Wyszukiwanie palindromów

Autor artykułu: mgr Jerzy Wałaszek

Wprowadźmy symbol s^R, który oznacza łańcuch znakowy o odwróconej kolejności znaków w stosunku do łańcucha s. Przykładowo, jeśli $s = \text{ABCD}$, to $s^R = \text{DCBA}$. Łańcuch znakowy s nazwiemy *palindromem*, jeśli czyta się tak samo w obu kierunkach, tzn. $s = s^R$. Zauważmy, że łancuch s jest palindromem, jeśli

- $s = ww^R$, gdzie w jest łańcuchem; wtedy s nazwiemy palindromem *parzystym*
- $s = wXw^R$, gdzie w jest łańcuchem, a X dowolnym symbolem alfabetu; wtedy s jest palindromem *nieparzystym.*

Przykładowo, $\text{ABCDDCBA} = \text{ABCD} + \text{DCBA}$ jest palindromem parzystym, a $\text{ABCDADCBA} = \text{ABCD} + \text{A} + \text{DCBA}$ jest palindromem nieparzystym.

Mówimy, że palindrom s w łańcuchu t jest maksymalny, jeśli fragment t zawierający jedną literkę przed i jedną literkę po s nie jest już palindromem. Przykładowo, palindrom ABA nie jest malsymalny w łańcuchu XCABACYZ, bo jest zawarty w dłuższym palindromie CABAC; z kolei palindrom CABAC już jest maksymalny.

Palindromy pojawiają się w genetyce (łańcuchy DNA, RNA), w tekstach, muzyce, matematyce, geometrii, fizyce itd. Stąd duże zainteresowanie informatyków w efektywnych algorytmach ich znajdowania. W badaniach genetycznych często szuka się tzw. przybliżonych palindromów (ang. aproximate palindromes), tzn. palindromów, w których do k-znaków może być błędnych, czyli nie pasujących do dokładnego palindromu (ang. exact palindrome). Takie palindromy występują w łańcuchach DNA, w których wystąpiły różnego rodzaju błędy genetyczne. Problemem palindromów przybliżonych nie zajmujemy się w tym opracowaniu.

Celem zadania jest napisanie procedury, która znajdzie wszsystkie maksymalne palindromy w zadanym łańcuchu t składające się z przynajmniej dwóch znaków. Dodatkowo, rozwiązanie musi działać w czasie liniowym względem długości t.

Algorytm Manachera

Rozwiązanie problemu wyszukiwania wszystkich palindromów w łańcuchu znakowym s opiera się na własnościach palindromów. Przedstawiony tutaj algorytm został opracowany w 1975 przez Glenna Manachera z Computer Center and Department of Information Engineering, University of Illinois, Chicago, IL. Do opisu algorytmu Manachera wprowadzimy kilka nowych pojęć.

Niech p_P będzie palindromem parzystym o postaci $p_P = ww^R$, gdzie w jest niepustym podłańcuchem. Niech p_N będzie palindromem nieparzystym o postaci $p_N = wXw^R$. *Promieniem* r_p palindromu p będziemy nazywali długość podsłowa w, czyli $r_p = |w|$. Palindrom parzysty p_P ma zawsze długość $|p_P| = 2r_p$; palindrom nieparzysty p_N ma zawsze długość $|p_N| = 2r_p + 1$.

Środkiem palindromu p jest pozycja $i_s = r_p$ – jest to pozycja pierwszego znaku za słowem w (można również definiować środek palindromu jako pozycję ostatniego znaku podsłowa w, lecz sądzę, iż nasz sposób jest lepszy, gdyż nie wymaga wprowadzania żadnych zmian dla palindromów nieparzystych). Dla palindromu parzystego środek wypadnie na pierwszym znaku w^R, natomiast dla palindromu nieparzystego środek wypadnie na znaku X: $p_P[r_p] = w^R[0]$, $p_N[r_p] = X$.

