Tiedonhakumenetelmät

Helena Ahonen-Myka

Kevät 2004, osa 5

Täsmäytysmenetelmät (1/2)

Tässä osassa

- täystäsmäytys
  - Boolen haku
- osittaistäsmäytys
  - vektorimalli, samanlaisuusmittoja

Täystäsmäytys: Boolen haku

- Boolen kysely:
  - lista termejä, jotka on yhdistetty loogisilla konnektiiveilla AND, OR ja NOT
  - vastauksena ovat ne dokumentit, jotka täyttävät kyselyn määrittelemät ehdot
  - text AND compression AND retrieval
    - dokumentti otetaan mukaan vastaukseen, jos kukin näistä 3 termistä esiintyy dokumentissa (järjestys on vapaa)

Täystäsmäytys: Boolen haku

- "...the **compression** and **retrieval** of large amounts of **text** is an interesting problem..."
- "...this **text** describes the fractional distillation scavenging technique for **retrieving** argon from **compressed** air"...

Boolen kyselyn käsittely

- kysely: "text AND compression AND retrieval"
- kukin kyselytermi (mahdollisesti muokattuna) etsitään sanakirjasta
  - sanakirja kertoo, kuinka monessa dokumentissa termi esiintyy (df)
    - text: 8
    - compress: 4
    - retrieve: 6
- termit järjestetään df:n mukaan nousevaan järjestykseen: compress, retrieve, text

Boolen kyselyn käsittely

- luetaan harvimmin esiintyvän termin käänteislista
- käänteislista = kandidaattien joukko (dokumentit, joita ei ole vielä eliminoitu ja jotka voivat olla vastauksia kyselyyn)
- kaikkien jäljelläolevien termien käänteislistat lomitetaan vuorollaan tämän kandidaattien listan kanssa
  - termit käsitellään frekvenssin mukaan nousevassa järjestyksessä

Esimerkki

- termin 'compress' käänteislista:
  - <4; 2, 5, 12, 16>
- termin 'retrieve' käänteislista:
  - <6; 2, 7, 12, 16, 20, 21>
- "compress AND retrieve"
  - <3; 2, 12, 16>
- termin 'text' käänteislista:
  - <8; 1, 4, 8, 12, 16, 20, 21, 30>
- "compress AND retrieve AND text"
  - <2; 12, 16>

7

AND-kysely

- AND-kyselyssä dokumentti ei voi kuulua vastaukseen, ellei se esiinny kaikissa käänteislistoissa
  - → kandidaattien joukko ei voi kasvaa kyselyn käsittelyn aikana
  - kun termi t käsitellään, käydään läpi kandidaattien joukko ja poistetaan dokumentit, jotka eivät esiinny termin t käänteislistassa
  - kandidaattien joukko voi tyhjetä, ennen kuin kaikki termit on käsitelty
- kun kaikki termit on käsitelty, jäljellejääneet kandidaattijoukon dokumentit muodostavat vastauksen

8

OR-kysely

- "text OR data OR image"
- termit voidaan käsitellä samanaikaisesti: käänteislistojen lomituksessa otetaan kukin dokumentti mukaan vain yhden kerran
  - text: <8; 1, 4, 8, 12, 16, 20, 21, 30>
  - data: <12; 2,4,7,8,10,12,13,15,19,20,21,28>
  - image: <5; 4,5,9,11,12>
- vastaus: <1,2,4,5,7,8,9,10,11,12,13,15,16,19,20,21,28,30>

9

Disjunktioiden konjunktio

- disjunktioiden konjunktio on yleinen kyselytyyppi
  - "(text OR data OR image) AND (compression OR compaction) AND (retrieval OR indexing OR archiving)"
  - kandidaattijoukon alkuarvoksi valitaan "pienimmän" konjunktin dokumenttijoukko: kokoarvio esim. laskemalla yhteen termien df-arvot
    - tämä on pessimistinen arvio: ei ota huomioon joukkojen mahdollista päällekkäisyyttä
  - lomitetaan seuraavaksi "pienin" kandidaattijoukon kanssa jne

10

Yleisemmät kyselyt

- yleisemmät Boolen kyselyt voidaan muuntaa disjunktioiden konjunktioksi
- "(information AND (retrieval OR indexing)) OR ((text OR data) AND (compression OR compaction))"
- → "(information OR text OR data) AND (retrieval OR indexing OR text OR data) AND (information OR compression OR compaction) AND (retrieval OR indexing OR compression OR compaction)"

