawiono również charakterystykę Grupy OLX Sp. z o.o., w ramach której miały zostać wdrożone modele opracowane w niniejszej rozprawie – w podrozdziale 1.3. („*Rozwój rekomendacji ofert pracy w OLX*"), Doktorant sytetycznie omawia wdrożone modele rekomendacyjne oraz przedstawia ich wpływ na serwisy oferowane przez Grupę OLX Sp. z o.o. W podrozdziale 1.4. („*Cele rozprawy*"), Autor przedstawia cel główny rozprawy oraz trzy cele pomocnicze – do tych celów Autor będzie odwoływał się w dalszej części dysertacji. Rozdział pierwszy zwieńczony jest opisem struktury rozprawy.

Rozdział drugi rozpoczyna się od dokładniejszego przedstawienia systemów rekomendacyjnych, w którym omówione zostały dwa podejścia (predykcyjne i rankingowe) definiowania takich systemów. Następnie Autor przedstawia typy systemów rekomendacyjnych oraz ich możliwe podziały, które są tworzone ze względu na jawność ocen czy metody uczenia się rangowania. W kolejnej części tego rozdziału zostały dokładniej omówione wyzwania związane z tworzeniem, oceną i wdrażaniem modeli rekomendacji. Szczególną uwagę Autor poświęca metodom ewaluacji systemów rekomendacyjnych (*online* i *offline*) oraz metrykom, które są wykorzystywane do oceny jakości modeli rekomendacyjnych – odpowiednie zaprojektowanie procedury walidacyjnej i rzetelny dobór metryk są

niezwykle istotne, ponieważ bezpośrednio wpływają na ocenę (i ostatecznie wybór) tworzonych modeli uczenia maszynowego (nie tylko w zagadnieniach związanych z systemami rekomendacji).

W **rozdziale 3.** omówiony został zbiór danych (*OLX Jobs Interactions*), który został stworzony i opublikowany (na platformie Kaggle) w ramach prac podjętych w niniejszej rozprawie – opracowanie i przedstawienie zbioru było jednym z celów pomocniczych. Autor porównuje zaproponowany zbiór danych z trzema innymi, znanymi z literatury (RecSys16, RecSys17 oraz CareerBuilder12) podkreślając najistotniejsze różnice pomiędzy tymi zbiorami oraz to, że dwa pierwsze są już niedostępne i nie mogą być wykorzystywane do oceny opracowywanych modeli rekomendacyjnych. W rozdziale tym Doktorant przedstawia analizę ilościową stworzonego zbioru, a także (bardzo) krótko omawia podział tego zbioru na podzbiory treningowe i testowe, które będą wykorzystane do oceny algorytmów rekomendacyjnych w dalszej części rozprawy doktorskiej. Uważam, że opracowanie rzeczywistego zbioru danych jest istotnym wkładem w rozwój dziedziny tworzenia modeli rekomendacyjnych, ponieważ może pozwolić innym grupom badawczym porównać opracowywane techniki wykorzystując ten sam podział danych oraz ten sam zestaw metryk. Zapewnienie takiej możliwości jest obecnie szczególnie istotne, ze względu na tzw. „kryzys reprodukowalności" badań, widoczny w różnych gałęziach nauki i przemysłu [1].

W **rozdziale 4.** Doktorant skupia się na analizie teoretycznej i eksperymentalnej metod rekomendacji znanych z literatury. Omówione zostały także aspekty implementacyjne – warto podkreślić, że Autor udostępnił swoje implementacje (przez repozytorium `GitHub`), starając się zapewnić reprodukowalność otrzymanych wyników eksperymentalnych i jednocześnie „zachęcając" inne grupy badawcze do wykorzystania opracowanych implementacji. W połączeniu z przygotowaniem i upublicznieniem nowego zbioru danych, uważam że jest to istotny krok, który może pozwolić na ustandaryzowanie (i zobiektywizowanie) sposobu oceny istniejących i nowych modeli rekomendacyjnych. W dalszej części rozdziału Autor przekrojowo opisuje ocenę rozważanych modeli, zarówno w podejściu *offline* jak i *online*, a istotność statystyczna otrzymanych wyników eksperymentalnych została zweryfikowana z wykorzystaniem odpowiednich testów statystycznych.