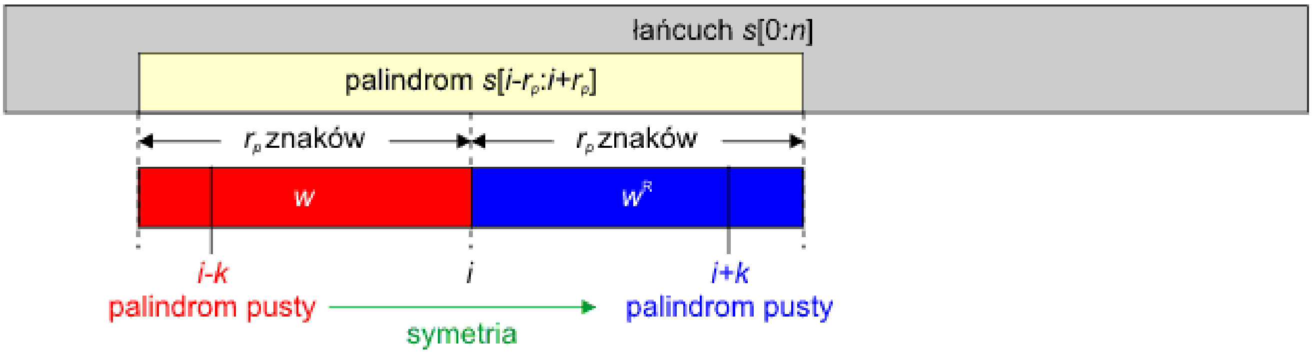
Algorytm Manachera nie wyznacza wszystkich palindromów, jak robiłby to algorytm naiwny, lecz maksymalne palindromy, których środki występują na kolejnych pozycjach znakowych przeszukiwanego łańcucha s. Dzięki takiemu podejściu redukujemy złożoność obliczeniową fazy przeszukiwania łańcucha s.

Dla danego łańcucha s algorytm Manachera tworzy tablicę dwuwymiarową R, gdzie

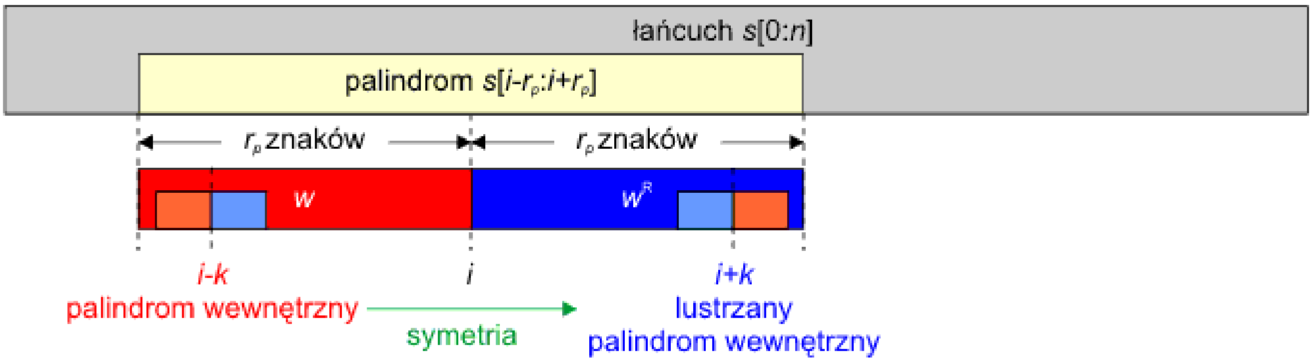
- $R[0, ...]$ – promienie palindromów parzystych
- $R[1, ...]$ – promienie palindromów nieparzystych

Indeksy tych tablic określają kolejne pozycje znakowe w łańcuchu s, natomiast elementy tablic zawierają maksymalne promienie palindromów o środkach na danej pozycji znakowej.

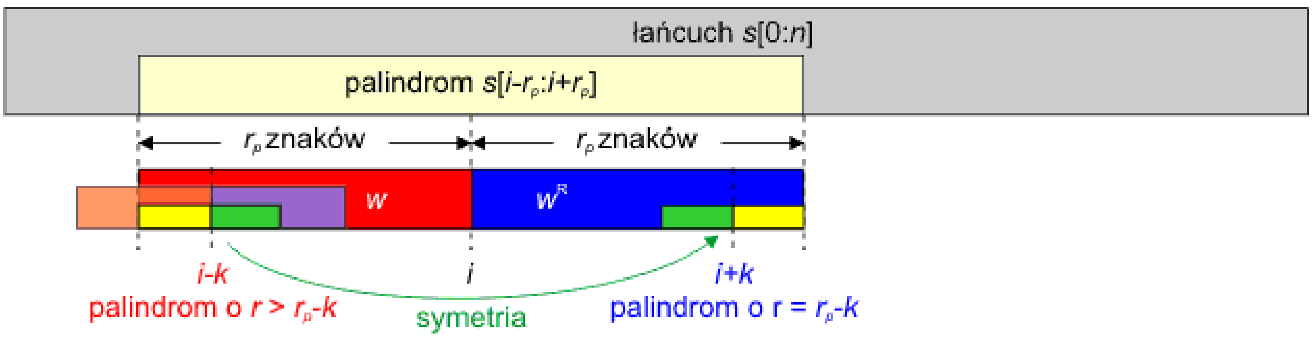
Używając w odpowiedni sposób tablicy R oraz własności symetrii palindromu algorytm Manachera wykorzystuje sprytnie informację o wcześniej wyznaczonych promieniach palindromów maksymalnych do wyszukiwania następnych palindromów. Otóż po wyznaczeniu promienia r_p palindromu na pozycji i-tej w łańcuchu s, sprawdzane są promienie palindromów na kolejnych pozycjach poprzedzających pozycję i-tą w obszarze podsłowa w. Tutaj algorytm wymaga dwóch wersji – osobnej dla palindromów parzystych i osobnej dla nieparzystych. Zasada jest identyczna dla obu wersji. Rozważmy zatem możliwe przypadki (dla palindromu parzystego), patrz rysunki 1, 2, 3 i 4.



