SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Implementacija FM-index algoritma

Ivan Borko, Sofia Čolaković, Florijan Stamenković
Voditelj: *doc. dr. sc. Mirjana Domazet Lošo*

Zagreb, siječanj 2015.

SADRŽAJ

1. Uvod i problematika

Pretraživanje teksta česta je praktična potreba mnogih informacijskih sustava. Pod terminom "pretraživanje teksta" podrazumijevamo pronalazak svih pojavljivanja nekog niza znakova Q (engl. *query*) unutar drugog niza znakova S (engl. *string*). Tipično se rezultat pretraživanja R formulira kao niz indeksa (rednog broja znaka) unutar niza S na kojem počinje pojavljivanje niza Q. Primjerice, za niz $S =$ *"Žuti pas je opasan kad je opasan remenom oko pasa"* i niz $Q =$ *"pas"* rezultati pretraživanja bili bi $R = \{6, 14, 28, 46\}$.

U području bioinformatike pretraživanje teksta koristi se u za pronalazak specifičnih sekvenci unutar zadanog genoma. Definicija pretraživanja je jednaka. Specifičnost bioinformatičkog pretraživanja jest da su nizovi koji se pretražuju iznimno velike duljine. Primjerice, ljudski genom tipično sadrži oko 3.3×10^9 znakova, što bi otisnuto na A4 stranice fontom veličine 10pt rezultiralo s otprilike milijun stranica.

Postoje mnogi algoritmi pretraživanja teksta koji na jednostavan način ispunjavaju definirane zahtjeve. Iz perspektive računalne složenosti algoritama, jednostavno je implementirati pretraživanje teksta koje radi u linearnom vremenu. Nažalost, za nizove vrlo velike duljine linearno vrijeme znači praktično predugo trajanje pretraživanja. Otud potreba, pogotovo u području bioinformatike, za vremenski sub-linearnim algoritmima pretraživanja.

U ovom projektu razmatramo implementaciju pretraživanja teksta koja se bazira na konceptu FM-indexa. Konkretna implementacija bazira se na binarnim stablima valića (engl. *wavelet-trees*). Teoretsko razmatranje i praktično testiranje pokazuju da ovakva implementacija pretraživanja ima vremenski sub-linearnu složenost.

2. FM-index algoritam

3. Implementacija i testiranje

4. Zaključak

Zaključak.

5. Literatura

6. Sažetak

Sažetak.