Warszawa, 12.12.2016

dr hab. inż. Arkadiusz Orłowski, prof. SGGW
Katedra Informatyki
Wydział Zastosowań Informatyki i Matematyki SGGW
ul. Nowoursynowska 159
02-787 Warszawa

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy:	***Integracja metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji***
Autor rozprawy:	**mgr inż. Marcin Gibert**
Promotor:	**dr hab. Bożena Śmiałkowska**
Promotor pomocniczy:	**dr inż. Jarosław Jankowski**

Recenzję sporządzono na wniosek WI/Dokt-160/2016 Dziekana Wydziału Informatyki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie.

Recenzowana praca liczy 164 strony. Składa się z obszernego szesnastostronicowego Wprowadzenia, pięciu rozdziałów merytorycznych (w tym odrębnego rozdziału poświęconego dyskusji wyników i weryfikacji hipotezy badawczej), trzystronicowego Podsumowania, listy Referencji zawierającej wykorzystaną literaturę (6 stron), Spisu rysunków, Spisu tabel oraz Spisu symboli.

Rozprawa ma charakter teoretyczny, z elementami opisowymi. Zawiera także testy różnych wariantów opracowanej przez Autora procedury integracji dwóch metod eksploracji danych. Zasadnicze wyniki rozprawy zawarte są w Rozdziałach 4, 5 i 6.

W pierwszej części Wprowadzenia omówione są różne etapy procesu podejmowania decyzji oraz najważniejsze metody eksploracji danych wykorzystywane w tym procesie. Podano krótkie uzasadnienie, dlaczego w dalszej części pracy skoncentrowano się na metodzie klasyfikacji. Zbiór wszystkich danych przedstawiono jako teoriomnogościową sumę zbiorów danych różnych typów, podkreślając praktyczne znaczenie danych tekstowych i danych numerycznych. Omówiono problem reprezentacji danych i ich wpływ na wynik eksploracji. Opisując model przestrzeni wektorowej, najczęściej używane podejście do eksploracji danych tekstowych, wymieniono trzy propozycje reprezentacji danych, w zależności od bazy przestrzeni cech. Przedstawiono i zdefiniowano pojęcie nośności informacyjnej danych, podkreślając trudności z praktycznym wyznaczeniem tej wielkości bezpośrednio z podanej definicji, sugerując wykorzystanie w tym celu metod porównawczych. Przedyskutowano problem pomiaru jakości decyzji, wyróżniając metody bazujące na wynikach klasyfikacji. W miarę szczegółowo omówiono miary jakości wykorzystujące „macierz pomyłek". Krótko uzasadniono wybór współczynników całkowitej dokładności (ACC) i całkowitego poziomu błędu (ERR) jako miar oceny jakości decyzji wykorzystywanych w rozprawie.

Część druga Wprowadzenia poświęcona jest sformułowaniu problemu badawczego, związanego z jednoczesną eksploracją danych tekstowych i danych numerycznych. Sformułowano sześć pytań związanych ze zidentyfikowanymi wadami bądź brakami istniejących metod eksploracji danych w procesie podejmowania decyzji i na podstawie analizy czterech z nich przyjęto główny cel rozprawy, którym jest „*opracowanie procedury integracji metod analizy fleksyjnej tekstu oraz metod eksploracji danych numerycznych*". Analiza dwóch pozostałych pytań pozwoliła sformułować hipotezę badawczą, która mówi, że „*integracja metod analizy fleksyjnej tekstu oraz eksploracji danych numerycznych zwiększy nośność informacyjną danych w wielokryterialnym procesie wspomagania decyzji*".

W trzeciej, ostatniej, części Wprowadzenia zwięźle i precyzyjnie omówiono zakres i strukturę pracy.

