Zadania domowe. Zestaw 2.3

Maciej Poleski

29 kwietnia 2013

1

Tworzymy automat Aho-Corasick przy użyciu danego zbioru słów (algorytm został opisany na wykładzie). Poszukiwane słowo istnieje jeżeli z korzenia możemy dotrzeć do cyklu nie wchodząc do stanu akceptującego (terminalnego) (również w cyklu). Takie słowo składa się z liter na ścieżce do cyklu po czym następują kolejne powtórzenia kolejnych liter cyklu. Z drugiej strony jeżeli istnieje nieskończone słowo nie zawierające żadnego ze słów ze zbioru A, to istnieje spacer z korzenia drzewa nie przechodzący przez żaden stan akceptujący. Skoro słowo jest dowolnie długie a automat skończony - w którymś momencie spaceru dotrzemy do wierzchołka już odwiedzonego tym samym znajdując cykl.

Pozostaje więc sprawdzić czy możemy dotrzeć z korzenia do cyklu. Możemy to zrobić za pomocą algorytmu DFS. Na jego potrzeby sąsiadami stanu są jego dzieci w drzewie A-C oraz stan wskazany krawędzią MSM wyłączając te stany które są akceptujące. Jeżeli korzeń jest akceptujący - poszukiwane słowo nie istnieje (bo puste słowo jest pod słowem każdego słowa).

Złożoność DFS jest liniowa ze względu na rozmiar drzewa A-C, czyli sumę długości wzorców. Podobnie konstrukcja drzewa A-C.

2

Rozwiązujemy problem wyszukiwania wielu wzorców w tekście (np. algorytmem Aho-Corasick). A jest zbiorem wzorców. Jako tekst bierzemy napis składający się z wszystkich wzorców oddzielonych od siebie znakami nie należącymi do alfabetu wzorców. Jeżeli jakiś wzorzec a jest pod słowem innego wzorca b, to w tekście w fragmencie b dopasujemy wzorzec b oraz jego pod słowo a. Jeżeli znajdziemy gdzieś takie dopasowanie - wyznacza ono parę wzorców z których jeden jest pod słowem drugiego. Ponieważ wzorce są rozdzielone znakami nie należącymi do alfabetu nie ma ryzyka dopasowania wzorca do sklejonej pary wzorców. Inaczej mówiąc wszystkie dopasowania mieszczą się w fragmentach odpowiadającym wzorcom.

Złożoność taka jak w algorytmie Aho-Corasick czyli liniowa ze względu na sumę długości wzorców (długość tekstu jest liniowa ze względu na sumę długości wzorców).