Efektywne algorytmy związane z kombinatoryczną strukturą słów

Autoreferat rozprawy doktorskiej

Marcin Piątkowski

Uniwersytet Mikołaja Kopernika
Wydział Matematyki i Informatyki

Badanie struktury powtórzeń jest jednym z podstawowych problemów kombinatoryki i algorytmiki słów, patrz [9], [11], [12]. Początki badań w tej dziedzinie datują się na około sto lat temu, patrz [18] (1906) oraz [19] (1912). Najważniejszymi typami rozważanych powtórzeń są **kwadraty** (powtórzenia postaci zz) związane z kompresowalnością słów, oraz **maksymalne powtórzenia** mające wiele zastoswań algorytmicznych, patrz [10].

Maksymalnym powtórzeniem w słowie s nazywamy podsłowo w słowa s postaci $w = u^k v$, gdzie $k \geqslant 2$, u jest minimalne o tej własności ($u \neq y^l$), zaś v jest właściwym prefiksem u, które nie może być rozszerzone w żadną stronę bez zmiany okresu (u).

Na przykład w słowie *abaababa* mamy 3 maksymalne powtórzenia: *abaaba*, *ababa*, *aa* oraz 4 parami różne kwadraty: *abab*, *baba*, *aa*, *abaaba*.

Oznaczmy przez $\boldsymbol{sq(w)}$ liczbę parami różnych kwadratów, zaś przez $\boldsymbol{\rho(w)}$ liczbę maksymalnych powtórzeń w słowie w; dodatkowo niech $\boldsymbol{sq(n)}$ oraz $\boldsymbol{\rho(n)}$ oznaczają odpowiednio maksymalną liczbę parami różnych kwadratów oraz maksymalnych powtórzeń w słowach długości n. Mimo wielu badań w tej dziedzinie dokładne asymptotyczne oszacowanie dla $sq(n)$ oraz $\rho(n)$ nadal nie jest znane, zaś algorytmy wyznaczania $sq(w)$ oraz $\rho(w)$ są skomplikowane.

W rozprawie zbadana została struktura wystąpień powtórzeń dla klasy standardowych słów Sturma $\mathcal{S}$ – jednej z intensywniej badanych klas w kombina-

toryce słów, zobacz na przykład [1], [4], [5], [6], [14], [16], [17], [20]. Podobne rezultaty uzyskuje się dla słów Christoffela, będących pewnymi cyklicznymi przesunięciami słów standardowych, oraz mających wiele zastosowań geometrycznych, patrz [6], [7], [8].

Slowa standardowe posiadają zwarte reprezentacje w postaci ciągów dodatnich liczb całkowitych. Dla ciągu $\gamma = (\gamma_0, \ldots, \gamma_n)$ słowo standardowe $\mathrm{Sw}(\gamma) = x_{n+1}$ określamy za pomocą rekurencji:

$$\begin{aligned} x_{-1} &= b, & x_0 &= a, \\ x_1 &= x_0^{\gamma_0} x_{-1}, & x_2 &= x_1^{\gamma_1} x_0, \\ &\vdots & &\vdots \\ x_n &= x_{n-1}^{\gamma_{n-1}} x_{n-2}, & x_{n+1} &= x_n^{\gamma_n} x_{n-1}. \end{aligned}$$

Ciąg γ możemy traktować jako skompresowaną reprezentację słowa standardowego. Długość słowa $N = |\mathrm{Sw}(\gamma)|$ jest wykładniczo zależna od długości ciągu $(n+1) = |\gamma|$. W szczególności dla ustalonego n najmniesza długość słowa jest równa długosci słowa Fibonacciego $F_n = \mathrm{Sw}(1, 1, ..1)$, a więc $|\mathrm{Sw}(\gamma)| \geqslant c \cdot \phi^n$, gdzie $\phi = \frac{1+\sqrt{5}}{2}$.

Wyniki prezentowane w rozprawie opierają się na opisanej powyżej reprezentacji słów standardowych. Mamy więc do czynienia z badaniem struktury powtórzeń w skompresowanej wersji słów, wykładniczych względem rozmiaru wejścia, co dodatkowo komplikuje konstrukcję algorytmów działających w czasie liniowym ze względu na rozmiar wejścia.

W **rozdziale 1** zdefiniowana została klasa $\mathcal{S}$ standardowych słów Sturma – nieokresowych słów nad alfabetem binarnym o minimalnej złożoności kombinatorycznej. Na wstępie przedstawiona została definicja rekurencyjna, prowadząca do efektywnej skompresowanej reprezentacji w postaci ciągów liczb całkowitych o długości logarytmicznie zależnej od długości słowa. Następnie zaprezentowana została geometryczna definicja słów Christoffela – pewnych cyklicznych przesunięć słów standardowych – jako dyskretyzacji prostych na płaszczyźnie dyskretnej oraz jako etykietowania grafu Cayley'a pewnej grupy skończonej związanej z liczbą liter w generowanym słowie. Zaprezentowana również została konstrukcja słów standardowych za pomocą rodziny morfizmów oraz ich arytmetyczny opis za pomocą ułamków łańcuchowych.

