

Projektisuunnitelma

Ilari Nieminen

Aihe:

Boyer-Moore-tyyppinen algoritmi monen hahmon sovitukseen

Toteutustapa:

Toteutetaan tuunaamattomat mutta optimaaliset perusversiot sekä *Aho-Corasick*-algoritmista sekä tehtävänannossa mainitusta *Boyer-Moore* tyyppisestä algoritmista.

Algoritmeihin sisäänrakennetaan laskurimekanismi, joka laskee algoritmin suorittamien operaatioiden lukumäärän merkkijonohaun aikana. Operaatioiksi lasketaan sijoituslauseet (esim $x=1$), vertailut (esim $x<1$), funktiokutsut, *if*-, *for* ja *while* komennot sekä yksinkertaiset laskutoimitukset (esim $x+1$). Esimerkiksi lausekke *if* ($x<y+1$) *then* $z=3$ katsotaan neljän operaation väärtiksi. Laskuri huomioi ainoastaan ongelman ratkaisemisen kannalta välttämättömät operaatiot.

Algoritmien suorituskykyä verrataan näiden tekemien operaatioiden määrän perusteella. Tällöin toteutuksessa voi hyödyntää hitaita mutta helppokäyttöisiä työkaluja ja algoritmeja ei tarvitse tuunata suoritusympäristöön sopiviksi, ainoastaan minimimäärän operaatioita tekeviksi.

Testiaineistona käytetään valmiita luonnollisen kielen tekstejä sekä ohjelmallisesti generoituja aineistoja. Tavoitteena on tuottaa laaja vertailuaineisto parametrien *{tekstin pituus, hahmojen määrä, aakkoston koko, esiintymien määrä, likimääräisten esiintymien määrä}* yli ja selvittää millä parametriavaruuden alueilla jos millään Boyer-Moore tyyppinen algoritmi päihittää Aho-Corasick algoritmin.

Algoritmien lisäksi toteutetaan testiohjelma, joka ottaa syötteekseen tekstitiedoston sekä tiedoston haettavia merkkijonoja, suorittaa haun molemmilla algoritmeilla ja tulostaa vertailutietoja algoritmien suorituksesta.

Toteutusympäristö:

Algoritmit ja testiohjelma toteutetaan *Javalla* ilman graafista käyttöliittymää. Testitulokset kootaan, tiivistetään ja esitetään havainnollistavina diagrammeina.

Työnjako:

Meitsi tekee kaiken.