Algorytmy Wyszukiwania Wzorca

Robert Nazar

Instytut Matematyki

Katowice, 17 września 2015

Zawartość Pracy

- Problem wyszukiwania wzorca
- Pojęcie złożoności obliczeniowej
- Algorytm naiwny
- Algorytm Karpa-Rabina
- Wyszukiwania za pomocą automatów skończonych
- Algorytm Knutta-Morrisa-Pratta
- Krótki opis innych algorytmów wyszukiwania wzorca
- Aplikacja komputerowa

Wyszukiwanie wzorca

- Wzorzec długości m oznaczany jest $W[1...m]$
- Tekst o długości n, w którym szukany jest wzorzec oznaczany jest $T[1...n]$ lub $T[s+1...s+n]$ dla $s > 1$
- Celem jest znalezienie pozycji $s+1$, dla której $T[s+1, ...s+m] = W[1...m]$

Definicja

Złożoność obliczeniowa - zależność pomiędzy liczbą operacji elementarnych wykonywanych w trakcie przebiegu algorytmu a rozmiarem danych wejściowych.

Rzędy wielkości funkcji

1. Notacja "duże O"
   Mówimy, że f jest co najwyżej rzędu g, gdy istnieją takie stałe $n_0 > 0$ oraz $c > 0$, że

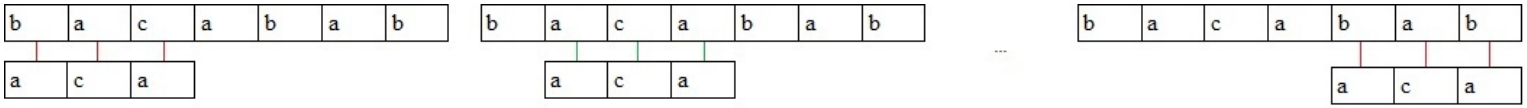
$$\forall_{n \geqslant n_0} \; f(n) \leqslant c \cdot g(n)$$

2. Notacja Θ
   Mówimy, że f jest dokładnie rzędu g, gdy istnieją takie stałe $n_0 > 0$ oraz $c_1 > 0, c_2 > 0$, że

$$\forall_{n \geqslant n_0} \; c_1 \cdot g(n) \leqslant f(n) \leqslant c_2 \cdot g(n)$$

Algorytm naiwny: Działanie

Algorytm porównuje wzorzec ze wszystkim fragmentami tekstu o długości zgodnej z długością wzorca.

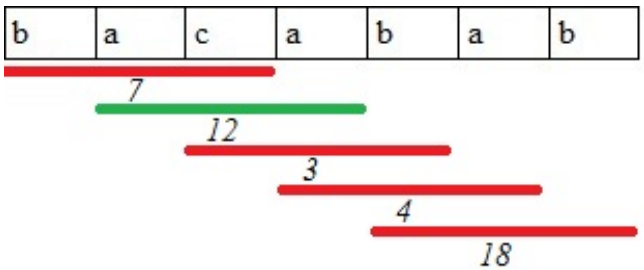


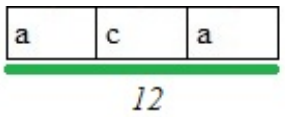
Złożoność obliczeniowa

- Algorytm pracuje w czasie $O((n - m + 1)m)$

Algorytm Karpa-Rabina: Działanie

Algorytm zamienia wzorzec i ciągi znaków tekstu w liczby nazywane haszami i następnie porównuje ich wartości. W przypadku identycznych haszy następuje porównywanie znaków podejrzanego fragmentu ze znakami wzorca.





Złożoność obliczeniowa

- Algorytm wykonuje przetwarzanie wstępne w czasie $\Theta(m)$, a operacje wyszukiwania w czasie $O((n-m+1)m)$

Definicja

Automatem skończonym M nazywamy uporządkowana piątkę $(Q, q_0, A, \Sigma, \delta)$, gdzie:

- Q jest zbiorem wszystkich stanów automatu,
- $q_0 \in Q$ jest stanem początkowym,
- $A \subseteq Q$ jest pewnym wyróżnionym zbiorem stanów akceptujących,
- Σ jest skończonym alfabetem wejściowym (zbiorem wszystkich znaków),
- δ jest funkcją $Q \times \Sigma \rightarrow Q$ zwaną funkcją przejść automatu M

Wyszukiwanie za pomoca automatów skończonych: Działanie

- Przetwarzanie wstępne polega na wyliczeniu **funkcji sufiksowej** danej wzorem $\sigma(x) = \max\{k : W_k \sqsupset x\}$, następnie korzystając z wyliczonej funkcji sufiksowej przeprowadzana jest operacja wyszukiwania.
- Funkcja przejść będzie zależna od znaków wzorca stąd $\delta(q, a) = \sigma(W_q a)$.

