Tiedonhakumenetelmät

Helena Ahonen-Myka

Kevät 2004, osa 9

Merkkijonohahmon etsintä selaamalla

Tässä osassa

- merkkijonohahmon etsintä selaamalla
  - käyttö: hakusanan etsintä tekstistä, kun dokumenttikokoelma on pieni
  - menetelmiä
    - suoraviivainen menetelmä
    - tehokkaampia menetelmiä: KMP, BM

Merkkijonohahmon etsintä selaamalla

- kun dokumenttikokoelmat ovat suuria, on käänteistiedosto yleensä aina paras vaihtoehto hakujärjestelmän toteuttamisessa
- jos dokumenttikokoelma on pienehkö ja mahtuu keskusmuistiin, voidaan hakusanan etsintä dokumenteista toteuttaa myös vertaamalla hakusanaa suoraan dokumentin tekstiin
- yleinen käyttötilanne: hakutulosten jälkikäsittely
  - esim. läheisyysoperaation toteuttaminen: esiintyvätkö hakusanat dokumentissa riittävän lähellä

Merkkijonohahmon etsintä selaamalla

- hakusanaa etsitään peräkkäishakuna vertaamalla hakusanan merkkejä dokumentin merkkeihin aloittaen dokumentin ensimmäisestä merkistä
- oletetaan, että dokumentti S on merkkijono
  - $S = s_1 s_2 ... s_n$, kukin s_i on aakkoston merkki
- ja etsittävä hahmo P on myös merkkijono
  - $P = p_1 p_2 ... p_m$, kukin p_j on aakoston merkki
- $m \leq n$

Esimerkki

- dokumentti S: abracabracadabra
- hahmo P: abracadabra

Suoraviivainen perusmenetelmä

```
abracabracadabra
abracadabra
 abracadabra
  abracadabra
   abracadabra
    abracadabra
     abracadabra
```

Suoraviivainen perusmenetelmä

- pahimmassa tapauksessa hahmo sopii tekstiin joka vertailussa aina viimeiseen merkkiin asti, mutta viimeinen merkki ei täsmää
- vertailuja tarvitaan n · m kappaletta → O(nm)
- S = aaaaaaaaaab, P = aab
- useimmiten tavallisessa tekstissä hahmon epäsopivuus ko. kohtaan todetaan jo muutaman merkin jälkeen

```
S – aaaaaaaaab
P – aab
    ++-
-----------------------------
S – aaaaaaaaab
P –  aab
      ++-
-----------------------------
...
S – aaaaaaaaab
P –          aab
             ++-
-----------------------------
S – aaaaaaaaab
P –           aab
              +++  (täsmää!)
```

Tehokkaampia menetelmiä

- perusmenetelmä siirtää hahmoa aina vain yhdellä merkillä eteenpäin
- menetelmä ei yritä käyttää hyväksi tietoa hahmon sisältämistä merkeistä
- tehokkaammat menetelmät analysoivat ensin hahmon ja tunnistavat merkkien toistumisen hahmossa
- analyysin perusteella voidaan hahmoa usein siirtää eteenpäin useampia merkkejä kerrallaan
- menetelmiä: KMP (Knuth-Morris-Pratt) ja BM (Boyer-Moore)

KMP (Knuth-Morris-Pratt)

- S = ... s_i s_{i+1} s_{i+2} s_{i+3} s_{i+4} s_{i+5} | s_{i+6} s_{i+7}...
- P = p_1 p_2 p_3 p_4 | p_5

- hahmon alkuosa $p_{1..4}$ löytyy tekstistä, mutta merkit s_{i+6} ja p_5 eivät ole samoja
- hahmon P esiintymä voi alkaa jo tutkitussa katkelmassa $s_{2..5}$ vain, jos jokin P:n alkuosa on myös tämän katkelman loppuosa

KMP (Knuth-Morris-Pratt)

- S = ...barba | papa...
- P = barba | ari
- P = ba rbaari

- S = ...sey | chellit...
- P = sey | moyr
- P = seymoyr

KMP (Knuth-Morris-Pratt)

- riittää tarkastella vain hahmoa, koska
  - hahmon alkuosa on täsmännyt tekstikatkelmaan
  - tekstikatkelman loppuosa on sama kuin hahmon alkuosan loppuosa ennen kohtaa, jossa merkit eroavat
- hahmon esikäsittely:
  - etsitään hahmon alimerkkijonot, jotka toistuvat hahmossa
  - muodostetaan next-taulukko
    - next[j] kertoo, mikä on hahmon osan $P_{1..j-1}$ pisin aito alkuosa, joka on myös osan $P_{1..i-1}$ loppuosa
    - lisäksi merkit, jotka seuraavat ko. alkuosaa ja loppuosaa, eivät ole samoja

