Määrittelydokumentti

Mikko Peltonen 31.10.2012

Helsingin yliopisto, Tietojenkäsittelytieteen laitos

Tietorakenteiden ja algoritmien harjoitustyö, syksy 2012, 2. periodi

ohjaaja Kristiina Paloheimo

Sisältö

1. Aiheen esittely

Harjoitustyöni aihe on Huffman-koodiin perustuva tiedon pakkaus- ja purkuohjelma. Ohjelma toimii ainoastaan komentoriviltä, ja sille annetaan syötteinä haluttu operaatio (pakkaus tai purku), ja operaatiosta riippuen joko pakattavan tai pakatun tiedoston nimi. Pakkausoperaatio ottaa syötteekseen pakattavan tiedoston nimen. Se käy läpi pakattavan tiedoston ensin kertaalleen, ja hakee huffman-koodit kullekin tiedostossa esiintyvälle merkille. Sen jälkeen ohjelma kirjoittaa pakkaustiedoston alkuun syötteestä generoidun sanaston (kaikki esiintyvät merkit ja kutakin merkkiä vastaavan Huffman-koodin). Ohjelma lukee sitten lähdetiedoston uudestaan, ja kirjoittaa pakkaustiedostoon alkuperäisen tiedon Huffman-prefiksikoodilla pakattuna. Purkuoperaatio ottaa syötteekseen purettavan tiedoston nimen. Se yrittää avaa annetun tiedoston, lukea sieltä ensin sanaston, ja sitten pakatun datan. Ohjelma purkaa datan annettua sanastoa käyttäen, ja kirjoittaa puretun datan toiseen tiedostoon.

Ohjelma kirjoitetaan Javalla, ja kehitysympäristönä käytän Eclipseä. Versionhallintana käytän GitHubia:

https://github.com/belzzi/TiRaLabra

2. Käytetyt tietorakenteet ja algoritmit

Ohjelmaan tarvitaan binäärihakupuu, jonka hakupolusta Huffman-koodi muodostuu. Prioriteettijono tarvitaan merkkien järjestämiseen esiintyvyyden perusteella. Prioriteettijonon toteutus tehdään minimikekona. Lisäksi työssä tarvitaan varmasti erilaisia taulukkorakenteita. Tässä vaiheessa suunnittelua en osaa vielä arvioida tarvitsenko tavallisten kaksiulotteisten taulukoiden lisäksi hajautustaulukoita tms., mutta saatan joutua toteuttamaan työn apuvälineiksi myöhemmin muita tukitietorakenteita.

3. Testaus

Kehityksenaikainen testaus suoritetaan JUnit-testauksella. Testitapaukset dokumentoidaan JavaDocilla. Lopputestaus suoritetaan vielä käsin useita erilaisia ja kokoisia tekstisyötteitä käyttäen. Esimerkiksi seuraavanlaiset syötteet voitaisiin testata:

1. Pienehkö tekstitiedosto (n. 1 kB)
2. Keskikokoinen tekstitiedosto (n. 20 kB)
3. Suuri tekstitiedosto (n. 1 MB)
4. Suurehko XML-tiedosto (n. 500 kB)

Testauksessa otetaan huomioon suoritusaika ja testausympäristö. Suoritusajan mittaamiseen käytetään ohjelmaan toteutettavaa laskuria, joka tulostaa aina ajoajan suorituksen lopuksi. Tällä mittaustavalla eliminoidaan Java-virtuaalikoneen käynnistymiseen ja alustamiseen kuluva aika.

4. Rajoitukset

Ohjelma rajoittuu tekstimuotoisten syötteiden käsittelyyn.

5. Mahdolliset laajennukset tulevaisuudessa

Ohjelma suunnitellaan siten, että pakkausalgoritmin voi vaihtaa helposti toteutusluokkaa vaihtamalla. Esim. Huffman-pakkausta käytettäessä valittaisiin pakkauksen toteutusluokaksi com.mtspelto.tiralabra.HuffmanPakkaaja, Lempel-Ziv 77:aa käytettäessä vastaava luokan nimi voisi olla com.mtspelto.tiralabra.LZ77Pakkaaja.

Toteutusluokan voisi valita esim. antamalla Javan tietyn system propertyn komentoriviltä (esim. java –Dpakkausalgoritmiluokka=com.mtspelto.tiralabra.HuffmanPakkaaja ....)

6. Tavoitteeliset aikavaativuudet

- Sekä minimikekoon lisääminen että keon pienimmän alkion palauttaminen alkio samalla poistaen ovat aikavaativuudeltaan logaritmisia, O(log n)
- Minimikeon pienimmän alkion palauttaminen alkio samalla poistaen on aikavaativuudeltaan O(log n)
- Huffman-algoritmi, joka ottaa syötteekseen järjestetyn listan merkkien esiintyvyyksistä (prioriteettijono) ja luo niistä Huffman-puun, aikavaativuus on O(n log n)
- Pakatessa lähdetiedoston toinen skannaus vie lineaarisen ajan O(n)
- Pakatessa kohdetiedoston kirjoittaminen vaatii Huffman-koodin hakemisen hakupuusta
- Purkaessa lähdetiedoston luku vie lineaarisen ajan O(n)
- Purkaessa Huffman-algoritmia ei ajeta, mutta jokainen merkki haetaan lähdetiedostosta luetusta Huffman-puusta – aikavaativuus logaritminen O(log n)
- Purkaessa kohdetiedoston kirjoittaminen vie lineaarisen ajan O(n)

Näin ollen sekä pakkaus- että purkuohjelman tavoitteellinen aikavaativuus voisi olla luokassa O(n).

7. Lähteet

1. Tietorakenteet-kurssin (syksy 2005) luentomoniste (Matti Luukkainen & Matti Nykänen)
2. http://en.wikipedia.org/wiki/Huffman_coding