Rozdział 2 poświęcony został dokładnemu zbadaniu struktury grafów podsłów słów standardowych. Struktura ta pozwala uzyskać nowe prostsze dowody wielu znanych faktów dotyczących słów standardowych oparte na strukturze grafów oraz prowadzi do prostych algorytmów wyznaczających takie

własności słów standardowych jak: liczba podsłów, punkt rozkładu krytycznego, leksykograficznie maksymalne sufiksy oraz zbiory wystąpień podsłów określonej długości wczasie liniowo zależnym od długości skompresowanej reprezentacji słowa (logarytmicznie zależnym od długości samego słowa).

W **rozdziale 3** na podstawie struktury grafów podsłów rozważanej w rozdziale 2 uzyskano prostą charakteryzację okresów maksymalnych powtórzeń w słowach standardowych. Prowadzi ona do zwartego wzoru na liczbę $\rho(w)$ maksymalnych powtórzeń w słowach $w \in \mathcal{S}$ oraz algorytmu liczącego $\rho(w)$ w czasie liniowo zależnym od rozmiaru skompresowanej reprezentacji w (zatem logarytmicznie zależnym od długości całego w). Dowiedziono również, że

$$\frac{\rho(w)}{|w|} \leqslant \frac{4}{5} \quad \text{dla dowolnego} \quad w \in \mathcal{S},$$

oraz skonstruowano nieskończony ciąg słów $w_k \in \mathcal{S}$ o ściśle rosnącej długości, dla którego

$$\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} \;=\; \frac{4}{5}.$$

Dokładne zbadanie struktury powtórzeń dla dużej klasy skomplikowanych słów $\mathcal{S}$ umożliwia lepsze zrozumienie tego problemu w szerszych klasach słów.

W **rozdziale 4** badanie powtórzeń zostało rozszerzone na kwadraty. Wykorzystane zostały wyniki pracy [13], gdzie zaprezentowane zostały skomplikowane wzory na liczbę kwadratów w słowach standardowych. W rozprawie uproszczono nieznacznie wzory, dowiedziono, że

$$sq(w) \leqslant \frac{9}{10}|w| \quad \text{dla dowolnego} \quad w \in \mathcal{S}$$

oraz skonstruowano nieskończony ciąg słów $w_k \in \mathcal{S}$ o ściśle rosnącej długości, dla którego

$$\lim_{k\to\infty} \frac{sq(w_k)}{|w_k|} \;=\; \frac{9}{10}.$$

Klasa słów standardowych $\mathcal{S}$ jest największą klasą słów, dla której znane są zwarty wzór oraz asymptotyczne oszacowanie na liczbę kwadratów.

W rozdziale tym dokonano również asymptotycznej analizy liczby maksymalnych powtórzeń oraz kwadratów w słowach Christoffela.

W **rozdziale 5** struktura skompresowanych grafów podsłów słów standardowych opisana w rozdziale 2 została wykorzystana do opisania dualnego systemu liczbowego Ostrowskiego. System ten może zostać zdefiniowany bez

odwoływania się do słów standardowych, jednak podejście grafowe prowadzi do prostej interpretacji reprezentacji dodatnich liczb całkowitych w tym systemie jako długości ścieżek w grafie podsłów odpowiedniego słowa standardowego. Wprowadzone również zostało nowe pojęcie związane ze słowami standardowymi – automat Ostrowskiego.

Analiza maksymalnych powtórzeń oraz kwadratów w bardzo długich słowach standardowych była w sposób istotny wspomagana komputerowo. Kilka przydatnych apletów związanych z problemami rozważanymi w rozprawie jest dostępnych na stronie:

`http://www.mat.umk.pl/~martinp/stringology/applets/`

Podsumowując, głównymi wynikami rozprawy są:

1. algorytm znajdowania $\rho(w)$ dla $w \in \mathcal{S}$ w czasie liniowo zależnym od jego (zazwyczaj logarytmicznej) skompresowanej reprezentacji γ (patrz [2]);

2. asymptotyczna granica $\rho(w) \leqslant 0.8|w|$ dla $w \in \mathcal{S}$ (patrz [2]);

3. algorytmiczna konstrukcja nieskończonego ciągu słów standardowych $\{w_k\}$ asymptotycznie osiągającego granicę $\rho(w_k) = 0.8|w_k| - o(|w_k|)$ (patrz [2]);