11

NOT-kyselyt

- NOT-kyselyt eivät voi esiintyä yksinään, vaan ne ovat oikeastaan AND NOT –kyselyjä
- "text AND NOT data"
  - text: <8; 1, 4, 8, 12, 16, 20, 21, 30>
  - data: <12; 2,4,7,8,10,12,13,15,19,20,21,28>
- lasketaan ensin "text AND data"
  - <4,8,12,20,21>
- lomitetaan termin "text" käänteislista ja "text AND data" –käänteislista siten, että poistetaan molemmissa esiintyvät
  - <1,16,30>

12

Täydellisen täsmäytyksen ongelmia

- kyselyyn lähes täsmääviä dokumentteja ei löydetä
- hakujärjestyksen järjestys on satunnainen
- Boolen kyselyitä ei ole helppo muodostaa
- tuloksen koon säätely on vaikeaa

13

Täydellisen täsmäytyksen ongelmia (tarkemmin)

- kyselyyn lähes tai osittain täsmääviä dokumentteja ei löydetä
  - tiedontarvetta voidaan harvoin esittää yksiselitteisesti hakuavainten avulla → jyrkkä raja täydellisesti täsmäävien ja osittain täsmäävien dokumenttien välillä ei ole perusteltu
- hakutuloksen järjestys on satunnainen
  - järjestys = esim. tietueiden tallennusjärjestys
  - olisi parempi saada hakutulos dokumenttien todennäköisen relevanssin mukaan laskevassa järjestyksessä

14

Täydellisen täsmäytyksen ongelmia

- Boolen kyselyitä ei ole helppo muodostaa
  - harjaantumaton käyttäjä tekee helposti virheitä kyselyjen muotoilussa
  - ”Ruotsin ja Norjan hiihtokeskukset” → ”(ruotsi OR norja) and hiihtokeskus”
- tuloksen koon säätely on vaikeaa
  - AND-kyselyjen tulos jää helposti niukaksi
  - OR-kyselyjen tulos voi räjähtää hyvin suureksi

15

Quorum-haku

- täystäsmäytyksen ongelmia voidaan yrittää ratkaista yleistämällä Boolen haku **quorum-hauksi**
- idea: automatisoidaan kyselyn laajennus tasoittain
- esim. käyttäjä antaa termit a,b,c ja d, järjestelmä muodostaa Boolen kyselyt
  - tiukka ehto → löysemmät ehdot

16

Esimerkki

- a and b and c and d
- (a and b and c) or (a and b and d) or (a and c and d) or (b and c and d)
- (a and b) or (a and c) or (a and d) or (b and c) or (b and d) or (c and d)
- a or b or c or d

17

Quorum-haku

- vastauksena palautettavien dokumenttien määrä kasvaa, kun siirrytään tasolta seuraavalle
  - ensimmäisellä tasolla dokumentteja vähän, mutta suhteellisesti enemmän relevantteja
  - ylemmillä tasoilla dokumentteja enemmän, relevantteja suhteellisesti vähemmän
- käyttäjä voi valita sopivan tason, joka tuottaa sopivan määrän dokumentteja sekä kohtuullisen saannin ja tarkkuuden

18

Osittaistäsmäytys

- osittaistäsmäytyksellä yritetään ratkaista täystäsmäytyksen ongelmat:
- kyselyyn osittain täsmäävät dokumentit voidaan löytää
- hakutuloksen järjestys perustuu kyselyn ja dokumentin täsmäävyyden asteeseen
  - hakutulos saadaan dokumenttien todennäköisen relevanssin mukaan laskevaan järjestykseen

19

Osittaistäsmäytys

- kyselyissä ei välttämättä tarvita lainkaan operaattoreita
  - mikä tahansa tekstikappale voidaan kopioida kyselyksi
- tuloksen koon säätely on helppoa
  - hakija ilmoittaa kuinka monta parasta vastausta hän haluaa

20

Vektorimalli

- **vektorimalliin** perustuva täsmäytys on yleisin osittaistäsmäytysmenetelmä
- aikaisemmin oletimme, että jos dokumenttikokoelmassa on t erillistä termiä, jokainen dokumentti kuvataan t termillä (termi + termin paino)
- Boolen haussa voidaan ajatella, että dokumenttia kuvataan t termin **joukolla**
- vektorimallissa kutakin dokumenttia (ja kyselyä) kuvataan t-dimensioisella **vektorilla**