Istotnym aspektem praktycznym związanym z wdrażaniem modeli rekomendacyjnych jest możliwość generowania rekomendacji w czasie rzeczywistym – w **rozdziale piątym**, Autor skupia się na opisie infrastruktury, która pozwoliła na wdrożenie takiego rozwiązania w serwisach Grupy OLX Sp. z o.o. Przedstawiono również nowy wariant modelu *RP3Beta* (*RP3Beta real-time*), który został z powodzeniem wdrożony we wspomnianej infrastrukturze. W podrozdziale 5.4., Doktorant krótko omawia ewaluację *online* modeli *RP3Beta* (modelu działającego w trybie wsadowym) i *RP3Beta real-time* (modelu generującego rekomendacje w czasie rzeczywistym) – wyniki badań eksperymentalnych wskazują na to, że wykorzystanie modelu *RP3Beta real-time* pozwoliło znacznie zwiększyć liczbę osób odpowiadających na rekomendowane oferty pracy. Warto podkreślić, że Doktorant wspomniał o dwóch istotnych aspektach (w tym o błędzie w opracowanej implementacji), które mogły wpłynąć na wyniki otrzymane w ramach eksperymentów – jest to dowód na naukową „szczerość" i dojrzałość Autora (błędy, które mogą pojawić się w czasie badań są czymś „normalnym" i bardzo istotnym jest to, żeby potrafić się do nich przyznać, wyciągnąć odpowiednie wnioski i poprawnie zinterpretować wpływ takich błędów lub niedociągnięć na otrzymywane wyniki).

Rozdział szósty zawiera opis modelu rekomendacyjnego *P3LTR*, będącego uogólnieniem modelu *RP3Beta*, wzbogaconego o możliwość uwzględnienia cech interakcji użytkowników i „przedmiotów". Autor omawia teoretyczne zalety opracowanego modelu, a w sekcji 6.3. („*Ewaluacja*") przedstawia wyniki jego ewaluacji *offline*.

W **rozdziale siódmym**, Doktorant jasno omawia najistotniejsze powody rozpoczęcia badań nad grafowymi sieciami neuronowymi w kontekście tworzenia systemów rekomendacyjnych. W rozdziale zawarto także bardzo krótkie wprowadzenie do grafowych sieci neuronowych, które stanowi wstęp do opisu modelu *P3GNN* opracowanego w ramach niniejszej rozprawy. Model ten został porównany z szeregiem innych algorytmów wspólnej filtracji opartych na grafowych sieciach neuronowych,

a także z modelem *RP3Beta*. W ramach badań opisanych w tym rozdziale Autor skupił się także na optymalizacji (licznych) hiperparametrów sieci grafowych, opracowaniu nowych funkcji straty oraz na procesie generowania przykładów negatywnych. Wyniki badań eksperymentalnych wykazały, że model *P3GNN* pozwala na osiągnięcie lepszych wyników niż z wykorzystaniem pozostałych rozważanych sieci grafowych.

W ostatnim **rozdziale 8.**, Autor podsumowuje rozprawę i odnosi się do celu głównego oraz celów pomocniczych, jasno omawiając w jaki sposób zostały one osiągnięte. Rozdział ósmy jest zwieńczony bardzo krótkim omówieniem najbardziej obiecujących kierunków dalszych prac badawczych i wdrożeniowych, które mogą zostać podjęte z wykorzystaniem wyników przedstawionych w rozprawie.

Rozprawa doktorska mgr. Roberta Kwiecińskiego prezentuje zarówno wiedzę teoretyczną jak i praktyczną Doktoranta w dyscyplinie *Informatyka*, a jej stronę merytoryczną oceniam pozytywnie. Na uwagę zasługuje fakt, że Autor udostępnił implementacje opracowywanych algorytmów oraz opracował rzeczywisty zbiór danych, który może zostać wykorzystany do rzetelnej oceny istniejących i nowych modeli rekomendacyjnych. Moim zdaniem zapewnienie reprodukowalności i przekrojowości eksperymentów jest obecnie szczególnie istotne, zwłaszcza w kontekście tzw. „kryzysu reprodukowalności" badań związanych z uczeniem maszynowym, z którym obecnie musimy się mierzyć [1, 2]. Warto również podkreślić to, że Autor z powodzeniem wdrożył część z opracowywanych technik i modeli w systemach Grupy OLX Sp. z o.o. – aspekty wdrożeniowe są często pomijane w literaturze w pracach dotyczących tworzenia modeli uczenia maszynowego, a istotnie wpływają na możliwość ich zastosowania w praktyce, zwłaszcza w przypadku metod o dużej złożonosci pamięciowej lub obliczeniowej (lub obu), które mają wypracowywać predykcje w krótkim czasie dla dużych i stale napływających danych.