Złożoność obliczeniowa

- Algorytm wykonuje przetwarzanie wstępne w czasie $O(m^3|\Sigma|)$, a operacje wyszukiwania w czasie $\Theta(n)$

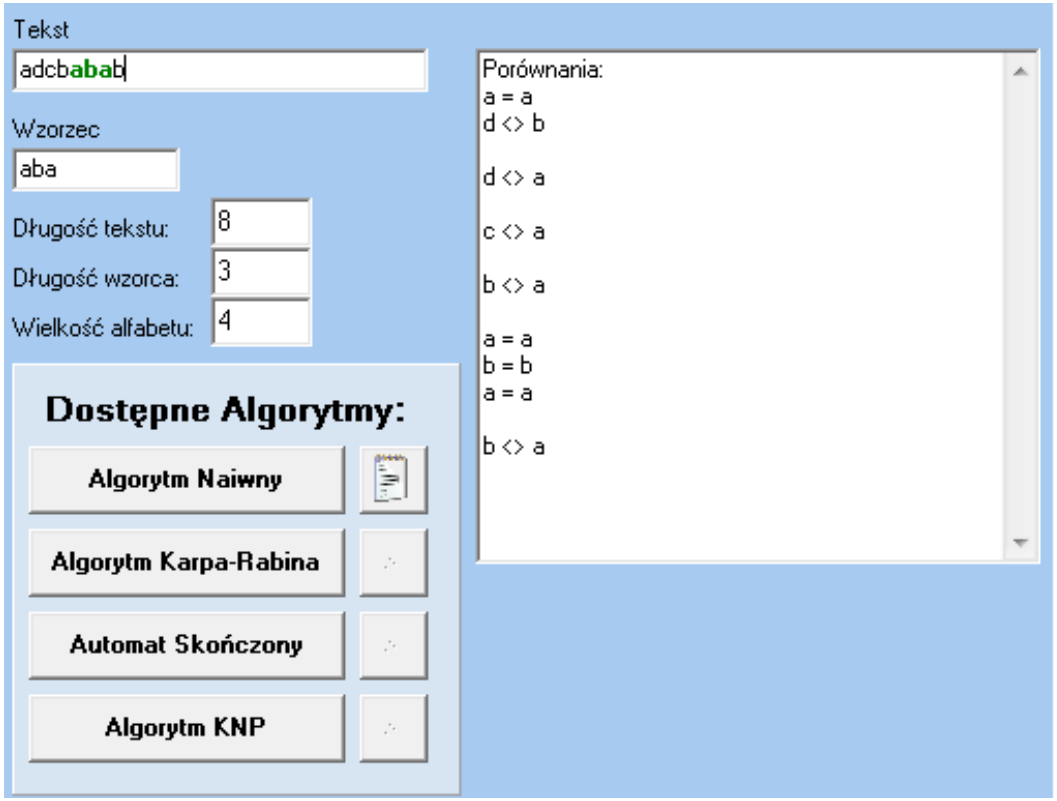
Algorytm Knutha-Morrisa-Pratta: Działanie

- Algorytm działa podobnie jak wyszukiwanie za pomoca automatów skończonych z tą różnicą, że w przetwarzaniu wstępnym nie jest wyliczana cała funkcja przejść δ wielkości $m|\Sigma|$ tylko przy pomocy funkcji pomocniczej π wyliczane są dynamicznie wartości funkcji δ tylko dla wybranego znaku a .
- Wyniki działaniu funkcji π danej wzorem $\pi[q] = max\{k : k < q \wedge W_k \sqsupset W_q\}$ zapisywane są w tablicy $\pi[1...m]$, z której korzysta się przy operacji wyszukiwania. Proces ten jest podobny do tego stosowanego w wyszukiwaniu za pomocą automatu skończonego.

Złożoność obliczeniowa

- Algorytm wykonuje przetwarzanie wstępne w czasie $\Theta(m)$, a operacje wyszukiwania w czasie $\Theta(n)$

Do pracy dołączony jest program komputerowy służacy do znajdowania wzorca w tekście przy pomocy wyżej opisanych metod



Dziękuje za uwagę.