Częstochowa, dn. 12 czerwca 2018 r.

Dr hab. inż. Rafał Scherer, prof. Politechniki Częstochowskiej
Instytut Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej mgra inż. Marcina Pietrasa, pt.: Syntaktyczna i semantyczna analiza danych tekstowych z wykorzystaniem modeli Markowa realizowanych sprzętowo.

Promotor: dr hab. inż. Przemysław Klęsk, prof. nadzw.

Niniejszą recenzję opracowano na zlecenie Dziekana Wydziału Informatyki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie, z dnia 18.04.2018 r.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Analiza tekstu i języka naturalnego są jednymi z głównych składników obecnej rewolucji związanej z rozwojem metod sztucznej inteligencji. Rozwój tych metod umożliwia coraz sprawniejszą interakcję z maszynami w sposób podobny do rozmowy z drugim człowiekiem. Rozwój metod naukowych wraz ze znaczącym ulepszeniem technologii obliczeniowych, szczególnie akceleratory GPU i technologie chmurowe, pozwolił na stworzenie systemów udających ludzkie odpowiedzi. Ciągle jednak stworzenie efektywnych i wydajnych sprzętowo systemów analizy tekstu stanowi ogromne wyzwanie.

Autor zauważa w tezie pracy, że *Sprzętowe realizacje ukrytych modeli Markowa o zredukowanej reprezentacji zmiennoprzecinkowej pozwalają na analizę sekwencji tekstowych z zadowalającą dokładnością wykorzystując logarytmowane wersje algorytmów Viterbiego oraz Forward-Backward.* Autor stworzył ukryte modele Markowa (HMM) do ekstrakcji tekstu ze stron WWW oraz do analizy syntaktycznej tekstu. Dokonał również umocowanej teoretycznie zlogarytmizowanej implementacji sprzętowej w układach FPGA.

2. Zawartość rozprawy

Recenzowana praca mgra inż. Marcina Pietrasa składa się z 8 rozdziałów zasadniczych, oraz dodatku, w którym umieszczono zbiory danych wykorzystane w rozprawie oraz wykonane

oprogramowanie. Praca zawiera również obszerną bibliografię, spisy rysunków i tabel. Dokument liczy 144 strony.

Rozdział pierwszy zawiera cel i tezę pracy oraz wprowadzenie do analiza syntaktycznej i semantycznej tekstu wraz z krótkim przeglądem literatury.

Rozdział drugi jest wprowadzeniem do ukrytych modeli Markowa. Opisano ogólną strukturę modelu wraz z *n*-gramową reprezentacja tekstu oraz modelem warstwowym. Następnie omawiana jest rozszerzona arytmetyka w dziedzinie logarytmicznej związana z zagadnieniami stabilności numerycznej będącej przedmiotem późniejszych rozdziałów. Pozwala ona uniknąć sytuacji, w której wartości prawdopodobieństwa będą równe zero. Zaprezentowano metody generowania modeli HMM z danych. Ostatecznie Autor opisuje autorski system do uczenia HMM i ich zapisu w formacie XML.

Rozdział 3 jest przeglądem metod ekstrakcji informacji ze stron WWW w formacie HTML. Autor zwraca uwagę na ograniczenia formatu HTML związane z pierwotnymi założeniami formatu, który miał służyć jedynie do prezentacji treści w sieci internet. Dlatego też, trudno jest wyekstrahować istotne informacje z całej zawartości strony WWW. Omówiono rodzaje reprezentacji danych, takich jak drzewo DOM. Zaprezentowano metodę ekstrakcji i modelowania wiedzy ze stron WWW za pomocą HMM.

W rozdziale czwartym Autor omawia syntaktyczną analizę treści, oraz oznaczone zbiory danych (korpusy), tzw. banki drzew, zawierające zdania przeanalizowane pod względem składni, a analiza ta jest reprezentowana za pomocą drzew. Analizowane są banki frazowe i zależnościowe oparte o różne modele. Omówiona jest również specyfika języka polskiego w konstruowaniu zależności banku. Autor omówił metodologię i stworzył modele HMM *n*-gramowe i oparte o afiksy dla języka polskiego oraz angielskiego wykorzystując najbardziej znane banki drzew. Podał też wyczerpujące analizy jakości i przykłady analizy zdań dla języka polskiego i angielskiego za pomocą stworzonych modeli HMM.

Rozdział piąty omawia aspekty analizy semantycznej na różnych poziomach. Autor przedstawia własna propozycję implementacji struktur semantycznych w systemie HMM-Toolbox.

