3. Dokumenttien ryhmittely

= clustering, ryvästäminen

= samanlaiset dokumentit lähekkäin!

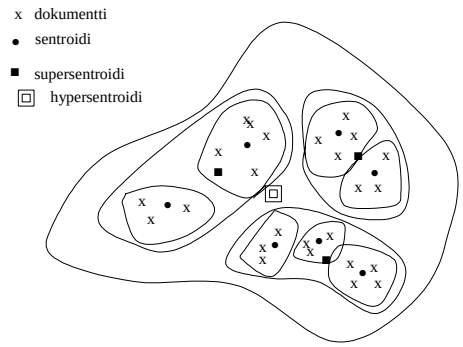
Miksi?
- lähekkäisten tarkastelu joukkona
- lähekkäisten relevanssin päättely
- lähekkäisten selaus voi olla kätevää
- saantitekniikka: lähekkäisten haku tehokasta

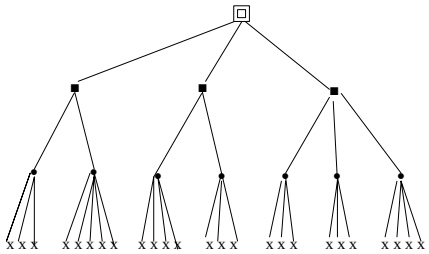
Ryvästäminen = luokittelu?
- tilastollinen tekniikka
- rypäiden määrä ja identiteetti ei etukäteen tiedossa

(relevantit / ei-relevantit = yksinkertainen ryvästys kahteen osaan)

Dokumenttien ryvästys / termien ryvästys
- termit: vrt. termien yhdistäminen indeksoinnissa (co-occurrence), termien käänteistiedostolistat

Rypäs = keskusalkio + ympäristö
=> hierarkkinen tulkinta luonnollinen





Menetelmiä:
- hierarkkinen, 'täydellinen' ryvästys
- heuristinen ryvästys

(+ paljon muita)

Dokumenttien samanlaisuus?
$sim(d_i,d_j)$ eli lähekkäisyys (tai etäisyys)

Rypäiden läheisyys: monia tulkintoja

1° väljä yhteys (single link)

rypäiden läheisyys = läheisimmän parin läheisyys

2° tiukka yhteys (complete link)

- kaukaisimman parin läheisyys ratkaisee (kaikki rypääseen kuuluvat riittävän läheisiä)

3° keskiarvoinen läheisyys (group average link)

- jäsenen keskimääräinen läheisyys ryhmän jäseniin > läheisyys toisen ryhmän jokaiseen jäseneen

1° Hierarkkinen ryvästys (bottom up -algoritmi)

- lähtökohtana termi/dokumentti-matriisi, josta dokumenttien kaikki parittaiset samanlaisuudet (lähekkäisyydet) => samanlaisuusmatriisi ($sim(d_i,d_j)$)

1. Laske kaikki parittaiset samanlaisuudet.

2. Aseta jokainen dokumentti omaksi rypääksi.

3. Yhdistä kaksi läheisintä ryvästä i,j yhdeksi sekä päivitä samanlaisuusmatriisi:
poista rivit ja sarakkeet i ja j,
lisää rivi rypäälle (i,j).

4. Toista askelta 3, kunnes vain yksi (?) ryväs jäljellä.

- mahdollisesti lopetus aikaisemmin; välivaiheissa saadaan joka tapauksessa erilaisia ryväskokoelmia (riippuen lähekkäisyydelle asetetusta vaatimuksesta)

(Esimerkkejä: ks. Salton, s. 330-337.)

Algoritmin toiminta (yhdistämisaskel):

- single link: jokainen pari kasvattaa jotakin ryvästä (paitsi jos samanlaisuusarvo redundantti)
  - matriisi: rypään ja ulkopuolisten samanlaisuus läheisimmän mukaan (= maksimi vaihtoehdoista)

- complete link: tulee tietää kaikki (syntyvän) rypään parien samanlaisuusarvot => enemmän kirjanpitoa
  - matriisi: rypään ja ulkopuolisten samanlaisuus kaukaisimman mukaan (= minimi vaihtoehdoista)

- group average link: välimuoto
  - matriisi: rypään ja ulkopuolisten samanlaisuudeksi keskiarvo

Käytäntö:

- single link
  - yksinkertaisempi laskea
  - tuottaa vähän, mutta 'heikkoja' rypäitä
  - tiedonhaun tarkkuus voi olla heikko

- complete link
  - laskennallisesti raskas
  - rypäät kiinteitä, lkm voi olla suuri
    - voi jäädä helposti jopa erillisiä dokumentteja
  - tiedonhaun tarkkuus hyvä

Päätösehto algoritmissa:
- rypäiden määrä haluttu (sopivan pieni)
- rypäiden koko sopiva
- samanlaisuudelle asetettu kynnysarvo: yhdistetään vain, jos on (vielä) riittävän samanlaisia

2° Hierarkkinen top down –algoritmi

- alussa kaikki samassa rypäässä
- jaetaan rypäitä peräkkäin osiin l. uusiin rypäisiin
  - periaatteessa selkeä, mutta millä perusteella?
  - tarvitaan mitta rypään kiinteydelle tms. (ainakin dokumenttien etäisyydet kaikista muista)
- käytetään vähän

3° Heuristinen ryvästys

- tavoitteena karkea jako halvalla

Perusalgoritmi (one pass):