W Rozdziale 2 przedstawiono różne problemy napotykane przy klasyfikacji danych tekstowych, z naciskiem na wykorzystanie modelu przestrzeni wektorowej VSM (w podejściu uczenia maszynowego) oraz znajomości reguł leksykalnych i składniowych języka i analizę znaczeniową tekstu (w podejściu eksperckim). Oba podejścia bardzo szczegółowo omówiono w dwóch kolejnych podrozdziałach. Na zakończenie Rozdziału 2, na podstawie istniejącej literatury przedmiotu, porównano trzy metody eksploracji danych tekstowych (automatyczna z VSM, automatyczna bez VSM, definiowanie reguł i wzorców przez eksperta), wykorzystując w tym celu analizę SWOT.

W Rozdziale 3 przedstawiono problem klasyfikacji danych numerycznych. Po krótkim przedstawieniu najważniejszych, znanych w literaturze, metod eksploracji danych numerycznych, skupiono się na podejściu wykorzystującym teorię zbiorów przybliżonych. Podkreślono wagę właściwego wyboru atrybutów oraz ich wpływ na postać reguł decyzyjnych. Przedyskutowano także znaczenie ich właściwej reprezentacji, która odgrywa dużą rolę w eliminacji (a przynajmniej redukcji) szumu informacyjnego związanego z eksploracją danych. Dużo uwagi poświęcono generowaniu reguł decyzyjnych metodą TZP. W ostatnim podrozdziale przeprowadzono badania istotności i zgodności wyników klasyfikacji za pomocą nieparametrycznego testu statystycznego McNemara (*de facto* test chi-kwadrat), który można stosować w przypadku prób zależnych z uwzględnieniem zmiennych dychotomicznych.

Rozdział 4 jest zasadniczą częścią pracy i szczegółowo omawia autorską procedurę integracji metod klasyfikacji dyskutowanych w dwóch poprzednich rozdziałach. Po przedstawieniu ogólnego schematu procedury integracji metod eksploracji danych tekstowych i numerycznych w procesie podejmowania decyzji, gruntownie i krytycznie omówiono poszczególne etapy. Szczegółowo opisano zarówno wstępną eksplorację danych tekstowych (w tym zdefiniowanie wzorców informacyjnych a następnie realizacja etapu 2 procedury: segementacja, lematyzacja, redukcja reprezentacji tekstu, *etc.*) jak też właściwą eksplorację danych tekstowych (etap 3 procedury: macierz reprezentująca dokumenty tekstowe, obliczenie wag, niejawna analiza semantyczna, obliczanie podobieństwa, klasyfikacja metodą kNN). Opracowano reprezentację danych numerycznych (dyskretyzacja, ustalenie wartości nominalnych, kodowanie). Na zakończenie omówiono klasyfikację danych w procesie decyzyjnym z użyciem eksploracji danych numerycznych.

W Rozdziale 5 Autor przeprowadza badania testowe. Rozważono cztery warianty eksploracji: (A) zintegrowana eksploracja danych tekstowych i numerycznych, (B) eksploracja wyłącznie danych numerycznych, (C) eksploracja wyłącznie danych tekstowych, (D) zintegrowany wynik eksploracji według

wariantów (B) i (C). W każdym z przypadków (A), (C) i (D) wykorzystano trzy różne reprezentacje danych tekstowych (unigramowa, n-gramowa, gamma-gramowa). Badania przeprowadzono niezależnie dla trzech różnych scenariuszy (procesów podejmowania decyzji), mających (przynajmniej potencjalnie) duże znaczenie praktyczne: wybór rentownych zamówień publicznych, sposób inwestowania na Warszawskiej Giełdzie Papierów Wartościowych i wyszukiwanie atrakcyjnych ofert pracy.

Rozdział 6 poświęcony jest pogłębionej dyskusji otrzymanych wyników oraz uzasadnieniu prawdziwości postawionej w rozprawie hipotezy badawczej w trzech analizowanych wcześniej przypadkach wykorzystania różnych wariantów procedury eksploracji danych. Jego zasadniczym elementem są tabele przedstawiające, między innymi, porównanie wartości miar jakości decyzji osiągniętych w różnych analizowanych przypadkach procesów decyzyjnych. Trzy obszerne punkty zamykające ten rozdział zawierają krytyczną analizę wspomnianych wyżej tabel i wynikające z tej analizy uzasadnienie prawdziwości hipotezy badawczej w każdym z badanych przypadków.

