

Warszawa, 26 sierpnia 2017

dr hab. inż. Szymon Jaroszewicz  
Instytut Podstaw Informatyki PAN  
ul. Jana Kazimierza 5  
Warszawa

## Recenzja rozprawy doktorskiej mgr Marii Małgorzaty Rafalak

Promotor: dr hab. Jerzy Nowacki, prof. PJATK  
Promotor pomocniczy: dr inż. Radosław Nielek

### Uwagi ogólne

Rozprawa Pani mgr Rafalak składa się z trzech artykułów w tym dwóch opublikowanych w czasopismach z listy A MNiSW<sup>1</sup> i jednego konferencyjnego indeksowanego w Web of Science. Dodatkowo praca zawiera krótkie wprowadzenie podsumowujące stan wiedzy i uzyskane wyniki, a także życiorys naukowy doktorantki.

Pod względem bibliometrycznym, dysertacja spełnia zatem, w połączeniu z pozostałym dorobkiem kandydatki (który omówię dalej), wymagania zwyczajowe oraz ustawowe stawiane pracom doktorskim. Należy też zaznaczyć, że doktorantka jest pierwszą autorką wszystkich publikacji wchodzących w skład dysertacji, nie mam więc wątpliwości co do jej znaczącego wkładu w owe prace.

---

<sup>1</sup>jedno z nich nie jest już wprawdzie na liście A ale wciąż jest ujęte w punktacji pięcioletniej

## **Ocena wkładu w dziedzinę informatyki**

Należy podkreślić interdyscyplinarny charakter pracy. Praca dotyczy dziedziny informatyki społecznej i występuje w niej znaczący element socjologiczny. Nie mam jednak wątpliwości, że można ją uznać za pracę przede wszystkim z dziedziny informatyki. Po pierwsze praca jest ściśle związana z technologiami internetowymi: analizą ocen w kwestionariuszach internetowych, klasyfikacją użytkowników serwisów internetowych itd. Praca stosuje też metody typowe dla informatyki takie jak metody oparte o metryki czy klasteryzację. Większość nowatorskich elementów pracy stanowi wkład w rozwój owych metod, pomimo iż odbywa się to w kontekście informatyki społecznej.

## **Ocena merytoryczna uzyskanych wyników**

Przejdę teraz do oceny merytorycznej wyników uzyskanych przez doktorantkę.

Tematyka pracy dotyczy kilku zagadnień związanych z analizą opinii pochodzących z kwestionariuszy internetowych. Analizowane są przede wszystkim trzy problemy: projektowania adaptacyjnych kwestionariuszy zmniejszających potencjalną liczbę pytań, analizy wrażliwości systemów opartych o kwestionariusze internetowe oraz metodę klasyfikacji użytkowników na podstawie ankiet. Jak widać doktorantka systematycznie analizuje całość procesu: od zbierania danych w ankietach, poprzez interpretację uzyskanych wyników aż po analizę ich wiarygodności. Poniżej przedstawię moją ocenę kolejnych trzech prac wchodzących w skład dysertacji.

Pierwsza z nich dotyczy problemu redukcji ilości pytań w kwestionariuszu internetowym, tak by zwiększyć szansę, że użytkownik wypełni go w całości. Redukcja odbywa się poprzez zastosowanie zbioru uczącego składającego się z kompletnych, wypełnionych kwestionariuszy. Na tym zbiorze uruchamiany jest algorytm indukcji reguł AQ, a wynik jego działania zostaje użyty do oceny istotności poszczególnych pytań kwestionariusza i wyboru ich podzbiór pozwalającego na osiągnięcie wysokiej dokładności testu. Procedura jest podobna do selekcji zmiennych, jednak ich rolę pełnią tu pytania. Należy podkreślić, że praca nie polega na prostym zastosowaniu algorytmu AQ, autorka wprowadziła szereg dodatkowych mechanizmów umożliwiających skuteczne zastosowanie tego algorytmu do testów internetowych: pytania są oceniane, sortowane i wybierane tak by zapewnić dobre pokrycie dla wszystkich możliwych wyników testu. Jest to, oprócz zastosowania algorytmu AQ, nowatorskim wkładem pracy w literaturę przedmiotu.