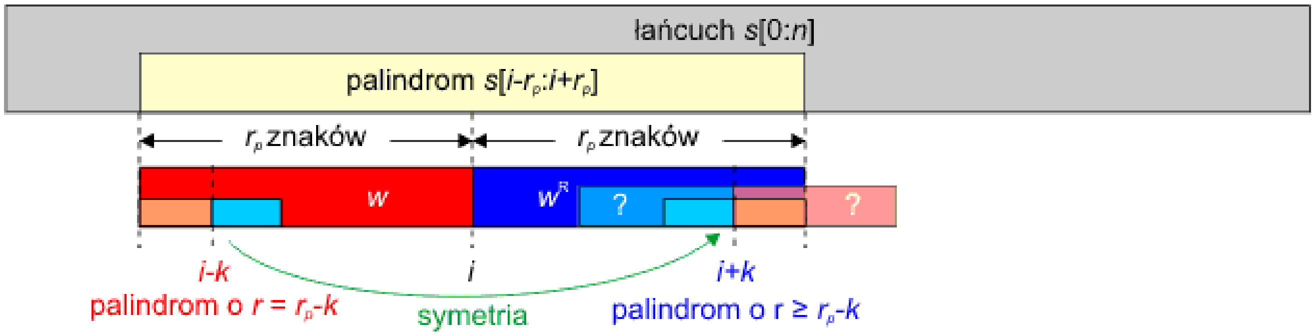
Rysunek 1: Na pozycji $i-k$, $k = 1, 2, ..., r_p$, promień palindromu wynosi 0 – czyli nie istnieje palindrom o środku na pozycji $i - k$. Skoro tak, to przez symetrię wnioskujemy, iż na pozycji lustrzanej $i + k$ również nie będzie żadnego palindromu. Pozycja $i + k$ możne zostać pominięta przy dalszym wyszukiwania palindromów.



Rysunek 2: Na pozycji $i - k$ jest palindrom o promieniu $r < r_p - k$. Taki palindrom w całości zawiera się wewnątrz rozważanego palindromu i co więcej, nie styka się z jego brzegiem. Poprzez symetrię wnioskujemy, iż na pozycji $i + k$ również musi występować taki sam palindrom, którego już dalej nie da się rozszerzyć. Pozycji $i + k$ nie musimy już dalej sprawdzać.



Rysunek 3: Na pozycji $i - k$ jest palindrom o promieniu $r > r_p - k$. Taki palindrom wykracza z lewej strony poza obszar rozważanego palindromu. Na pozycji $i+k$ znajduje się palindrom o promieniu $r = r_p - k$ i palindromu tego nie da się już rozszerzyć. Wyjaśnienie tego faktu jest bardzo proste – gdyby palindrom na pozycji $i + k$ posiadał większy promień niż wyliczone r, to również z uwagi na symetrię przeglądany palindrom posiadałby promień większy od r_p, a przecież jest to palindrom maksymalny. Pozycję $i + k$ również możemy pominąć.



Rysunek 4: Pozostał ostatni przypadek – na pozycji $i - k$ występuje palindrom o promieniu $r = r_p - k$. Taki sam palindrom musi być na pozycji $i + k$, jednakże w tym przypadku palindrom ten może być rozszerzalny. Pozycję $i + k$ musimy zatem sprawdzić na obecność palindromu o promieniu większym od r.

Z powyższych rozważań wynika liniowy algorytm, który wyznacza promienie maksymalnych palindromów o środkach w kolejnych pozycjach w zadanym tekście.

Punktacja

- (**1.5p**) – znajdowanie palindromów tylko jednej parzystości (parzystych lub nieparzystych).
- (**1p**) – znajdowanie wszystkich palindromów.

Uwagi

- Palindromy powinny być zwracane jako lista par (indeks pierwszego znaku, długość palindromu) – tzn. para (i, ℓ) oznacza, że pod indeksem i znajduje się pierwszy znak palindromu składającego się z d znaków.
- W liście wynikowej należy uwzględnić tylko palindromy o długości przyajmniej 2.
- Kolejność wyników nie ma znaczenia.
- Można założyć, że w tekście wejściowym nie występują znaki # i $ - można je wykorzystać w roli wartowników.