Pracę zamyka Podsumowanie, w którym Autor streszcza uzyskane rezultaty oraz uzasadnia, dlaczego jego zdaniem postawiony cel pracy został osiągnięty, a hipoteza udowodniona. W trzech punktach przedstawiono najważniejsze elementy autorskiego wkładu do omawianej w rozprawie problematyki. W ostatnim akapicie Podsumowania przedstawiona jest sugestia dalszych prac nad tymi zagadnieniami, w szczególności opracowanie systemu wspierającego proces decyzyjny, wykorzystującego opracowaną w Rozdziale 4 pracy procedurę eksploracyjną oraz automatyczne wyszukiwanie informacji dostępnych na stronach internetowych tworzonych w ramach Internetu semantycznego.

Bibliografia jest obszerna i zasadniczo kompletna - w rozprawie wykorzystano 110 pozycji literatury przedmiotu. Są wśród nich monografie, artykuły naukowe, prace przeglądowe, standardy oraz materiały zamieszczone na stronach internetowych, w tym trzy pozycje których Autor rozprawy jest autorem lub współautorem. Cytowane prace są w ogromnej większości istotne dla omawianej problematyki.

Uwagi krytyczne dotyczące rozprawy rozpocznę od obserwacji, że choć praca jest dobrze i przystępnie napisana z merytorycznego punktu widzenia, to zawiera bardzo dużo irytujących literówek, lapsusów gramatycznych i błędów interpunkcyjnych. Na szczęście nie obniżają one wartości naukowej rozprawy.

Przy formułowaniu celu pracy brakuje mi krótkiego uzasadnienia, wykraczającego poza odniesienie się do pytań 1, 2, 3 i 6, dotyczących braków i niedociągnięć istniejących metod eksploracji danych. Nie jest precyzyjnie wyjaśnione, poza wzmianką, że wynika to z analizy odpowiedzi na wyżej wspomniane pytania, po co tworzymy taką procedurę integracyjną. W pewnym sensie uzasadnieniem jest oczywiście sama hipoteza badawcza (a właściwie jej prawdziwość), ale lepiej byłoby to jawnie napisać.

Pojęcie nośności informacyjnej, zdefiniowane w pierwszej części Wprowadzenia, jest stosunkowo mało znane i przydałby się szerszy kontekst, zwłaszcza aplikacyjny i w odniesieniu do literatury anglojęzycznej. Warto byłoby też wyjaśnić różnice pomiędzy znaczeniem tego terminu w niniejszej rozprawie, a niezależnym choć pokrewnym i podobnie brzmiącym oraz znacznie bardziej popularnym w polskim środowisku statystycznym pojęciem metody wskaźników pojemności informacji (metoda optymalnego doboru zmiennych objaśniających w modelach ekonometrycznych, autorstwa Zdzisława Hellwiga).

Nie jest jasne, czy i jak w problemach klasyfikacji dokonano sprawdzianu krzyżowego, np. podziału na zbiór uczący i zbiór walidacyjny. Trudno więc ocenić niektóre aspekty procedury, np. ryzyko przeuczenia. Przykładowo, na stronie 99 czytamy: „zbiór 200 przypadków zamówień publicznych poddano klasyfikacji". Ani słowa o tym, czy przeprowadzono kroswalidację, choćby w formie losowego podziału na zbiór uczący (treningowy) i walidacyjny (testowy). Podobnie na stronie 103 Autor pisze, że badanie „przeprowadzano na zbiorze 200 testowych przypadków transakcji z wykorzystaniem 11 przypadków treningowych dla kategorii oznaczającej spadek kursu akcji". Nie jest jasne jak to dokładnie rozumieć. Czy pozostałe 189 przypadków należało do kategorii oznaczającej wzrost kursu akcji? Czy też może przypadki te były zróżnicowane (zarówno wzrost jak i spadek) i część z nich (jaka?) posłużyła jako zbiór testowy? Czy 11 elementów wylosowano? Tego typu nieprecyzyjnych stwierdzeń jest więcej.