4. asymptotyczna granica $sq(|w|) \leqslant 0.9|w|$ dla $w \in \mathcal{S}$ ($\mathcal{S}$ stanowi największą klasę słów, dla której znane jest dokładna asymptotyczna granica, dla ogólniejszych klas słów najlepszym oszacowaniem jest $n \leqslant sq(n) \leqslant 2n$, a uzyskanie dokładniejszego ograniczenia jest bardzo trudne) (patrz [15]);

5. algorytmiczna konstrukcja nieskończonego ciągu słów standardowych $\{w_k\}$ asymptotycznie osiągającego granicę $sq(w_k) = 0.8|w_k| - o(|w_k|)$ (patrz [15]);

6. zbadanie struktury skompresowanych grafów podsłów dla słów standardowych – rozmiar tych grafów jest liniowo zależny od γ, przy ich pomocy uzyskano kilka efektywnych algorytmów związanych z kombinatorycznymi własnościami słów standardowych (patrz [3]);

7. związek między strukturą grafów podsłów słów standardowych a pewnymi systemami liczbowymi oraz pewną szczególną klasą automatów skończonych (patrz [3]).

Wyniki te zostały opublikowane w pracach [2], [3] oraz [15].

Literatura

[1] J. Allouche and J. Shallit. *Automatic Sequences. Theory, Applications, Generalizations.* Cambridge University Press, 2003.

[2] P. Baturo, **M. Piątkowski**, and W. Rytter. The number of runs in Sturmian words. In *Proceedings of the 13th international conference on Implementation and Applications of Automata*, volume 5148 of *Lecture Notes in Computer Science*, pages 252–261. Springer, 2008.

[3] P. Baturo, **M. Piątkowski**, and W. Rytter. Usefulness of directed acyclic subword graphs in problems related to standard Sturmian words. *International Journal of Foundations of Computer Science*, 20(6):1005–1023, 2009.

[4] P. Baturo and W. Rytter. Compressed string-matching in standard Sturmian words. *Theoretical Computer Science*, 410(30–32):2804–2810, 2009.

[5] J. Berstel. Sturmian and Episturmian words: a survey of some recent results. In *Proceedings of the 2nd international conference on Algebraic informatics*, volume 4728 of *Lecture Notes in Computer Science*, pages 23–47. Springer, 2007.

[6] J. Berstel, A. Lauve, C. Reutenauer, and F. Saliola. *Combinatorics on Words: Christoffel Words and Repetitions in Words.* CRM monograph series. Providence, R.I: American Mathematical Society, 2009.

[7] E. B. Christoffel. Observatio arithmetica. *Mathematische Annalen*, 6:145–152, 1875.

[8] E. B. Christoffel. Lehrsätze über arithmetische eigenshaften du irrationnalzahlen. *Annali di Mathematica Pura ed Applicata, Series II*, 15:253–276, 1888.

[9] M. Crochemore, L. Ilie, and W. Rytter. Repetitions in strings: Algorithms and combinatorics. *Theoretical Computer Science*, 410(50):5227–5235, 2009.

[10] M. Crochemore, C. S. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, and T. Walen. Extracting powers and periods in a string from its runs structure. In *Proceedings of 17th International Symposium on String Processing and Information Retrieval*, volume 6393 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2010.

[11] M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. *Algorithmica*, 13(5):405–425, 1995.

[12] M. Crochemore and W. Rytter. *Jewels of Stringology: text algorithms*. World Scientific, 2003.

[13] D. Damanik and D. Lenz. Powers in Sturmian sequences. *European Journal of Combinatorics*, 24(4):377–390, 2003.

[14] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of mathematics and its application*. Cambridge University Press, 2002.

[15] **M. Piątkowski** and W. Rytter. Asymptotic behaviour of the maximal number of squares in standard Sturmian words. In *Proceedings of the 14-th Prague Stringology Conference*, pages 237–248. Czech Technical University, 2009. accepted to International Journal of Foundations of Computer Science.

[16] M. Sciortino and L. Zamboni. Suffix automata and standard Sturmian words. In *Proceedings of the 11th International Conference on Developments in Language Theory*, volume 4588 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2007.

[17] J. Shallit. Characteristic words as fixed points of homomorphisms. Technical Report CS-91-72, University of Waterloo, Department of Computer Science, 1991.

[18] A. Thue. Über unendliche zeichenreihen. *Norske Vid. Selsk. Skr. I Math-Nat.*, Christiana 7:1–22, 1906.

[19] A. Thue. Über die gegenseitige lage gleicher teile gewisser zeichenreihen. *Norske Vid. Selsk. Skr. I Math-Nat.*, Christiana 10:1–67, 1912.

[20] H. Uscka-Wehlou. *Digital lines, Sturmian words, and continued fractions*. PhD thesis, Department of Mathematics, Upspsala University, 2009.