- dokumentit mielivaltaisessa järjestyksessä, samanlaisuus olemassaoleviin rypäisiin määriteltävissä

- dokumentti liitetään rypääseen, jos se on riittävän lähellä
  - rypään jotakin alkiota
  - rypään sentroidia
  - (- rypään kaikkia alkioita)

- tulos:
  - rypäillä voi olla yhteisiä alkioita
  - voi tulla suuria rypäitä
  - voi jäädä irrallisia alkioita
  - riippuu alkioiden käsittelyjärjestyksestä
  - kokeiden mukaan tulokset hyviä

Näitä voidaan kontrolloida:
- max-koko => rypään jako kahtia
- päällekkäisyys: riittävä kynnysarvo mahdollisesti muuttuva
- irralliset alkiot: yhdistäminen (jopa vain 'teknisesti', kynnysarvosta tinkien)

Muita menetelmiä:

- dokumenttiavaruuden tiheyden selvitys
  - valitaan tiheältä alueelta 'sentroidi' ja kootaan ryväs sen ympärille

- dokumentin siirto toiseen rypääseen jonkin alkutilanteen jälkeen (refinement)
  - esim. vertailu kaikkien rypäiden sentroideihin, siirto läheisimpään

- dokumenttien yhteyteen perustuvat siirrot
  - saman kyselyn tulosjoukossa (1,2,...)
  - sama relevanssi(palaute)arvio

Haku rypäiden avulla

- ryväshypoteesi: rypään dokumenttien relavanssi kyselyyn samansuuntaista
- sentroidi rypään edustajana
- top down - ja bottom up –menetelmät mahdollisia
- karkeasti: kaikki rypään alkiot tulokseen
- jatko: alkiot tutkitaan erikseen

Top down - haku:

Oletetaan kysely q sekä ryväshierarkia, sen sisältö vähitellen 'työlistaan'.
Halutaan tulokseen p dokumenttia.

1. Vie huippusentroidi työlistaan. k <-- p.
2. Ota listasta korkeinta sim(q,C) -arvoa vastaava sentroidi.
   Jos vastaavan rypään alkioiden määrä n(C) <= k, vie alkiot tulokseen ja aseta k <- k-n(C);
   muuten korvaa sentroidi työlistassa jälkeläisillään (ja vie mahdolliset jälkeläis-dokumentit tulokseen).
3. Jos k > 0, toista askel 2.

Bottom up - haku:

- käydään läpi alimman tason sentroideja sim(q,C) -järjestyksessä, kunnes saadaan haluttu määrä alkioita tulokseen
  - useita tapoja valita tulokseen:
    koko ryväs,
    parhaat alkiot useasta (kärkipään) rypäästä, ...

Ryvästyshakujen käyttö?

- suurten tiedostojen ongelmat (paljon sivuhakuja verrattuna käänteis-hakemistoihin)
- rypäiden ylläpito (jos/kun kokoelma ei ole staattinen)
- paljon sentroidivertailuja
- selausmahdollisuus
  - alkioiden valinnassa (ellei oteta koko ryvästä)
  - rypäiden valinnassa (heuristiikat)
    vrt. Scatter/Gather-tekniikka
  - selauksen hyöty vaikea mitata …

Scatter/gather-tekniikka

- käytetään ryvästystä iteratiivisesti tavanomaisen kyselyn tuloksen ryhmittelyyn:
  - järjestelmä tekee ensimmäisen ryhmittelyn (scatter-vaihe) ja kuvailee rypäitä asiasanoilla
  - käyttäjä ilmaisee, mitkä rypäät ovat kiinnostavia gather-vaihe
  - järjestelmä tekee uuden ryhmittelyn jne

Esimerkki:

- kysely 'star' (taivaalta, viihdemaailmasta, ...)

- tulos (rajoitettu): 250 dokumenttia, 5 rypäässä:

cluster 1 size: 8 key army war francis banner air ...
  - dokumenttien otsikot ('star' symbolina …)
cluster 2 size:68 film player career win television ...
  (elokuva-, tv-tähdet)
cluster 3 size: 97 bright magnitude constellation period
  (astrofysiikkaa)
cluster 4 size: 67 astronomer observatory telescope …
  (tähtitiedettä, astrofysiikkaa)
cluster 5 size: 10 family species flower animal arm …
  (eläimiä, kasveja; 'star' esim. nimessä)

- käyttäjä voi esim. pyytää rypään 2 uudelleen käsittelyä, 3 rypääseen:

cluster 1 size: 14 player league hit game national set ...
  (urheilijoita)
cluster 2 size: 47 role stage broadway comedy …
  (elokuva, tv, …)
cluster 3 size: 7 music country jazz …
  (musiikki)

Huom.
- Uusi ryvästys voi tuoda näkyviin uusia asioita (kaikkia ei näytetty / katsottu).
- Käyttäjä voi jatkaa edelleen poimimalla jonkin rypään (tai useita) ensimmäisestä tai toisesta ryvästyksestä:
  esim. ensi vaiheen suuri ryväs 3 (97 dokumenttia) jne

Ominaisuuksia:
- rypäiden koko ja lkm? alunperin lkm piti ilmoittaa
- mitä kaikkea kannattaa näyttää: rypäät, niiden kuvailu?
- kuinka rypään 'vähemmistödokumenteista' voitaisiin ilmaista riittävästi (ilman uutta ryvästystä)?

-- tarkemmin harjoitustehtävänä --