21

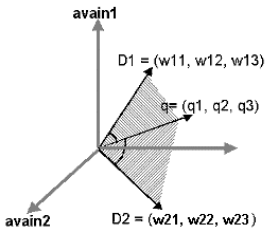
Vektorimalli

- tehdään yksinkertaistava oletus: termit ovat riippumattomia toisistaan → dimensiot ovat ortogonaalisia toisiinsa nähden
- on määriteltävä samanlaisuusfunktio, joka kuvaa dokumentin ja kyselyn (tai kahden dokumentin välistä) samanlaisuutta
- useat vektorimallissa käytettävät samanlaisuusfunktiot perustuvat vektorien sisätuloon:

$$sim(d_i, q_j) = \sum_{k=1}^{t} d_{ik} \cdot q_{jk}$$

22

Vektorimalli



23

Vektorien sisätulo

- jos dokumenttivektorissa termien painot ovat binäärisiä (0 tai 1)
  - sisätulo: vastinparien lkm (vain ykköset)

$$sim(d_i, q_j) = \sum_{k=1}^{t} d_{ik} \cdot q_{jk}$$

  - dokumentti i: (1,0,1) ja kysely j: (0,1,1)
    - sisätulo: 0+0+1 = 1

24

Vektorien sisätulo

- v = (1,1,1,0,0,0,0,0)
- w = (1,1,1,0,0,0,0,0)
- v = (1,1,1,1,1,1,1,1)
- w = (1,1,1,0,0,0,0,0)
- molemmissa tapauksissa sisätulo on 3

25

Vektorien sisätulo

- jos painot eivät ole binäärisiä
  - sisätulo: vastinparien tulojen summa

$$sim(d_i, q_j) = \sum_{k=1}^{t} d_{ik} \cdot q_{jk}$$

  - dokumentti i: (0.9, 0.1, 0.9) ja kysely j: (0.1, 0.8, 0.9)
    - sisätulo: (0.9 · 0.1) + (0.1 · 0.8) + (0.9 · 0.9) = 0.09 + 0.08 + 0.81 = 0.98
  - kysely j': (0.9, 0.2,0.8)
    - sisätulo: (0.9 · 0.9) + (0.1 · 0.2) + (0.9 · 0.8) = 0.81 + 0.02 + 0.72 = 1.55

26

Kosinifunktio

- sisätulo ei anna ylärajaa samanlaisuusarvolle
- usein sisätulo normeerataan vektorien pituuksilla, jolloin funktio kuvaa vektorien välisen kulman kosinia
  - kaksi samanlaista vektoria → kulma on 0°, kosini 1
  - täysin erilaiset vektorit → kulma on 90°, kosini 0
- kosinifunktio:

$$\cos(d_i, q_j) = \frac{\sum_{k=1}^{t} d_{ik} \cdot q_{jk}}{\sqrt{\sum_{k=1}^{t} (d_{ik})^2 \cdot \sum_{k=1}^{t} (q_{jk})^2}}$$

27

Overlap-funktio

- jos dokumentit ovat hyvin pitkiä, saa kosinifunktio pieniä arvoja
  - dokumentin pituus vaikuttaa suoraan nimittäjään
  - koska kysely yleensä on lyhyt, ei osoittaja kasva vastaavasti
  - voidaan määritellä funktio, joka ei heikennä pitkien dokumenttien merkitystä:

$$overlap(d_i, q_j) = \frac{\sum_{k=1}^{t} \min(d_{ik}, q_{jk})}{\min(\sum_{k=1}^{t} d_{ik}, \sum_{k=1}^{t} q_{jk})}$$

28

Vektorimallin merkitys

- vektorimallin etuja
  - käsitteellinen yksinkertaisuus
  - termien painot luontevasti mukana
  - samanlaisuusjärjestys
  - vektoreita on helppo muokata hakuprosessin kuluessa
- vektorimallin ongelmia
  - mallissa oletetaan, että termit ovat riippumattomia, vaikka ne eivät ole
  - samanlaisuusmitat ovat heuristisia: ei ole olemassa teoreettisia perusteluja jonkin tietyn mitan käyttämiseen jossakin tietyssä tilanteessa (tai aina)

29

Tässä osassa

- täystäsmäytys
  - Boolen haku
  - quorum-haku
- osittaistäsmäytys: vektorimalli
  - samanlaisuusmitat: vektorien sisätulo, kosinifunktio, overlap-funktio

30