KMP (Knuth-Morris-Pratt)

- P = a b r a c a d a b r a
- next = 0 0 0 0 0 1 0 1 0 0 0 0 4

- S = abraca | bracadabra
- P = abraca | d
- P = a bracadabra

KMP (Knuth-Morris-Pratt

- hahmon esikäsittely voidaan tehdä O(m) ajassa
- algoritmi tutkii jokaisen tekstin merkin ja jokaista merkkiä kohti vähintään yhden hahmon merkin → vähintään 2n vertailua
- → O(m + n)
- käytännössä KMP ei välttämättä toimi paremmin kuin suoraviivainen menetelmä
- menetelmä on helposti laajennettavissa tilanteeseen, jossa hahmoja on useita
  - kaikkien hahmojen esiintymät haetaan samanaikaisesti

BM (Boyer-Moore)

- hahmon ja tekstin vastaavuutta voidaan tarkastella myös hahmon lopusta alkuun päin
  - kun KMP-algoritmissa tarkasteltiin kullakin hetkellä hahmon alkuosaa, tarkastellaan BM-algoritmissa hahmon loppuosaa
- kaksi periaatetta siirtää hahmoa tekstin suhteen
  - täsmäytysperiaate (match heuristics)
  - esiintymäperiaate (occurrence heuristics)
- kumpikin periaate antaa siirtoaskelien lukumäärän
- suurempi lukumäärä voittaa

BM (Boyer-Moore)

- täsmäytysperiaate
  - vastaa KMP-algoritmin next-taulukkoa
  - talletetaan jokaiselle hahmon loppuosalle tieto siitä, toistuuko loppuosa hahmossa
  - kun hahmon lopusta alkuun kuljettaessa tulee vastaan kohta, jossa hahmon ja tekstin merkit eivät täsmää, voidaan turvallisesti siirtää hahmon edellinen ko. loppuosan esiintymä tähän tekstikohtaan

BM (Boyer-Moore)

- esiintymäperiaate
  - oletetaan, että ”c” on tekstissä esiintyvä merkki, jonka kohdalla hahmon loppuosa ei täsmää tekstiin
  - jos ”c” esiintyy hahmossa, voidaan hahmoa siirtää niin, että hahmon ”c” tulee tekstin ”c”:n kohdalle
  - jos ”c” ei esiinny hahmossa, voidaan hahmo siirtää tekstin ”c”:n ohi oikealle puolelle

BM (Boyer-Moore)

- S = abracababra...
- P = abracadabra

- b ≠ d
- täsmäytysperiaate
  - ”abra” löydetty → hahmoa voidaan siirtää turvallisesti 7 askelta (hahmon 1. ”abra” tekstin käsittelykohtaan)

- S = abracababra...
- P = abracadabra

BM (Boyer-Moore)

- esiintymäperiaate
  - jos ”b” on osa hahmoa, voidaan hahmon lähin vasemmalla oleva ”b” siirtää tekstin ”b”:n kohdalle → voidaan siirtää 5 askelta
- S = abracababra...
- P = abracadabra
- valitaan suurempi askelmäärä, eli 7 askelta

BM (Boyer-Moore)

- BM ei välttämättä tutki tekstin jokaista merkkiä
- keskimääräinen vertailujen määrä O(n log(m) / m), pahin tapaus O(mn)
- useita muunnelmia
  - käytetään vain esiintymäperiaatetta
  - käytetään vain esiintymäperiaatetta, mutta sovelletaan sitä siihen merkkiin, jota verrataan hahmon viimeiseen merkkiin
  - kuten edellä, mutta sovelletaan merkkiin, joka seuraa hahmon viimeisen merkin kohdalla olevaa merkkiä

Läheisyysoperaatio

- on haettava useampien sanojen esiintymiä siten, että sanojen tulisi esiintyä toistensa läheisyydessä
- jos etsitään fraasia kuten ”tietojenkäsittelytieteen laitos”, voidaan etsintä tehdä samoin kuin yksittäisten sanojen etsintä (välilyönti on vain yksi merkki)
- jos sanojen järjestys ja etäisyys voi vaihdella, kannattaa hakea ensin sanaa, joka esiintyy harvimmin ja/tai on pisin
  - muiden sanojen esiintymät tarkistetaan ensimmäisen sanan ympäristöstä

Tässä osassa

- merkkijonohahmon etsintä selaamalla
  - suoraviivainen menetelmä
  - KMP-algoritmi
  - BM-algoritmi