2.3. Vektorimalli

· dokumenttien ja kyselyiden esitys termivektoreina
→ vektoriavaruus

Mitä saavutetaan?

- dokumenttien lähekkäisyys hallintaan
  tiheys, ryhmittely

- kyselyn ja dokumentin vastaavuus (samanlaisuus)
  - dokumenttien järjestys
  - valinta kynnysarvon perusteella
  - kyselyn muokkaus parhaiten sopivien dokumenttien perusteella
  - dokumenttien joukon 'keskipiste' (sentroidi) voidaan määrittää

$d_i = (d_{i1}, d_{i2}, \ldots, d_{it})$

$q_i = (q_{i1}, q_{i2}, \ldots, q_{it})$

d_{ij}, q_{ij} joko esiintymiä (0/1) tai painoja (w_{ij})

Dokumentin ja kyselyn samanlaisuus (vastaavasti kahden dokumentin välinen samanlaisuus, 'etäisyys'):

_ perustuu dokumenttivektorin ja kyselyvektorin sisätuloon:

$$sim(d_r, q_s) = \sum_{i,j} d_{ri} \cdot q_{sj}$$

- termien riippumattomuus ?

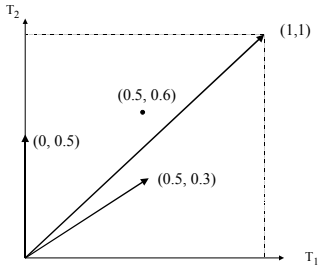
Sisätulon ominaisuuksia
· binääritapaus: vastinparien lkm
· painot: tulojen summa

Sisätuloon perustuvia samanlaisuusmittoja:

- normeeraus kyselyllä:

$$sim(d_r, q_s) = \left(\sum_{i,j} d_{ri} \cdot q_{sj} \right) / \sum_{j} q_{sj}$$

luonnollinen
arvot välille [0,1]



kosinikerroin: normeerataan sisätulo kyselyn ja dokumentin termien neliösummien neliöjuurella

$$sim(d_r, q_s) = \frac{\sum d_{rj} q_{sj}}{\sqrt{\sum d_{rj}^2 \sum q_{sj}^2}}$$

- kyselyyn kuulumattomat termit pienentävät arvoa
- dokumenttien termien määrän vaihtelu vaikuttaa

päällekkäisyyskerroin (overlap c.):

$$sim(d_r, q_s) = \frac{\sum d_{rj} q_{sj}}{\min(\sum d_{rj}, \sum q_{sj})}$$

Dicen kerroin:

Jaccardin kerroin:

(+ kymmeniä muita ...)

- eri kertoimet => (hieman) erilaisia samanlaisuusjärjestyksiä

Vektorimallin etuja:
_ käsitteellinen yksinkertaisuus
_ termien painot luontevasti mukana
_ samanlaisuusjärjestys

Vektorimallin ongelmia:
· termien riippuvuus toisistaan
· mitat heuristisia (teoreettiset perustelut ?)
· ei yhteyttä Boolen mallin tyyppiseen termien yhdistämiseen

Kyselyvektorin muuntaminen eli relevanssipalaute (relevance feedback)

· tiedonhaku käytännössä:
peräkkäisiä hakuja, iterointia
(saanti/tarkkuus - ongelma)

Yleinen perusoletus:
relevantit dokumentit lähellä toisiaan

- tämän hyödyntäminen
(puoli)automaattisesti?

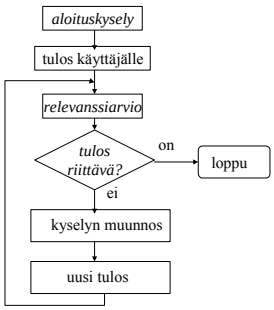
- muunnetaan kyselyä lisäämällä sen samanlaisuutta relevanttien dokumenttien kanssa

Oletetaan: ideaalinen kysely, joka
- maksimoi samanlaisuuden
relevantteihin dokumentteihin
- minimoi samanlaisuuden
ei-relevantteihin dokumentteihin

kuviteltu Q_{opt} =

$$k\left\{\frac{1}{R}\sum_{rel}\frac{d_i}{|d_i|}-\frac{1}{N-R}\sum_{nonrel}\frac{d_i}{|d_i|}\right\}$$

R = relevanttien lkm (tuntematon!),
N-R = ei-relevanttien lkm



Relevantit määrätään jollakin tavalla; tulos syötetään palautteena järjestelmälle:

E aloitetaan sopivalla kyselyllä
· approksimoidaan R tuloksen perusteella käyttäjä arvioi: R' relevanttia, N' = N - R'

uusi kysely vektorina:

$$q^{(i+1)} = q^{(i)} + \frac{1}{R'}\sum_{d_i \in R'} D_i - \frac{1}{N'}\sum_{d_i \in N'} D_i$$

$$= \quad q^{(i)} + \alpha \sum_{d_i \in R'} d_i - \beta \sum_{d_i \in N'} d_i$$

Kertoimet α ja β voidaan valita monella tavalla:

$\alpha = \beta = 0.5$?

$\alpha = 1, \beta = 0$?

"dec hi" : kaikki relevantit huomioon,
vain ensimmäinen ei-relevantti pois

Aloituskysely, tuloksen arviointi?

· pitää saada ei-tyhjä tulos
· satunnainen tulosjoukko?
· jos tarpeen, vähennetään and-ehtoja

Relevanssipalautetekniikan ongelmia:

· dokumenttijoukon jakauma
relevantit hajallaan
epärelevantteja relevanttien lähellä

ryvästäminen;
jos muodostuu useita relevanttien dokumenttien ryppäitä, kysely jakaantuu osakyselyihin

· dokumenttien arviointi
relevanssijärjestyksessä luonnollista, samojen dokumenttien toistuva arviointi on tarpeetonta; menettelyjä:

- kiinnitetään relevantit paikoilleen
('rank freezing')
- poistetaan epärelevantit

Muita muunnostapoja:

- dokumenttijoukon jako
  testi- ja kontrollikokoelmaan

- dokumenttiavaruuden muunnos
indeksejä
  parantamalla
  - lisätään relevantteihin
dokumentteihin
    kyselyvektorin termejä
  - poistetaan termejä ei-relevanteista
  - vain pienin muutoksin …
    yhden kyselyn vaikutukset /
      globaali tilanne ?

  (idea lähellä suosittelujärjestelmiä ?)