

Abstrakt

Algoritmy na vyhľadávanie vzorky v texte majú široké využitie napríklad v textových procesoroch, vyhľadávačoch, či bioinformatike. V tejto bakalárskej práci sa zaoberáme algoritmami na presné vyhľadávanie všetkých výskytov jednej vzorky v texte. Poskytujeme prehľad, vysvetlenie a odhady zložitosti nasledujúcich algoritmov: naivný, algoritmus so základným predspracovaním vzorky, Knuth-Morris-Pratt algoritmus a jeho online verzia, algoritmus Boyer-Moore a jeho Apostolico-Giancarlo verzia. Všetky uvedené algoritmy, okrem naivného, používajú predspracovanie vzorky s cieľom zlepšenia časovej zložitosti. Cieľom tejto práce bolo všetky spomenuté algoritmy implementovať, testovať a porovnať z hľadiska počtu explicitných porovnaní dvoch znakov a z hľadiska reálneho času behu programu na rôznych dvojiciach vzorka-text. Okrem prípadu, keď sa text aj vzorka skladali z jediného znaku, dosahoval najlepšie výsledky algoritmus Boyer-Moore. Dobré výsledky dosahovali aj samostatné implementácie pravidiel Zlého znaku a Dobrého suffixu, ktoré využíva algoritmus Boyer-Moore. V praxi je teda najlepšou voľbou implementovať algoritmus Boyer-Moore, alebo jedno z jeho pravidiel.