Jestem przekonany, że Doktorant jest dobrze przygotowany do tego, żeby prowadzić dalsze badania związane z projektowaniem, implementowaniem, weryfikacją, walidacją i wdrażaniem systemów wykorzystujących techniki klasycznego i głębokiego uczenia maszynowego, zwłaszcza systemów rekomendacyjnych.

2.1 Pytania i uwagi

W trakcie lektury rozprawy nasunęły mi się następujące pytania i uwagi:

1. Pewien niedosyt pozostawia opis procesu anonimizacji opracowanego zbioru danych *OLX Jobs Interactions* (na str. 52), zwłaszcza punkty dotyczące odfiltrowania pewnej części interakcji oraz dodania pewnej liczby nieprawdziwych interakcji. Naturalne pytania, które od razu nasuwają się w trakcie lektury dotyczą tego, dlaczego niektóre interakcje zostały usunięte, a inne (losowe?) zostały dodane – być może celem było dodatkowe zaszumienie rzeczywistych danych zebranych w ramach Grupy OLX Sp z o.o. Niestety Autor podkreśla, że nie został upoważniony do ujawniania szczegółów tych operacji.

2. Na str. 53 Doktorant zauważa, że dwa z trzech benchmarkowych zbiorów danych (RecSys16 oraz RecSys17) nie są obecnie dostępne. Dlaczego trzeci ze wspomnianych zbiorów, CareerBuilder12, nie został wykorzystany w badaniach eksperymentalnych prowadzonych w ramach podjętych prac? W Tabeli 3.2 widzimy, że zbiór *OLX Jobs Interactions* jest zdecydowanie bogatszy niż pozostałe zbiory danych benchmarkowych, ale – jak sam Autor podkreśla na str. 53 – zbiór CareerBuilder12 jest jednym z najczęściej wykorzystywanych zbiorów danych zawierających interakcje użytkowników z ofertami pracy. Uważam, że wykorzystanie tego zbioru w badaniach eksperymentalnych (lub przynajmniej w wybranych eksperymentach) mogłoby pozwolić na bardziej „przekrojowe" porównanie opracowanych modeli z tymi znanymi z literatury.

3. Autor zaproponował pojedynczy podział zbioru na podzbiory treningowe i testowe (w podzbiorze testowym znalazło się 20% najnowszych interakcji), a dla zbioru testowego wykonano

dodatkowe operacje, które wpłynęły na jego zawartość (str. 61). W pracy zabrakło mi dokładniejszej analizy (ilościowej) obu podzbiorów – Autor przeprowadził taką analizę dla pełnego zbioru danych (przed podziałem na podzbiory treningowe i testowe). Doktorant wskazał również, że istnieją także inne podziały, które mogłyby być przydatne np. do oceny modeli wykorzystywanych do generowania rekomendacji w trakcie sesji użytkowników serwisu – czy Autor rozważał opracowanie takich dodatkowych podziałów zbioru *OLX Jobs Interactions*? Chciałbym również poznać opinię Doktoranta odnośnie walidacji krzyżowej – czy opracowanie kilku podziałów danych i przeprowadzenie walidacji krzyżowej mogłoby istotnie wpłynąć na ocenę modeli rekomendacyjnych? Czy Doktorant mógłby zaproponować scenariusz walidacji krzyżowej dla zbioru *OLX Jobs Interactions*?

4. Czy Autor (wraz z zespołem) rozważał zorganizowanie konkursu z wykorzystaniem zbioru *OLX Jobs Interactions*? Uważam, że taki konkurs, w ramach którego uczestnicy tworzyliby modele rekomendacyjne, byłby istotnym wkładem w rozwój takich systemów – opracowane modele mogłyby zostać przekrojowo zweryfikowane i porównane ze sobą z wykorzystaniem precyzyjnie zdefiniowanych zbiorów testowych i metryk oceniających zarówno ich aspekty funkcjonalne jak i niefunkcjonalne (np. czas inferencji, wymagania pamięciowe itd.).

5. Metody, które zostały wybrane do porównań (zebrane w Tabeli 4.1) należą do różnych grup metod rekomendacji, ale nie należą do najnowszych (2008–2016). Czy Autor rozważał porównanie opracowanych modeli rekomendacji z nowszymi technikami rekomendacji?

6. W sekcji 4.4.4, w której Autor omawia skalowalność analizowanych metod, czytamy: „W przypadku dwóch ostatnich, konieczna byłaby optymalizacja przed wdrożeniem w serwisach Pracodawcy, gdzie jak wspomnieliśmy wcześniej, modele wspólnej filtracji trenowane są nawet kilka razy dziennie." – czy Doktorant mógłby zaproponować kryteria akceptacji odzwierciedlające akceptowalne wymagania pamięciowe i obliczeniowe wdrażanych metod? Co oznacza, że „Wymagana pamięć jest stosunkowo duża, lecz dostatecznie mała, aby rozwiązanie mogło być wdrożone." (str. 81)?

