Warszawa, 28.12.2020

prof. Wojciech Plandowski
Instytut Informatyki
Uniwersytet Warszawski

Recenzja pracy doktorskiej mgr Michała Gańczorza

Na pracę doktorską składajaą sie cztery artykuły naukowe:

1. Michał Gańczorz, Entropy bounds for grammar compression, preprint, *CoRR abs/1804.08547*;
2. Michał Gańczorz, Entropy lower bounds for dictionary compression, *Proc. CPM'19*, 2019;
3. Michał Gańczorz, Towards better compressed representations, *Proc. DCC'20*, 2020;
4. Michał Gańczorz, Using statistical encoding to achieve tree succinctness never seen before, *Proc. STACS'20*, 2020.

W pierwszej z nich autor rozważa reprezentacje bitowe gramatyk bezkontekstowych generujących pojedyncze słowo. Wyróżnia cztery takie reprezentacje: całkowicie naiwną, naiwną, entropijną i przyrostową.

Autor rozważa dwa algorytmy tworzące gramatykę: algorytm Re-Pair i algortm Greedy.

Autor dowodzi, że dla gramatyk budowanych metodą Re-Pair słowa bitowe reprezentujące gramatykę mają rozmiar co najwyżej $1.5|S|H_k(S) + o(|S| \log \sigma)$ gdzie $H_k(S)$ jest entropią empiryczną rzędu k oraz $k = o(\log_\sigma |S|)$ o ile gramatyka jest kodowana metodami entropijną lub przyrostową. Jeśli jednak zatrzymamy algorytm Re-Pair w odpowiednim momencie gdy rozmiar słowa do kompresji wynosi $|S|^c$, dla pewnej stałej $c < 1$, to rozmiar ten nie przekracza $|S|H_k(S) + o(|S| \log \sigma)$.

Dla gramatyk budownych algorytmem Greedy wyniki są podobne. Mianowicie rozmiar słowa jest ograniczony przez $2|S|H_k(S) + o(|S| \log \sigma)$ dla kodowania

całkowicie naiwnego i $1.5|S|H_k(S) + o(|S| \log \sigma)$ dla kodowania entropijnego. Gdy algorytm jest zatrzymany po $|S|^c$ iteracjach, dla pewnego $c < 1$, to rozmiar ten jest ograniczony przez $|S|H_k(S) + o(|S| \log \sigma)$.

Autor zauważył ciekawą rzecz: zatrzymując algorytmy Re-Pair i Greedy w odpowiednim momencie można uzyskać lepsze ograniczenie na rozmiar generowanego słowa bitowego niż gdy algorytmy działają do końca. Poza tym pesymistyczny czas działania algorytmu Greedy to $O(|S|^2)$. Jeśli go zatrzymamy po $|S|^c$ iteracjach to czas ten maleje do $O(|S|^{1+c})$. Mamy więc szybszy algorytm dający teoretycznie lepszą kompresję.

Dowody tych wszystkich rezultatów wymagają udowodnienia szeregu lematów i zaawansowanych rachunków. Cała praca liczy 38 stron. Gdyby tylko ona wchodziła w skład pracy doktorskiej to uznałbym pracę za bardzo dobrą.

Druga praca dowodzi dolnych granic kompresji dla dużej klasy metod słownikowych. Klasa ta nazwana algorytmami naturalnymi zawiera większość słownikowych metod kompresji. Są w niej LZ77, LZ78, i metody gramatykowe takie jak Re-Pair, Greedy, Sequitur i Sequential. Praca ta została przyjęta na konferencję CPM, która jest średniej klasy konferencją poświęconą tekstom.

Niech A będzie algorytmem naturalnym. W pracy dowodzi się, że dla pewnej rodziny tekstów S_n dolne ograniczenie na na rozmiar skompresowanego tekstu S_n algorytmem A wynosi $|S_n|H_k(S_n) + \Omega(|S_n|k \log \sigma / \log_\sigma |S_n|)$ bitów, dla $k \leq \log_\sigma |S_n| - \frac{1}{2}$.

Niech α będzie liczbą wymierną z przedziału $(0, 1)$, zaś A algorytmem naturalnym. Drugim głównym rezultatem pracy jest, że dla dostatecznie dużego σ istnieje rodzina tekstów S_n nad alfabetem σ-literowym taka, że jeśli algorytm A skompresuje tekst do rozmiaru $\beta|S_n|H_k(S_n) + o(|S_n| \log \sigma)$ bitów, dla każdego n, i dla $k = \alpha \log_\sigma |S_n|$, to $\beta \geq \frac{1}{1-\alpha}$.

