Prof. dr hab. inż. Sebastian Deorowicz,
Instytut Informatyki
Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska
44-100 Gliwice, ul. Akademicka 16

Gliwice, 5.02.2019 r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy:

Efficient Data Structures for Internal Queries in Texts
(Efektywne struktury danych dla zapytań wewnętrznych w tekstach)

Autor:

mgr Tomasz Kociumaka

Promotor:

Prof. dr hab. Wojciech Rytter

I. Problematyka naukowa oraz przedmiot rozprawy

Tekst jest naturalnym sposobem reprezentowania wiadomości stosowanym przez ludzi od wieków. Także w przypadku danych przechowywanych w pamięciach komputerów ta forma reprezentacji jest również często spotykana, zarówno bezpośrednio (teksty) jak i pośrednio (np. sekwencje biologiczne zapisywane jako ciągi symboli reprezentujących nukleotydy czy też aminokwasy). Czasami w taki sposób zapisuje się także liczby, co ułatwia ich analizę przez człowieka.

Kiedy mówimy o reprezentowaniu informacji w formie tekstowej, to w naturalny sposób rodzi się pytanie o bardziej dalekosiężny cel. Oczywiście samo tylko przechowywanie takowym nie jest. Ważniejsza jest możliwość wydobycia czy wyszukania informacji, które mogą być dla nas ważne. Biorąc pod uwagę to, że długości tekstów, które można mieć na myśli są liczone w milionach, miliardach czy nawet bilionach znaków, to rozważania dotyczące optymalnych algorytmów czy też struktur danych rozwiązujących rozmaicie postawione problemy wyszukiwania w tekście stanowią niezaprzeczalnie bardzo ważną tematykę badań nie tylko z punktu widzenia teoretycznego, ale także praktycznego.

Oceniana rozprawa stanowi właśnie przykład wszechstronnego i dogłębnego spojrzenia na problemy wyszukiwania, czy też bardziej ogólnie, wykonywania zapytań wewnętrznych w tekstach. Autor pokazuje tutaj liczne techniki algorytmiczne i interesujące struktury danych, których czasy konstrukcji, realizacji zapytań są optymalne, bądź konkurencyjne do znanych technik.

II. Analiza treści rozprawy oraz uzyskanych wyników

1. Treść rozprawy

Doktorant postawił sobie w pracy ambitne cele znalezienia lepszych niż znane struktur danych i algorytmów na nich operujących, które umożliwiają efektywną realizację zapytań wewnętrznych w tekstach.

Rozprawa napisana jest w języku angielskim i składa się 10 rozdziałów. W pierwszym rozdziale Doktorant wprowadza podstawowe pojęcia, którymi posługuje się w dalszej części pracy. Definiuje też szczegółowe problemy, które będą rozważane, formułując przy tym główne twierdzenia dowiedzione w kolejnych rozdziałach. Autor krótko przedstawia tutaj także zaproponowane bądź ulepszone techniki, które zostały wykorzystane do opracowania struktur danych będących wynikiem pracy.

Rozdział drugi stanowi wprowadzenie w tematykę algorytmów operujących na tekstach. Przedstawione są tutaj podstawy algorytmów kombinatorycznych, okresów, kompresji. Nie jest to oczywiście przegląd kompletny, ale w zupełności sprawdza się jako podstawa dla dalszego wywodu prowadzonego w rozprawie. Dodatkowo omówione tutaj są także m.in. takie struktury danych jak tablice czy drzewa sufiksów.

W rozdziale trzecim znajdujemy dalszy ciąg wprowadzenia. Tym razem dotyczy on jednak przyjętego modelu obliczeń (Word RAM), upakowanej reprezentacji tekstów. Ponadto omawiane są drzewa fuzji czy drzewa falkowe, a także techniki realizacji zapytań o rangę.

Rozdział czwarty rozpoczyna część dotyczącą nowych wyników Doktoranta. Dotyczy on zapytań o najdłuższy wspólny prefiks w przypadku tekstów nad małym alfabetem. Po krótkim przeglądzie istniejących rozwiązań Autor pokazuje jak można skonstruować strukturę danych wielkości $O(n \log \sigma)$ bitów zapewniającą realizację zapytań o najdłuższy wspólny prefiks w czasie stałym. Czas konstrukcji tej struktury jest $O(n/\log_\sigma n)$. Struktura ta poparta odpowiednim twierdzeniem dotyczącym złożoności stanowi pierwszy z głównych wyników rozprawy.