Jako miary decyzji stosowane są ACC i ERR, zdefiniowane w Rozdziale 3 na stronie 63. Autor nie zauważył jednak (a przynajmniej o tym nie wspomniał), że ich jednoczesne stosowanie jest w redundantne, ponieważ ich suma jest zawsze równa 1. Według Autora ich główna zaletą jest fakt, że w obu przypadkach wykorzystane są wszystkie współczynniki (składowe *confusion matrix*) charakteryzujące klasyfikację (TP, TN, FP, FN). Dlatego rezygnuje z innych miar, które powinny być stosowane parami. Uważam jednak, że w

niektórych sytuacjach wykorzystanie innych miar, jak na przykład para *czułość* i *specyficzność*, jest równie, a może nawet bardziej, uzasadnione. Przykład inwestowania na WGPW omawiany w Podrozdziale 5.3 należy, moim zdaniem, do tej kategorii.

Co oznacza znak zapytania w drugiej kolumnie tabeli 3 ze strony 36? Brak wyekstrahowanego wyrazu?

Na stronie 86 czytamy: „Dla określonego poziomu istotności α=0,05 przeważająca większości wyników badań za pomocą testu McNemara posiada wartość *p-value* poniżej wartości α …". Z załączonych tabel 15-17 wynika, że jest tak dla wszystkich przypadków.

Nagłówek tabeli 44 jest na wcześniejszej stronie niż reszta tabeli. W opisie kolumn tabeli 66 pisałbym raczej o liczbie niż o ilości (chodzi o sztuki). W niektórych miejscach nie jest jasne czy Autor mówi o procentach czy o punktach procentowych.

Nie sposób cytować wszystkie książki na określony temat, ale trochę szkoda, że Autor nie dotarł do ciekawego podręcznika wydanego w 2009 roku przez PWN (Z. Markov, D. T. Larose, *Eksploracja zasobów internetowych*), który porusza sporo bardzo zbliżonych zagadnień.

Na stronie 11 rozprawy czytamy: „autorzy artykułu „A Roadmap for Web Mining: From Web to Semantic Web" wskazują …". Zwykle Autor rozprawy cytuje literaturę przedmiotu podając jedynie numer pozycji w spisie Referencji. Tym razem zrobił wyjątek. Pozycja [7] o którą najwyraźniej chodzi, ma jednak nieco inny tytuł.

Mimo powyższych uwag krytycznych oceniam, że uzyskane przez autora wyniki zostały zaprezentowane właściwie. Cel postawiony na początku rozprawy został osiągnięty: Autor opracował procedurę integracji metod eksploracji danych tekstowych i danych numerycznych w procesie podejmowania decyzji. Przeprowadzone testy wykazały, że taka integracja istotnie zwiększa zdefiniowaną w rozprawie nośność informacyjną, co należy uznać za wystarczające uzasadnienie prawdziwości postawionej w rozprawie hipotezy badawczej. Tematyka rozprawy bez wątpienia należy do dyscypliny naukowej *Informatyka* w dziedzinie nauk technicznych a przedstawione rozwiązanie postawionego w niej problemu naukowego świadczy o zdolności jej Autora do prowadzenia pracy naukowej.

Podsumowując, uważam, że rozprawa doktorska magistra inżyniera Marcina Giberta stanowi interesujący wkład w rozwój ważnej, aktualnej i wciąż zyskującej na znaczeniu problematyki wykorzystania metod eksploracji danych do wspomagania procesu podejmowania decyzji. Tematyka podjęta w rozprawie z pewnością będzie intensywnie rozwijana, zwłaszcza w kontekście podejmowania decyzji politycznych (np. konieczność bieżącej analizy komentarzy na blogach i forach internetowych w trakcie kampanii wyborczych) i decyzji biznesowych. Poziom naukowy recenzowanej rozprawy doktorskiej oceniam pozytywnie i, wobec wypełnienia wymogów ustawowych, wnoszę o dopuszczenie jej autora do kolejnych etapów przewodu doktorskiego.

A Ostrowski