Merkkijonomenetelmät (syksy 2008)

Harjoitus 2

Niko Välimäki   ti 16.09. klo 10.15–12   CK111

1. Luennolla mainittu Horspool-algoritmi käyttää vain esiintymäheuristiikkaa. Jos aakkosto on pieni, keskimääräinen siirtymä jää lyhyeksi. Tilannetta voi parantaa tekemällä siirto yhden merkin sijasta q:n merkin perusteella.

   (a) Analysoi algoritmin keskimääräistä aikavaatimusta q:n merkin esiintymäheuristiikalla. (Asymptoottinen aikavaatimus riittää.)

   (a) Mikä on optimaalinen q:n arvo? (Asymptoottinen arvo riittää.)

   Tehtävässä oletetaan, että aakkoston Σ kukin merkki esiintyy samalla todennäköisyydellä $1/\sigma$ jokaisessa P:n ja S:n positiossa.

2. Triviaali- ja KMP-algoritmit ovat *aakkostoriippumattomia*: ne tekevät vain merkkivertailuja ja niiden toiminta ei riipu aakkostosta, erityisesti niiden pahimman tapauksen aikavaatimus ei riipu aakkoston koosta. BM-algoritmin toiminta sen sijaan riippuu aakkostosta.

   (a) Anna esimerkkejä tilanteista, joissa aakkostoriippuvuus rajoittaa BM-algoritmin käyttökelpoisuutta.

   (b) Saadaanko BM-algoritmista aakkostoriippumaton, jos toinen sen heuristiikoista poistetaan?

   (c) Ovatko Karp-Rabin-algoritmi ja shift-or-algoritmi aakkostoriippumattomia? Onko jollakin niistä aakkostoriippuvuus erilaista eli enemmän tai vähemmän rajoittavaa kuin BM-algoritmilla?

3. Hahmottele seuraavien algoritmien yleistys monen hahmon tapaukseen:

   (a) Karp-Rabin

   (b) Shift-or

   Aikavaatimuksen tulee olla parempi kuin kunkin hahmon etsiminen erikseen saman algoritmin perusversiolla.

4. Muodosta AC-algoritmin mukainen jononsovitusautomaatti hahmojoukolle $\{\texttt{SAKARI}, \texttt{SAKU}, \texttt{ARI}, \texttt{MARKO}, \texttt{ILMARI}\}$.

5. Kehitä menetelmä, jolla NFA, joka säännölliselle lausekkeelle A löytää $L(A)$:n esiintymät tekstistä S, pystyy tulostamaan myös ne $L(A)$:n jonot, jotka ko. paikassa esiintyvät. Esim. hahmolla $P = b(\texttt{ac} \cup \texttt{b})^*\texttt{ba}$ tekstistä $S = \texttt{ccbbbacacbacc}$ tulisi löytää 5 esiintymää: $S[2\ldots5]$, $S[3\ldots5]$, $S[2\ldots10]$, $S[3\ldots10]$ ja $S[4\ldots10]$.