W rozdziale piątym uwaga skupiona jest na zapytaniach dotyczących okresów w tekstach. W pewnym uproszczeniu, słowo jest okresowe jeśli można przedstawić go jako konkatenację pewnej liczby kopii innego, krótszego słowa (ostatnia kopia może być niekompletna). Doktorant zajmuje się tutaj dwoma problemami szczegółowymi. Wynikiem rozważań nad pierwszym z nich jest struktura danych wielkości $O(n)$, która odpowiada na zapytania o okresy w czasie $O(\log |x|)$, gdzie n jest długością tekstu a x jest wybranym podsłowem tego tekstu. Drugi problem szczegółowy dotyczy zapytań o prefikso-sufiksy, a więc wyszukiwania takich fragmentów tekstu, które w jednym ze wskazanych fragmentów występują jako prefiksy a w innym jako sufiksy. W obu przypadkach nie tylko takie struktury danych zostały zaproponowane, ale także zostały sformułowane i udowodnione stosowne twierdzenia dotyczące złożoności czasowych i pamięciowych.

Rozdział szósty dotyczy wewnętrznego wyszukiwania wzorca, a więc problemu, w którym wyszukiwane są wystąpienia jednego podsłowa w innym podsłowie. Sformułowano tutaj dwa główne twierdzenia. Pierwsze z nich dotyczy możliwości konstrukcji w czasie liniowym struktury danych realizujących wewnętrzne wyszukiwanie wzorca w czasie stałym, przy czym struktura ta zajmuje pamięć liniową względem długości tekstu. Drugie twierdzenie dotyczy nieco ogólniejszego problemu, w którym poszukiwany jest najdłuższy prefiks jednego podsłowa, który występuje w innym podsłowie.

W rozdziale siódmym Doktorant przedstawia zastosowania wewnętrznego wyszukiwania wzorca w problemie kompresji tekstów. Podane jest tu wykorzystanie opracowanej struktury

do ulepszenia algorytmu kompresji opartego na faktoryzacji metodą LZ77. Wyniki podsumowane są odpowiednim twierdzeniem.

Rozdział ósmy przedstawia nową strukturę danych jaką jest falkowe drzewo sufiksowe. Po jej zdefiniowaniu pokazane jest jak można jej użyć do realizacji zapytań kompresji podsłów metodą BWT+RLE. Poparte jest to stosownym twierdzeniem.

W rozdziale dziewiątym Doktorant skupia się na zaproponowaniu optymalnej struktury danych dla realizacji zapytania o najmniejszy sufiks i najmniejsza rotację cykliczną. Główne zastosowania tych zapytań dotyczą wyznaczania faktoryzacji Lyndona dla wybranego podsłowa oraz klasyfikacji podsłów ze względu na cykliczną równoważność. Jak w poprzednich rozdziałach, uzyskane wyniki podsumowane są za pomocą twierdzeń.

Ostani rozdział stanowi zwięzłe podsumowanie uzyskanych wyników.

Rozprawa kończy się bogatym spisem cytowanej literatury liczącym 146 pozycji.

2. Najważniejsze wyniki przedstawione w rozprawie

Doktorant zawarł w pracy wiele nowych wyników dotyczących realizacji zapytań wewnętrznych w tekstach. Rozważane problemy z pewnością nie należą do trywialnych a uzyskane rezultaty są bardzo wartościowe. Godna pochwały jest dbałość o precyzyjne formułowanie problemów i analizę złożoności opracowanych przez siebie struktur danych i algorytmów. Spośród tych licznych wyników za najbardziej wartościowe i interesujące (przynajmniej z mojego punktu widzenia) uważam:

- nową strukturę danych realizującą zapytania o najdłuższy wspólny prefiks,
- nową strukturę danych realizującą zapytania o prefikso-sufiksy,
- nową strukturę danych realizującą zapytania wewnętrznego wyszukiwania wzorca,
- nową strukturę danych realizującą rangowanie i selekcję sufiksów podsłów.

Pozostałe rezultaty także są interesujące, a mój wybór podyktowany jest zapewne tym, że wymienione wyżej wyniki są bliższe tematyce, której dotyczą moje zainteresowania naukowe.

3. Uwagi merytoryczne

Praca jest napisana bardzo starannie, z dużą dbałością o precyzję. Robi to bardzo dobre wrażenie i dobrze świadczy o dojrzałości Doktoranta. Z pewnością imponująca jest liczba twierdzeń i ich dowodów, a także swoboda, z jaką Autor rozważa kwestie złożoności czasowej i pamięciowej. Tym niemniej w każdej większej pracy nietrudno o jakieś przeoczenia czy też niekonsekwencje. Poniżej wymienię kilka z nich, które zwróciły moją uwagę:

1. W polskiej wersji autoreferatu występuje Twierdzenie 5, którego nie znalazłem w rozprawie. Prawdopodobnie w tym miejscu powinno się znaleźć Twierdzenie 1.1.5 z rozprawy, które nie jest wspomniane w autoreferacie.
2. Na stronie 8. rozprawy znajduje się Twierdzenie 1.1.6 (bez dowodu). Można się domyślać, że jego sformułowanie stanowi pewną kombinację Twierdzeń 8.4.4 oraz 8.5.1 (str. 80. rozprawy). Tym niemniej lepiej by było, aby we wstępie zawierającym bardzo ładny „przewodnik" po wynikach opisanych w rozprawie sformułowanie było takie samo jak w dalszej części.

3. Podobnie Twierdzenia 1.1.7, 1.1.10 oraz Stwierdzenie 1.1.8 także zawarte są tylko we wstępie. Z pewnością umieszczenie ich w odpowiednich miejscach rozdziałów 8 i 9 pozytywnie wpłynęłoby na spójność wywodu i łatwość jego śledzenia.

4. Uwagi redakcyjne

Od strony redakcyjnej rozprawa została napisana precyzyjnym językiem, z dużą dbałością o szczegóły. Rysunki ilustrujące proponowane struktury danych zostały przygotowane starannie. Skład dokumentu także nie budzi poważniejszych zastrzeżeń. Tym niemniej, jak w każdej większej pracy, można znaleźć pewne niedociągnięcia, np.:

- W spisie treści należy umieścić też bibliografię (str. vii).
- W sformułowaniu Twierdzeń 1.1.2 oraz 1.1.7 występuje symbol „x", którego znaczenie jest wprawdzie wyjaśnione wcześniej, ale dla czytelnika z pewnością byłoby korzystniej gdyby znaczenie tego symbolu było zawarte także w sformułowaniu samego twierdzenia.
- Indeksy „opisowe" takie jak np. „rsucc" powinny być składane fontem prostym (np. str. 5). W części przypadków tak Autor robi, ale nie zawsze.
- Można się zastanowić czy do algorytmu RLE nie dać odwołania do pozycji literaturowej (str. 19).
- Sformułowanie „non-overlapping periodic progressions" (str. 61) wydaje się być nieco mylące (chodzi o samą nazwę). Zgodnie z definicją „the last term of $\mathbf{p}$", którym jest p_{k-1} musi być mniejszy niż p'_0 (w jednym z przypadków), co oznacza, że symbol tekstu o indeksie p_{k-1} nie może należeć ani do fragmentu $\mathbf{p}$ ani też do fragmentu $\mathbf{p}'$. W potocznym rozumieniu znaczenia nakładania się fragmentów można dopuścić, aby $\mathbf{p}$ i $\mathbf{p}'$ były „bliżej siebie" i jeszcze się nie nakładały. Domyślam się, że wybrana definicja ma związek z faktoryzacją LZ77, ale dla przejrzystości wywodu wydaje się celowe zwrócenie uwagi czytelnika w tym miejscu pracy na tę kwestię.

5. Podsumowanie

Przedstawiona do oceny rozprawa zawiera wiele bardzo wartościowych wyników, co starałem się podkreślić we wcześniejszej części recenzji. Prace te zostały zaprezentowane na znaczących konferencjach naukowych takich jak SODA (2 prace), CPM (1 praca). Co więcej, Doktorant uzyskał na konferencji CPM nagrodę dla najlepszego artykułu.

Rozprawa zawiera drobne niedociągnięcia, co wskazałem w niniejszej recenzji. Tym niemniej nie umniejszają one mojej bardzo wysokiej oceny. Życzyłbym sobie, aby wszystkie recenzowane przeze mnie rozprawy doktorskie stały na tak wysokim poziomie.

Na zakończenie dodam, że choć nie uważam, aby wskaźniki bibliometryczne były kluczowe na tak wczesnym etapie kariery naukowej, to jednak wobec faktu, że są one imponujące, to je przytoczę. Baza Scopus zawiera 71 prac, których współautorem jest Doktorant. Tylko te 3 prace, których wyniki stanowią sedno recenzowanej rozprawy były według niej cytowane już 27 razy.

III. Konkluzja

Podsumowując stwierdzenia zawarte w mojej recenzji, bez wątpienia mogę stwierdzić, że zgodnie z „Ustawą o Stopniach naukowych...” recenzowana rozprawa mgra Tomasza Kociumaki nt. „Efektywne struktury danych dla zapytań wewnętrznych w tekstach (Efficient data structures for internal queries in texts)” stanowi oryginalne rozwiązanie problemu naukowego oraz wykazuje ogólną wiedzę teoretyczną kandydata w dyscyplinie informatyka oraz umiejętność prowadzenia pracy naukowej. Zawarte w recenzji nieliczne uwagi krytyczne w żadnej mierze nie wpływają na moją bardzo wysoką ocenę rozprawy. Wnoszę zatem o dopuszczenie wspomnianej rozprawy do publicznej obrony.

Ponadto, biorąc pod uwagę wysoką jakość uzyskanych wyników oraz fakt, iż rezultaty prac były już prezentowane na renomowanych konferencjach i spotkały się z bardzo dobrym odbiorem wnoszę o wyróżnienie pracy.