7. Z czego wynika tak duża różnica w liczności grupy kontrolnej i testowej w eksperymencie A/B (Tabela 4.7, str. 83). Czy nie było możliwe zapewnienie równoliczności tych grup?

8. W sekcji 6.2 Autor skupia się na zaletach modelu P3LTR – jakie, według Doktoranta, są najistotniejsze *wady* tego modelu?

9. Na str. 110 czytamy: „Optymalne hiperparametry dla obu modeli raportujemy w tabeli 6.1." – czy możliwe jest formalne udowodnienie „optymalności" tych wartości hiperparametrów?

10. Według mnie, w pracy brakuje rysunków poglądowych, prezentujących np. architektury (i komponenty) grafowych sieci neuronowych, które znacznie ułatwiłyby zrozumienie i przyswojenie treści dysertacji.

11. W literaturze znane są algorytmy automatyczego tworzenia architektur głębokich i optymalizacji ich hiperparametrów – na jakiej postawie Autor wybrał algorytm optymalizacji takich hiperparametrów (str. 132)? Czy według Doktoranta można byłoby wykorzystać tutaj inne techniki optymalizacji hiperparametrów (lub automatycznego tworzenia architektur sieci głębokich)? Jeśli tak, to jakie?

12. Algorytm oceny istotności hiperparametrów (fANOVA) powinien zostać opisany w pracy – Autor zawarł tylko odniesienie do artykułu [65] (str. 133). Dlaczego Doktorant zdecydował się akurat na wykorzystanie podejścia fANOVA?

13. Na str. 135 czytamy: „Być może liczba parametrów tych modeli była zbyt duża w stosunku do zbioru danych, przez co model nadmiernie dopasował się do danych treningowych" – czy Autor w jakiś sposób weryfikował tę hipotezę?

14. Doktorant wskazuje, że najistotniejszą metryką, która jest głównym kryterium wyboru modelu w serwisach ogłoszeniowych Grupy OLX Sp. z o.o. jest *Precision@k*. W przypadku większości eksperymentów, Doktorant prezentuje wyniki eksperymentalne w postaci tabel, w których zawarte są wartości szeregu metryk dla zbioru testowego. Być może warto byłoby przedstawić (na wykresie) zależności pomiędzy metryką *Precision@k* i np. pokryciem zbioru testowego, entropią Shannona lub liczbą parametrów modelu?

15. W rozprawie miejscami pojawiają się niejasne sformułowania – na przykład we wprowadzeniu (str. 21) czytamy: „W momencie rozpoczęcia współpracy w serwisach Grupy OLX nie istniały modele rekomendacyjne dedykowane tej kategorii." (tj. kategorii „praca") – czy dla innych kategorii takie modele istniały? Na str. 46 Autor podkreśla, że „Właściwe zdefiniowanie i mierzenie metryk z pozostałych grup jest czasochłonnym zadaniem w stosunku do oczekiwanego przez nas wpływu." – wpływu czego na co?

16. W sekcji 1.3. („*Rozwój rekomendacji ofert pracy w OLX*") warto byłoby bardzo krótko (w 1-2 zdaniach) opisać, na czym polegały wspomniane testy A/B (uważam, że przeprowadzenie testów A/B jest bardzo istotnym aspektem prowadzonych badań i bez wątpienia są ciekawym elementem rozprawy).

3 Ocena strony formalnej rozprawy doktorskiej

3.1 Ocena układu pracy i uwagi redakcyjne

Układ rozprawy doktorskiej jest poprawny i logiczny, a kolejne rozdziały wprowadzają czytelnika w najistotniejsze aspekty prac – zarówno naukowych jak i wdrożeniowych – Doktoranta. W rozprawie zabrakło mi jednak rozdziału lub podrozdziału, np. w rozdziale 1 („*Wprowadzenie*"), w którym Autor opisałby dotychczas opublikowane artykuły dotyczące tematyki rozprawy, i których Doktorant jest (współ)autorem. W załączonej dokumentacji (oświadczenia współautorów) możemy zauważyć, że powstały trzy takie artykuły – jeden opublikowany w czasopiśmie *IEEE Access* (obecny współczynnik wpływu, z ang. *Impact Factor*, IF=3,9; 100 punktów ministerialnych), jeden przedstawiony na międzynarodowej konferencji *7th Conference on Computer Science and Intelligence Systems (FedCSIS 2022)* i jeden zgłoszony do publikacji w czasopiśmie *Expert Systems with Applications* (IF=8,5, 200 punktów ministerialnych). Uważam, że w pracy warto byłoby podkreślić aktywność publikacyjną Doktoranta.