W rozdziale 6 Autor rozważa zagadnienia stabilności numerycznej oraz proponuje algorytmy wykorzystujące zlogarytmizowaną, zredukowaną wersję obliczeń. Udowodniono jednocześnie bezpieczeństwo numeryczne zlogarytmizowanych modeli Markowa. Należy podkreślić, że wyniki prezentowane w tym rozdziale zostały opublikowane w czasopiśmie z listy JCR.

Rozdział 7 stanowi dokumentację autorskich algorytmów akceleracji sprzętowej za pomocą układów FPGA. Ponieważ Autor wykazał w poprzednim rozdziale, że zlogarytmizowanie obliczeń zapewnia stabilność numeryczną, użył tu jednostce aproksymacji wielu funkcji logarytmicznych i wykładniczych „MLEAU".

W rozdziale ósmym Autor podsumowuje pracę oraz przedstawia nowatorskie elementy pracy i najważniejsze oryginalne wyniki.

Do pracy dołączona jest bardzo obszerna bibliografia i dodatki w formie elektronicznej zawierające zbiory danych oraz autorskie oprogramowanie.

Ogólnie zasadnicze i oryginalne rezultaty pracy można podsumować następująco:

- Analiza współczesnych trendów w analizie tekstu,
- Stworzenie modeli Markowa dla ekstrakcji istotnego tekstu ze stron WWW,
- Opracowanie modeli Markowa dla analizy części mowy i zdania,
- Nauczenie modeli z wykorzystaniem dwóch oznaczonych korpusów,
- Sprzętowa implementacja wymienionych wyżej metod w układach FPGA.

Wymienione oryginalne metody przedstawione w pracy zostały opublikowane w kilku artykułach naukowych, głównie w materiałach konferencyjnych. Lemat z dowodem „o długości sekwencji niebezpiecznej numerycznie" został opublikowany w czasopiśmie z Listy Ministerialnej A. Zaprezentowany materiał pokazuje, że Doktorant udowodnił tezę pracy.

3. Uwagi krytyczne i wskazówki dotyczące rozprawy

Praca napisana jest wyjątkowo schludnie i przejrzyście. Każdej metodzie towarzysza obszerne wyjaśnienia i czytelne rysunki oraz schematy.

Rozumiejąc specyfikę podjętej tematyki, czy możliwe jest porównanie jakości i szybkości działania przedstawionych metod z innymi w literaturze?

Czy możliwe jest porównanie zaprezentowanych metod z wynikami osiąganymi obecnie przez sieci neuronowe?

Patrząc pod kątem edytorskim należy podkreślić, że w tak obszernym dokumencie występuje jedynie kilka nieistotnych błędów: tzw. literówek lub błędów interpunkcyjnych, np.:

str. 44, brak przecinka „... w głównych językach (angielski chiński i arabski). „

str. 28, „ogólne" w „Dodatkowym celem może być ogólne semantyczna kategoryzacja składowych treści."

str. 30, „muszę" w „Nie dziwi więc fakt, że przeglądarki internetowe muszę w dużym stopniu tolerować błędy i luźną składnię HTML."

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że Pan mgr inż. Marcin Pietras w rozprawie doktorskiej „Syntaktyczna i semantyczna analiza danych tekstowych z wykorzystaniem modeli Markowa realizowanych sprzętowo":

- Zrealizował cel i zweryfikował hipotezę rozprawy,
- Uzyskał oryginalne rezultaty naukowe dotyczące syntaktycznej i semantycznej analizy danych tekstowych,
- Dokonał analizy współczesnych trendów w analizie tekstu,
- Stworzył modele Markowa dla ekstrakcji istotnego tekstu ze stron WWW,
- Opracował modele Markowa dla analizy części mowy i zdania,
- Nauczył zaprojektowane modele z wykorzystaniem dwóch oznaczonych korpusów,
- Dokonał sprzętowej implementacji wymienionych wyżej metod w układach FPGA.
- wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i bogatą wiedzą w dziedzinie uczenia maszynowego oraz programowalnych układów FPGA.

Recenzowana praca spełnia wymagania ustawy o tytule i stopniach naukowych w dyscyplinie naukowej Informatyka. Wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony. Jednocześnie, ze względu na wyjątkowo wysoki poziom naukowy i aplikacyjny, wnioskuję o wyróżnienie rozprawy o ile jest to zgodne z zasadami przyjętymi przez Radę Wydziału Informatyki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie.