dr hab. Szymon Grabowski, prof. ucz.
Politechnika Łódzka
Instytut Informatyki Stosowanej
90-924 Łódź, Stefanowskiego 18/22

Łódź, dn. 7.09.2021 r.

Recenzja rozprawy doktorskiej mgr. Wiktora Zuby nt. *Efficient enumeration in words*, przygotowanej pod kierunkiem prof. dr. hab. Wojciecha Ryttera

1. Informacje wstępne

Podstawę wykonania recenzji stanowi uchwała Rady Naukowej Dyscyplin Matematyka i Informatyka UW z dnia 24 czerwca 2021 r., powołująca mnie na recenzenta rozprawy doktorskiej mgra Wiktora Zuby w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie informatyka.

Rozprawa ma postać cyklu czterech publikacji, w języku angielskim, a także zawiera 13-stronicowy autoreferat w wersji angielskiej (*extended abstract*) i polskiej.

2. Ocena zawartości rozprawy

Publikacje tworzące cykl mają charakter teoretyczny (kombinatoryka na słowach), są bez wyjątku zespołowe i zostały przedstawione na dobrych (choć nie z najwyższej półki) konferencjach. Dokładniej, przedstawiają się one następująco (litera A, B lub C po nazwie konferencji charakteryzuje jej aktualną rangę wg CORE Conference Portal):

- „Efficient Representation and Counting of Antipower Factors in Words" (LATA'19, C, 6 autorów; w doktoracie zaprezentowana w wersji pokonferencyjnej (special issue) przyjętej do pisma „Information & Computation"),
- „Counting Distinct Patterns in Internal Dictionary Matching" (CPM'20, B, 8 autorów),
- „The Number of Repetitions in 2D-Strings" (ESA'20, A, 5 autorów),
- „Efficient Enumeration of Distinct Factors Using Package Representations" (SPIRE'20, B, 6 autorów).

Taka sytuacja wymaga ścisłej charakterystyki wkładu Doktoranta i pozostałych autorów (spośród których Wojciech Rytter, Jakub Radoszewski i Tomasz Waleń są współautorami wszystkich czterech artykułów, zaś Panagiotis Charalampopoulos i Tomasz Kociumaka trzech z nich). Otrzymane przeze mnie oświadczenia są dość precyzyjne w tej materii, ale niestety, bodaj najmniejszą precyzją charakteryzuje się oświadczenie samego Doktoranta. Co bowiem dokładnie znaczy *my most visible individual contributions were (...) in creation of (...), in designing of the algorithm (...)* (w odniesieniu do, odpowiednio, pierwszego i drugiego

z wyżej wymienionych artykułów)? Czy np. algorytm znajdowania różnych faktorów kwadratowych (ang. *distinct square factors*) w publikacji z CPM'20 opracował sam Doktorant, czy też tylko miał (bliżej nieznany) udział w jego opracowaniu? Chodzi tu chyba o samodzielne otrzymanie wyników z sekcji 6 („Internal Counting of Distinct Squares") – potwierdza to raczej fakt, iż pozostali autorzy nie wspominają o niej w swoich oświadczeniach; niemniej, jednoznaczność sformułowań by na pewno pomogła.

Przedstawię teraz pokrótce te cztery publikacje. Praca „Efficient Representation and Counting of Antipower Factors in Words" prezentuje trzy wyniki dla tzw. k-antypotęg, tj. słów składających się z k parami różnych faktorów o tej samej długości. Pierwszy wynik to zliczanie i raportowanie, dla słowa o długości n, wszystkich jego podsłów będących k-antypotęgami w czasie $O(n \log k + |\text{output}|)$, drugi z przedstawionych wyników wyznacza liczbę różnych k-antypotęg w słowie w czasie $O(n\, k^4 \log k \log n)$ (ten wynik został poprawiony w ostatnim artykule w cyklu), a trzeci jest parametryzowanym indeksem: dla dowolnego $1 \le r \le n$ wybieranego przez użytkownika buduje strukturę o rozmiarze $O(n^2/r)$ (w czasie liniowym w jej rozmiarze) sprawdzającą czy dane zapytanie jest k-antypotęgą w czasie $O(r)$. Osiągnięcia te są lepsze niż zaprezentowane w pracy wcześniejszej (Badkobeh i in., IPL 2018) przynajmniej dla pewnych zależności między n, k i r (np. wynik ostatni, tj. indeks, wygrywa z analogicznym indeksem ze wspomnianej pracy, tj. osiąga interesujący kompromis między pamięcią a szybkością obsługi zapytań, gdy r jest (znacznie) mniejsze od k).

Artykuł „Counting Distinct Patterns in Internal Dictionary Matching" przedstawia również kilka wyników, tym razem dla problemu zliczenia kluczy ze słownika D w wskazanym fragmencie (przedziale) tekstu T, z użyciem pomocniczej struktury danych. Podstawowy wynik Autorów to budowa struktury w czasie i pamięci $O((n+d)\,\text{polylog}(n+d))$ i 2-aproksymacyjna obsługa zapytań w czasie stałym (analogiczny wcześniejszy wynik, częściowo tego samego zespołu Autorów, osiągał (log n)-aproksymację). Możliwe są też warianty z dokładną obsługą zapytań, ale w czasie $O(m)$, gdzie m jest parametrem struktury danych (im większe m, tym, naturalnie, mniejsza i szybciej budowana struktura).

Przedmiotem artykułu „The Number of Repetitions in 2D-Strings" są powtórzenia w tekstach dwuwymiarowych. Doktorant słusznie wskazuje na trudności takiej generalizacji względem tekstu jednowymiarowego, np. tablica $n \times n$ może zawierać aż $\Theta(n^4)$ kwartyk (kwartyką są cztery identyczne teksty ułożone w kratę 2×2); *różnych* kwartyk musi jednak być znacznie mniej. Autorzy wykazali, że liczba dwuwymiarowych „runów" w tablicy $n \times n$ to $O(n^2 \log^2 n)$, czym znacznie zbliżyli się do znanego ograniczenia dolnego ($\Omega(n^2)$), podali też takie samo ograniczenie na liczbę różnych kwartyk. Autorzy przedstawiają też algorytmiczne zastosowania nowego ograniczenia górnego na liczbę runów 2D odnoszące się do kwartyk. Niejako „po drodze" rozwiązany został interesujący (choć nietrudny) problem raportowania wymiarów maksymalnych białych prostokątów w tablicy „schodkowej" w liniowym czasie (lemat 21), co wg mnie może mieć zastosowanie np. w kompresji pewnych klas obrazów.

Praca ostatnia, „Efficient Enumeration of Distinct Factors Using Package Representations", proponuje tzw. reprezentację pakietową dla podzbiorów słów, która w szczególnych przypadkach (rozważane w pracy przypadki to potęgi oraz k-antypotęgi) pozwala na zwięzłą ich reprezentację. Idea ta pozwoliła na poprawę czasu zliczenia

wszystkich k-antypotęg w słowie do $O(nk^2)$. Praca ta w mniejszym stopniu niż poprzednie odwołuje się do problemów postawionych wcześniej w literaturze.

Wszystkie publikacje tworzące cykl cechują się dużą trudnością i technicznością otrzymanych wyników, z których największe znaczenie dla dziedziny ma chyba artykuł o tekstach 2D, poprawiający lub rozszerzający wyniki innych zespołów, raportowane m.in. pracami (Amir i in., 2018, 2020), (Deza i in., 2015), (Bannai i in., 2017). Doktorant pracuje w zasadniczo tym samym, dość licznym i mocnym naukowo zespole.

Pewien niedosyt pozostawia u mnie Autoreferat. Owszem, poprawnie definiuje on rozważane w doktoracie problemy, podając odpowiednie definicje i skrótowy stan wiedzy, a także w wystarczającym stopniu streszcza osiągnięte wyniki. Z drugiej strony, pewna hermetyczność poruszanej problematyki i wysoka techniczność i zwięzłość publikacji tworzących przedłożony cykl (ich zwięzłość może się wiązać np. z limitami stron w publikacjach konferencyjnych) zachęcają do zmiany ujęcia w autoreferacie na bardziej popularyzatorski (ułatwiający zrozumienie treści, np. poprzez podanie większej ilości przykładów), a także wskazujący na szerszy kontekst badań i ich motywację. Żeby podać konkretny przykład: ostatnia publikacja z cyklu („Efficient Enumeration of Distinct Factors...") wiąże się z reprezentacją podzbiorów podsłów (w niektórych przypadkach jest to reprezentacja oszczędna), co powinno mieć interesujące konsekwencje dla niektórych innych problemów analizy tekstów czy np. bioinformatyki poza zastosowaniami algorytmicznymi z sekcji 3 artykułu; niestety, brakuje takiej dyskusji.

Wspomnę jeszcze o pozostałych publikacjach Doktoranta, spoza cyklu. Jest ich 10, również mają charakter zespołowy (grono współautorów bardzo zbliżone do już wymienionego) i były prezentowane na takich konferencjach stringologicznych / algorytmicznych jak SPIRE (2018, 2019, 2020), CPM (2019, 2020), LATA'19, WALCOM'20, jak również w czasopismach Theoretical Computer Science i Journal of Computer and System Sciences. Podsumowując, dorobek Doktoranta jest bardzo dobry, na tym etapie Jego kariery naukowej, tak pod względem jakościowym, jak i ilościowym.

3. Uwagi szczegółowe

W pracach teoretycznych zawsze warto podać przyjęty model obliczeniowy (np. Word-RAM), a nie widzę takich wzmianek ani w artykułach wchodzących w skład cyklu, ani w autoreferacie. Podobnie, podsumowując wyniki, warto też napisać czy obowiązują one dla najgorszego przypadku, czy też niektóre z algorytmów są zrandomizowane.