Pewnym mankamentem pracy jest porównanie jedynie z losową selekcją

pytań a nie z bardziej zaawansowanymi metodami opisanymi w rozdziale 2.3. Nie zgadzam się też z uwagą, zwartą na stronie 19, że analiza czynnikowa jest trudna w interpretacji, praca zyskałaby na porównaniu z nią.

Kolejna praca dotyczy klasyfikacji respondentów do zdefiniowanego zestawu kategorii na podstawie wypełnionych przez nich kwestionariuszy. Podejście jest oparte o metody metryczne. W pracy przedstawiono szereg nowatorskich rozwiązań dostosowujących te metody do danych pochodzących z kwestionariuszy co stanowi istotny wkład w dziedzinę informatyki społecznej. Po pierwsze zaproponowano dwie ogólne miary odległości: jedną traktującą każdą odpowiedź oddzielnie i drugą podobną do klasycznej metody najbliższych sąsiadów opartą o porównywanie wektorów wszystkich odpowiedzi danego testu. Po drugie testowano dwie miary odległości: klasyczną odległość kosinusową oraz rzadziej stosowaną w tym kontekście metrykę Wessersteina.

Zaproponowano również metodę poprawiającą separację klas polegającą na przybliżaniu przykładów uczących do najbardziej reprezentacyjnego dla nich wzorca. Metoda jest interesująca i wyraźnie poprawia wyniki eksperymentalne. Trochę niejasności wprowadza określanie tego podejścia metodą obwiedni (*envelope*) co nie oddaje jej istoty. Dodatkowo definicja obwiedni nie jest precyzyjna (*area representing the distribution*).

Jednym z najważniejszych elementów nowości tej pracy są metody oceny separacji klas. Jest to bardzo przydatne w analizie danych z kwestionariuszy pozwala bowiem ocenić ich wartość do oceny konkretnej cechy. Miary są oparte o odległość kosinusową oraz o metrykę Wessersteina. Wykorzystanie tych odległości do liczbowego oszacowania trudności klasyfikacji nie było trywialne i opracowanie miar separacji klas stanowi istotny wkład w dziedzinę analizy danych ankietowych. Znajduje to potwierdzenie w wynikach eksperymentów, w szczególności zawartych w tabelach 2 i 3 zestawionych z tabelą 4. Wyniki te pokazują, że miary te dobrze korelują się z ostateczną dokładnością klasyfikacji.

Brakuje nieco analizy teoretycznych własności tych miar, jednak spowodowałyby to zapewne nadmierny wzrost objętości artykułu.

Trzecia praca wchodząca w skład rozprawy ukazała się w prestiżowym czasopiśmie ACM Transactions on the Web i dotyczy zagadnienia dokonywanych przez użytkowników ocen treści internetowych. Idea proponowanej metody polega na porównywaniu empirycznych rozkładów ocen przy pomocy metryki EMD (*Earth Movers Distance*) czyli szczególnego przypadku metryki Wessersteina. Jest to podejście odmienne – oferujące bogatszą semantykę – od dotychczasowych, które bazowały na parametrach rozkładów takich jak średnia czy odchylenie standardowe.

Podejście oparte jest o klasteryzację rozkładów ocen w czterech różnych zbiorach danych. Dwa spośród zbiorów danych użytych do badań zostały

opracowane w PJATK i są interesujące z uwagi na fakt, że zawierają też dość szczegółowe dane użytkowników, w tym ich profile psychologiczne. Uzyskane klastry autorzy przypisali do klasy pozytywnej, negatywnej bądź kontrowersyjnej. Pozostałe przypadki przypisano do klasy neutralnej. Tak uzyskane klasy posłużyły do ilustracji zjawisk zachodzących w kwestionariuszach internetowych (np. *positive bias*) oraz jako punkt odniesienia do oceny proponowanych ogólnych definicji klas.