3.2 Ocena zastosowanego piśmiennictwa

Bibliografia zawiera 139 pozycji, które zostały uporządkowane alfabetycznie. Dobór prac jest poprawny i jest jednym z dowodów na to, że Autor rozprawy dobrze orientuje się w aktualnym nurcie badań związanych z budowaniem i wdrażaniem systemów rekomendacyjnych. W trakcie analizy wykorzystanego piśmiennictwa zauważyłem niewielkie uchybienia:

- W niektórych wpisach w bibliografii możemy zauważyć niepoprawne wykorzystanie wielkich i małych liter, np. zamiast „*The use of the area under the ROC curve in the evaluation of machine learning algorithms*" czytamy „*The use of the area under the roc curve in the evaluation of machine learning algorithms*" (w [19]), czy zamiast „*A social network-based recommender system (SNRS)*" mamy „*A social network-based recommender system (snrs)*" (w [61]).

- W bibliografii możemy znaleźć niekompletne wpisy, w których brakuje np. stron czy nazwy czasopisma (lub konferencji) – np. [34]. Brakuje też spójności zapisu nazw konferencji i czasopism – niektóre są zapisane tylko skrótem: Int J Soft Comput Eng (w [38]), ACM Trans. Inf. Syst. (w [23]), a inne pełną nazwą.

3.3 Uwagi redakcyjne

Rozprawa jest napisana starannie, a tekst jest poprawny pod względem językowym, stylistycznym i interpunkcyjnym (zauważyłem tylko drobne błędy interpunkcyjne, np. nadmiarowe przecinki). W pracy zauważyłem drobne uchybienia, które jednak nie wpływają negatywnie na odbiór tekstu:

1. W streszczeniu (w j. angielskim) Autor używa zarówno brytyjskiego (np. „*optimised*") jak i amerykańskiego (np. „*familiarize*") angielskiego – powinno to zostać uspójnione.
2. W rozprawie możemy zauważyć niepotrzebne kalki z j. angielskiego, np. „adresować problem" (lepiej byłoby „podjąć" lub „rozwiązać" problem) lub „raportować wyniki" (przedstawić).
3. Autor używa dywizu zamiast (pół)pauzy (myślnika).
4. W pracy zauważyłem nieliczne literówki, np. „Dzięki temu możemy przejść do przeprowadzenie testów post-hoc (...)" (str. 112), „(...) w których wartość funkcja interakcji osiągałaby duże wartości (...)" (str. 125).
5. Wszystkie osie zawsze powinny być podpisane (np. na Rys. 4.4, 7.3).
6. Raczej unikałbym wykorzystywania skrótów jako tytułów (pod)sekcji, np. *4.1.1. ALS*, *4.1.3. SLIM*, itd.
7. Zachowanie tej samej kolejności modeli na rysunkach i w tabelach (np. na Rys. 4.2, Rys. 4.4, Rys. 4.5 i w Tabelach 4.3 czy 4.5) znacznie ułatwiłoby ich analizę.

Powyższe uchybienia nie wpływają na mój pozytywny odbiór rozprawy doktorskiej.

4 Konkluzja

Z pełnym przekonaniem stwierdzam, że recenzowana dysertacja Pana mgr. Roberta Kwiecińskiego **spełnia** wymagania stawiane rozprawom doktorskim przez Ustawę – praca prezentuje ogólną wiedzę teoretyczną Doktoranta w dyscyplinie *Informatyka* oraz umiejętność samodzielnego prowadzenia pracy naukowej, a przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie precyzyjnie zdefiniowanego problemu naukowego. W związku z powyższym, **wnioskuję o przyjęcie rozprawy doktorskiej oraz o dopuszczenie mgr. Roberta Kwiecińskiego do publicznej obrony**.

Jakub Nalepa

Literatura

[1] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns (Aug. 2023). doi:10.1016/j.patter.2023.100804.
URL `https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9`

[2] S. Kapoor, E. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik, P. Nanayakkara, R. A. Poldrack, I. D. Raji, M. Roberts, M. J. Salganik, M. Serra-Garcia, B. M. Stewart, G. Vandewiele, A. Narayanan, REFORMS: Reporting Standards for Machine Learning Based Science (2023). arXiv:2308.07832.