W dowodzie lematu 2.4 jest wykorzystany „bucket sort", z odnośnikiem do monografii Cormena i in. (wyd. 3 z 2009 r.). Jednak bucket sort opisany w rozdziale 8.4 Cormena jest algorytmem zrandomizowanym i czas oczekiwany jego pracy będzie liniowy przy rozkładzie jednostajnym. Jeśli pamięć (która nie jest tu wspomniana) nie jest problemem, można było użyć sortowania przez zliczanie (*counting sort*), a lepiej zapewne 1/eps (dla dowolnej stałej eps > 0) przejść sortowania pozycyjnego (*radix sort*) przy redukcji pamięci pomocniczej do $O(n^{\text{eps}} + m)$.

Odwołania do monografii powinny być, w miarę możliwości, sprecyzowane podaniem odpowiedniego rozdziału czy nawet strony. Umiarkowanie pomoże czytelnikowi np. takie odwołanie (przykład z artykułu „The Number of Repetitions in 2D-Strings"): *we use a variant of the Dictionary of Basic Factors in 2D (2D-DBF in short) that is similar to the one presented in [25]*, gdzie [25] to „Jewels of stringology" Crochemore'a i Ryttera.

Stwierdzenie, iż policzenie liczby elementów LPF[$i..i+k$] mniejszych od l można zrealizować przy pomocy zapytań zakresowych (ang. *range queries*) w czasie $O((n+m)\log^{1/2} n)$ nie zostało opatrzone odnośnikiem literaturowym.

W końcówce sekcji 4.3 rozdziału 6 niejasne jest dla mnie przejście z pary zagnieżdżonych sum o zakresach: x od 1 do $\log_3 n - 1$ oraz l od $3x$ do $9x-1$ do końcowej sumy o zakresie od 3 do n. Wprawdzie mamy tu notację $O(.)$, ale pewne składniki $|F_l| \log n$ powtarzają się wiele razy (inaczej mówiąc, gdyby zastąpić $|F_l|$ dowolnymi y_l, to wzór nie byłby poprawny) – wdzięczny będę Autorowi za wyjaśnienie.

W rozdz. 1.3.3 na str. 9 Autor wspomina technikę zamiatania: *This problem can be solved with the use of a sweeping line technique ([7]), which allows us to count set-theoretic sums of many families of rectangles on an $n \times n$ grid in $O(n + r +$ output) total time (r denotes the number of all the rectangles)*. Nie mam dostępu do pracy [7] (Bentley, unpublished notes, 1977), ale pośrednio wiem, że algorytm Bentleya liczy pole pokryte n prostokątami (o bokach równoległych do osi) w czasie $O(n \log n)$. Moje pytanie do Doktoranta: czy mówimy o tym samym (lub równoważnym) problemie? Jeśli tak, to co z czynnikiem logarytmicznym? Być może jest to kwestia modelu obliczeniowego, ale proszę o wyjaśnienie.

Definicja „gapped (q, d)-square" (rozdz. 3, sekcja 1) jest trochę myląca; po co wprowadzać symbol q, gdy można, nie zmieniając tej definicji, napisać „gapped (k–2, d)-square".

Literówka (rozdz. 3, sekcja 1): a simpler the linear-time algorithm → a simpler linear-time algorithm.

Drobne uwagi językowe/edytorskie do Autoreferatu:

- co najmniej kilka błędów interpunkcyjnych (brakujące przecinki),
- drobne błędy latexowe (myślnik/półpauza to --, a nie -, cudzysłów otwierający w *na pytania "Czy dane podsłowo...* (str. 18, przedostatnia linia) to prawidłowo ``),
- pewna niekonsekwencja notacji w złożonościach wrażliwych na wielkość wyjścia (ang. *output-dependent*): zwykle Autor używa +wynik, ale na str. 18, linia 3, mamy: +rozmiar wyniku,
- mało zgrabne, pod względem szyku, zdanie z rozdz. 2.3.2: „W przypadku ostatniego pytania tylko $O(\log n)$ aproksymacja wyniku została zaprezentowana."

4. Wniosek końcowy

Biorąc pod uwagę przedstawiony jednotematyczny cykl publikacji pt. „Efficient enumeration in words", a także oświadczenia Współautorów i zamieszczony Autoreferat, stwierdzam, iż Doktorant wykazał się znaczącym (oraz wystarczająco samodzielnym) wkładem naukowym w dziedzinę kombinatoryki na słowach (czy też algorytmów tekstowych), prezentując oryginalne rozwiązania kilku powiązanych tematycznie problemów naukowych. Publikacje tworzące cykl cechują się wysokim poziomem merytorycznym, poprawiając wcześniejsze wyniki znane z literatury przedmiotu. Innymi słowy, uważam, że przedłożona rozprawa spełnia ustawowe i zwyczajowe wymogi stawiane rozprawom doktorskim. **Wnoszę zatem o dopuszczenie p. mgra Wiktora Zuby do dalszych etapów przewodu doktorskiego.**