Jednym z najważniejszych osiągnięć przedstawionych w pracy są ogólne, niezależne od konkretnego zbioru danych, definicje klasy pozytywnej, negatywnej, kontrowersyjnej oraz neutralnej. Klasy zdefiniowane zostały w oparciu o odległość EMD od rozkładów prototypowych oraz dodatkowy próg na prawdopodobieństwa niektórych kategorii. Wyjątkiem jest klasa neutralna, która zdefiniowana jest po prostu jako zbiór rozkładów nie należących do żadnej z pozostałych klas. Zaproponowano również zmodyfikowaną definicję dla klasy kontrowersyjnej wykorzystującą dodatkowo miarę zgodności Leika.

W przypadku klas o ogólnym przeznaczeniu kluczowa jest ich walidacja. Autorzy włożyli dużo wysiłku aby przekonać czytelnika, że zaproponowane definicje klas są naprawdę ogólne. Po pierwsze, porównano (tabela IV) przypisania do poszczególnych klas uzyskane na podstawie klasteryzacji przeprowadzonej oddzielnie na każdym zbiorze z przypisaniami uzyskanymi na podstawie proponowanych ogólnych definicji. Zbieżność przewidywań jest dość dobra i wydaje się być wystarczająca dla zastosowań praktycznych. Kolejnym etapem było porównanie z prostszymi definicjami opartymi o średnie ocen. Autorzy wykazali, że owe prostsze definicje są szczególnymi, suboptymalnymi, przypadkami definicji proponowanych przez nich. Podsumowując, wyniki walidacji są przekonujące zatem proponowane definicje należy uznać za znaczący wkład w dziedzinę informatyki społecznej.

Jeszcze bardziej przekonującą metodą walidacji byłoby przetestowanie proponowanych definicji na kolejnym, piątym zbiorze danych, który nie był używany w projektowaniu tych definicji. Szkoda, że autorzy nie zdecydowali się na takie rozwiązanie. Brak ten rekompensuje po części analiza wrażliwości opisana poniżej.

Obszerna część pracy poświęcona jest analizie wrażliwości ocen na szereg czynników socjologicznych i psychologicznych charakteryzujących użytkowników wystawiających oceny. Na podstawie porównania rozkładów odległości EMD ocen w danej podgrupie i całej populacji autorzy stwierdzają, że wpływ ten jest niewielki, a co za tym idzie, ogólne definicje klas powinny działać poprawnie również w podgrupach. Analiza jest bardzo interesująca i szczególnie. Pewnym mankamentem jest to, że nie przetestowano wpływu tych czynników na rozkład poszczególnych klas, co mogłoby być bardzo interesujące.

Na koniec mam uwagę dotyczącą stosowanego testu statystycznego. Wbrew obiegowej opinii, którą autorzy cytują na stronie 54, hipotezą zerową testu Kruskalla-Wallisa nie jest równość rozkładów we wszystkich grupach, lecz brak stochastycznej dominacji. Na przykład test ten nie jest w stanie rozróżnić rozkładów symetrycznych o zerowej medianie.

## Ocena pozostałego dorobku kandydatki

Pozostały dorobek kandydatki obejmuje pięć prac, w trzech z nich jest ona pierwszą autorką. Pani Rafalak była laureatką szeregu nagród międzynarodowych (m. in. stażu w firmie Huawei) i jest członkiem International Association for Computer Adaptive Testing.

## Inne uwagi

Na koniec wymienię kilka drobnych błędów językowych i terminologicznych:

str. 2 “if wanted to be used”

str. 2 “with to the average”

str. 4 “rule induction based method” powinno być “... methods”

publikacja 1, str. 16 “proposed method significantly reduce” powinno być “... reduces”

publikacja 1, str. 35 “geometrics spaces” powinno być “geometric spaces”

publikacja 2, str. 1262 “area representing the distribution” domyślam się, że chodzi o “support of the distribution”

publikacja 3, str. 58 termin “cumulative distributions” może być mylący bo są to raczej rozkłady uśrednione a nie dystrybuanty (cumulative distribution functions)

## Podsumowanie

Podsumowując, stwierdzam że praca spełnia z nawiązką wymagania ustawowe stawiane dysertacjom doktorskim w dziedzinie informatyki i stawiam wniosek o dopuszczenie doktorantki do publicznej obrony.

Szymon Jaroszewicz