Sanaindeksi
Mrittelydokumentti

Patrik Ahvenainen*
Helsingin yliopisto
Fysiikan laitos

1. syyskuuta 2012

Työssä tehdään sanojen indeksointiohjelma, joka indeksoi kaikista sille annetuista tiedostoista kaikki niistä löytyvät sanat. Sanoista talletetaan tiedoston järjestysnumero ja sanan rivinumero. Sama sana voi löytyä useammasta paikasta, jolloin tallennetaan kaikki rivit, joista sana löytyy. Ohjelman tulee osata etsi myös tietyllä merkkijonolla alkavia sanoja. Ohjelmalle voidaan antaa useita sanoja, jolloin se ilmoittaa löytyykö mistään tiedostosta (tai miltään riviltä) kaikkia annettuja sanoja.

Ohjelmalle annetaan syötteeksi (englanninkielisiä) tekstitiedostoja (tiedostojen nimi) ja ohjelma lukee niistä kirjain- ja numeromerkeistä koostuvat sanat sekä mahdolliset tavuviivat yhdyssanojen välistä. Ohjelman ei ole tarkoitus indeksoida erikois- ja välimerkkejä. Indeksoinnissa kaikkia kirjaimia käsitellään isoina kirjaimina.

Tietorakenteena käytetään kahta erilaista puurakennetta: trie-puuta sekä punamustaa puuta. Indeksiä rakennetaan kertaalleen tiedostoa luettaessa ja sen jälkeen hakuja voidaan tehdä nopeasti ilman pääsyä kovalevyllä olevaan tiedostoon. Tiedoston lukeminen sana kerrallaan toteutetaan erillisenä luokkana ja luetut sanat syötetään valinnan mukaan joko trie- tai punamustaan puuhun.

Trie-puussa jokaisen solmun arvo määräytyy solmusta itsestään ja kaikista sen vanhemmista. Tässä työssä siis solmun arvo on sana ja jokainen siirtymä solmujen välillä merkitsee yhtä merkkiä ja kulkemalla juuresta solmuun muodostuu sana (tai sanan alku). Lapsisolmuja Trie-puussa voi olla enemmän kuin kaksi. Koska hyväksyttyjä merkkejä on rajallinen määrä (aakkoset, numerot ja tavuviiva) käytetään trie-rakenteen lapsilistan toteutukseen tässä maksimitaulukkoa. Tämä mahdollistaa lapsilistan läpikäymisen vakioajassa, jolloin toteutuksessa painotetaan hakemisen nopeutta. Linkitetty järjestämätön lapsilista sopisi toteutukseen, jossa painotettaisiin indeksoinnin tekemisen nopeutta hakuajan kustannuksella. Koska

*patrik.ahvenainen@alumni.helsinki.fi / 013326292

Taulukko 1: Punamustan puun säännöt [1].

1. Solmu on joko punainen tai musta.
2. Juuri on musta.
3. Kaikki lehdet (None-tyyppisiä) ovat mustia.
4. Jokaisen punaisen solmun kummatkin lapset ovat mustia.
5. Jokainen suora polku solmusta mihin tahansa sen lehteen sisältää yhtä monta mustaa solmua.

jokaisen esiintymän paikka halutaan tallentaa on tilavaativuus lineaarisesti riippuva sanojen lukumäärän n ja keskimääräisen sanan pituuden m tulosta ($\mathcal{O}(mn)$). Lisäksi haun nopeuttamiseksi suuri osa lapsilistataulukosta on tyhjänä (maksimitaulukko). Aikavaativuus sanan, jonka pituus on l, haulle on $\mathcal{O}(l)$. Aikavaativuus sanan lisäykselle on myös $\mathcal{O}(l)$, jos oletaan, että lapsilistat ovat valmiiksi luotuja.

Punamusta puu on tasapainoitettu puu, joten siinä ideaaliset yksinkertaiset haku- ja lisäysalgoritmit tapahtuvat ajassa $\mathcal{O}(\log n)$ [2], jossa n on sanojen lukumäärä ja tilavaativuus on $\mathcal{O}(n)$. Sen muodostussäännöt ovat listattu taulukkoon 1. Koska sanat eivät ole yksinkertaisia vakioita, vaan merkkijonoja, vaikuttaa myös sanojen keskimääräinen pituus m tilavaativuuteen $\mathcal{O}(mn)$ ja aikavaativuuteen $\mathcal{O}(m \log n)$. Punamustan puun toteutukseen ei lisätä mahdollisuutta hakea sanan alun perusteella.

Käytännössä tekstin sanojen keskimääräinen pituus on lähinnä kirjoituskielestä riippuva vakio ($\mathcal{O}(m) = \mathcal{O}(1)$).

Lähteet:

1. Esa Junttila: Trie-rakenne. http://www.cs.helsinki.fi/u/ejunttil/opetus/tiraharjoitus/trie.html. 10.8.2005. Haettu 31.7.2012.
2. Matti Luukkainen: 58131 Tietorakenteet. Luentomuistiinpanot. Kevät 2011.
3. Wikipedia: Trie. http://en.wikipedia.org/wiki/Trie. Haettu 37.7.2012.
4. Wikipedia: Red-black tree. http://en.wikipedia.org/wiki/Red%E2%80%93black_tree. Haettu 37.7.2012.