Rodziny tekstów S_n są w obu twierdzeniach zbudowane w oparciu o ciągi de Brujin.

Niech m będzie liczbą naturalną wiekszą od 1. Trzecia praca traktuje o algorytmach kompresji opartych na faktoryzacjach słowa wejściowego S na faktory o długościach co najwyżej m. Takie faktoryzacje są nazywane m-faktoryzacjami. Pośród m-faktoryzacji słowa S należy znaleźć taką, która minimalizuje $|Y_S|H_0(Y_S)$ gdzie Y_S jest m-faktoryzacją słowa S, zaś H_0 empiryczną entropią rzędu 0 dla ciągu Y_S. Praca została przyjęta na konferencję DCC, która jest najlepszą konferencją specjalizującą się w kompresji danych.

Autorowi nie udaje się znaleźć dokładnego rozwiązania głównego problemu. Podaje za to heurystykę wynikające z dwu twierdzeń ograniczających z góry wartośc $|Y_S|H_0(Y_S)$.

Podobnie podaje heurystykę dla wyznaczenia m-faktoryzacji minimalizującej $|Y_S|H_1(S)$ na podstawie dwu twierdzeń ograniczających z góry wartość $|Y_S|H_1(S)$.

Obie heurystyki obliczają kwazi-optymalną faktoryzację metodą programowania dynamicznego. Okazuje się że heurystyki te są, dla tekstów słabookresowych, bardziej efektywne od gzip.

Rezultaty te wymagają nietrywialnych dowodów dwu lematów.

Ostatnia praca jest poświęcona kompresji drzew, których wierzchołki są etykietowane literami ze skończonego alfabetu. Została ona przyjęta na kon-

ferencję STACS, która jest, obok konferencji ICALP, najbardziej prestiżową konferencją europejską specjalizującą się w informatyce teoretycznej.

Dla drzewa $\mathcal{T}$ definiuje się entropię rzędu k oznaczaną $H_k(\mathcal{T})$ oraz entropię $H(\mathcal{T})$. Dotychczasowe wyniki dotyczące kompresji drzew używały tylko tych dwu parametrów do prezentacji głównych wyników prac. Autor proponuje dwa dalsze parametry $H_k(\mathcal{T}|L)$ i $H_k(L|\mathcal{T})$. Następnie, opisuje efektywność algorytmów w oparciu o te cztery parametry. W jego pracy drzewo jest kodowane za pomocą minimalnej liczby bitów z dwu wartości:

$$|\mathcal{T}|H_k(\mathcal{T}|L) + |\mathcal{T}|H_k(L) + O(|\mathcal{T}|k \log \sigma / \log_\sigma |\mathcal{T}|) + |\mathcal{T}| \log \log_\sigma |\mathcal{T}| / \log_\sigma |\mathcal{T}|$$

oraz

$$|\mathcal{T}|H_k(L|\mathcal{T}) + |\mathcal{T}|H(\mathcal{T}) + O(|\mathcal{T}|(k+1) \log \sigma / \log_\sigma |\mathcal{T}|) + |\mathcal{T}| \log \log_\sigma |\mathcal{T}| / \log_\sigma |\mathcal{T}|.$$

Taka struktura danych jest mniejsza od wcześniej znanych i pozwala na wykonywanie podstawowych operacji na drzewach w czasie $O(1)$ co poprawia wcześniejsze rezultaty. Co więcej, za cenę zwiększenia kosztu operacji do $O(\log |\mathcal{T}| / \log \log |\mathcal{T}|)$ można zredukować rozmiar drzewa o pierwszy składnik w $O()$ o ile $k = \alpha \log_\sigma |\mathcal{T}|$, dla pewnego $0 < \alpha < 1$.

Praca ta zawiera mnóstwo nietrywialnych rezultatów i jest bardzo zaawansowana. Gdyby tylko ona wchodziła w skład doktoratu, to uznałbym go za bardzo dobry.

Podsumowując, cztery prace wchodzące w skład rozprawy doktorskiej z wielkim naddatkiem wypełniają wymagania dotyczące rozpraw doktorskich. Wnioskuję o **wyróżnienie** rozprawy.

Wojciech Plandowski