Tiedonhakumenetelmät

Helena Ahonen-Myka

kevät 2004, osa 3

Indeksointi (1/2), harjoitustyön tehtävänmäärittely

Tässä osassa

- indeksointi
  - kuvaajien tyyppejä
  - indeksoinnin tavoitteet
  - termin frekvenssin vaikutus termin valintaan
  - termin erottelykyvyn vaikutus termin valintaan
- harjoitustyön tehtävänmäärittely

Indeksointi

- dokumenttien kuvaajien muodostaminen
  - yleensä valitsemalla joukko termejä, jotka otetaan mukaan kuvaajaan
- myös hakemiston rakentaminen ja tallettaminen (=hakutietorakenteen implementointi) on osa indeksointia
- indeksi = hakemisto
  - joukko dokumenttien kuvaajia
  - hakutietorakenne
- kyselyiden kuvaajat muodostetaan pääosin samoilla periaatteilla kuin dokumenttien kuvaajat

Kuvaajien tyyppejä

- termien valinta voi olla manuaalista tai automaattista
- termit voivat olla objektiivisia tai subjektiivisia
- termit voidaan valita kontrolloidusta sanastosta tai vapaasti (usein dokumentin tekstistä)
- termit voivat olla yksittäisiä sanoja tai fraaseja (sana + sen konteksti)

Manuaalinen vs. automaattinen

- aikaisemmin indeksointi tehtiin manuaalisesti
  - joko aihepiirin asiantuntijat
  - tai erityisen indeksoijien ammattikunnan edustajat
- nykyisin indeksointi tehdään useimmiten automaattisesti
  - automaattinen indeksointi ei voi koskaan olla täydellistä, mutta myös ihmisindeksoijat tekevät virheitä tai toimivat epäjohdonmukaisesti

Objektiiviset vs. subjektiiviset termit

- objektiiviset termit
  - julkaisun kirjoittaja, julkaisupaikka, sivumäärä yms. bibliografiset tiedot
- subjektiiviset termit
  - dokumentin sisältöä kuvaavat termit

Kontrolloitu vs. vapaa sanasto

- jos indeksointi tehdään manuaalisesti, indeksoijat käyttävät usein yhteisesti sovittua termistöä ja soveltavat termien käytöstä annettuja ohjeita
  - termien valinta yhtenäistyy
  - termejä voidaan hyödyntää myös hauissa
- automaattisessa indeksoinnissa kontrolloitua sanastoa ei ole niin helppo hyödyntää, joten yleensä termeinä käytetään dokumenttien omia sanoja
  - termien valikoima on suurempi
  - myös hauissa voidaan käyttää laajempaa sanastoa

7

Sanat vs. fraasit

- termeinä voidaan käyttää pelkkiä yksittäisiä sanoja
  - kuvaaja on joukko sanoja
  - kukin sana kuvaa pientä osaa sisällöstä
- termeinä voidaan käyttää myös sanaliittoja (fraaseja) tai sanaryhmiä, joissa sanojen väliset suhteet tunnetaan
  - monimutkaisten termien löytäminen tai muodostaminen on vaikeampaa kuin yksittäisten termien

8

Indeksoinnin tavoitteet

- indeksin tehokkuutta säätelee kaksi parametria
  - indeksoinnin tyhjentävyys (indexing exhaustivity)
  - termien spesifisyys (term specificity)

9

Indeksoinnin tyhjentävyys

- Kuinka suuressa määrin kaikki dokumentin käsittelemät asiat ja aihepiirit kuvataan indeksissä?
- kun indeksi on tyhjentävä, kutakin dokumenttia kohti on yleensä valittu paljon termejä ja pienimmätkin alijuonet on kuvattu
- kun indeksi ei ole tyhjentävä, indeksi kuvaa dokumenteista vain yleisimmät pääpiirteet (aihepiirin yms.)

10

Termien spesifisyys

- Kuinka laaja-alaisia tai kapea-alaisia termit ovat merkitykseltään?
- kun merkitykseltään laaja-alaisia termejä käytetään indeksoinnissa, suuri määrä käyttäjää kiinnostavia dokumentteja palautetaan haussa, mutta samalla myös mahdollisesti roskaa
  - → laaja-alaiset termit eivät pysty erottelemaan relevantteja ja epärelevantteja dokumentteja toisistaan
- kapea-alaiset termit palauttavat vähemmän dokumentteja, mutta useimmat niistä ovat todennäköisesti relevantteja

11

Saanti ja tarkkuus

- kun indeksin sanasto (= indeksiin valitut sanat) on kapea-alaista, tarkkuutta suositaan saannin kustannuksella
  - monet haitalliset termit on jätetty huomiotta, mutta samalla myös monet hyödylliset
- kun indeksin sanasto on laaja-alaista, suositaan puolestaan saantia tarkkuuden kustannuksella
- yleensä käyttäjä toivoo sekä kohtuullista saantia että kohtuullista tarkkuutta

12

Saanti ja tarkkuus

- indeksoinnin tyhjentävyys voi vaikuttaa sekä saantiin että tarkkuuteen samansuuntaisesti
- jos indeksointi ei ole tyhjentävä, saanti kärsii: jos kiinnostuksen kohdetta ei ole ollenkaan kuvattu indeksissä, dokumenttia ei löydetä
- tarkkuus voi kärsiä, jos indeksiin valitut (harvat) termit ovat laaja-alaisia eivätkä näin erottele relevantteja ja epärelevantteja dokumentteja toisistaan

13

Termin frekvenssi valintakriteerinä

- tavoitteena on valita indeksitermejä, joiden avulla pystytään erottelemaan relevantteja ja epärelevantteja dokumentteja toisistaan
- mutta indeksiä luotaessa ei vielä tiedetä, mitä hakuja käyttäjä tulee tekemään (eikä relevanssiarvioita)
- ei siis tiedetä, mikä tekee termistä relevantin tai epärelevantin
- voidaan kumminkin tutkia termien esiintymisfrekvenssiä kokoelmassa ja hyödyntää tätä tietoa

14

Termin frekvenssi valintkriteerinä

- kielessä on joukko sanoja, jotka esiintyvät hyvin usein ja tasaisesti kaikissa dokumenteissa
  - suomessa: ei, ja, on, se, että, …
  - englannissa: in, of, and, it,…
- näillä sanoilla on yleensä jonkinlainen toiminnallinen rooli, mutta ne eivät varsinaisesti kuvaa tekstin sisältöä
- usein tällaisista sanoista kootaan hukkasanalista (stopword list)

15

Termin frekvenssi valintakriteerinä

- muut kuin hukkasanalistan sanat kuvaavat paremmin dokumentin sisältöä
- nämä sanat eivät ole yleensä tasaisesti jakautuneita dokumenttikokoelmassa
- sanojen esiintymisfrekvenssiä voidaan käyttää valittaessa termejä
- idea: jos termi esiintyy usein dokumentissa, tämä termi kuvaa dokumentin keskeistä sisältöä

16

Termin frekvenssi valintkriteerinä

- mahdollinen indeksointimenetelmä
  - Poista dokumenteista hukkasanalistan sanat
  - Laske frekvenssi tf_{ij} kaikille jäljellejääneille termeille T_j jokaisessa dokumentissa D_i:
    - tf_{ij} = kuinka monta kertaa termi T_j esiintyy dokumentissa D_i
  - Valitse frekvenssille kynnysarvo K, ja valitse kunkin dokumentin D_i kuvaajaan kaikki termit T_j, joille $tf_{ij} > K$

17

Termin frekvenssi valintakriteerinä

- jos otetaan huomioon vain termin esiintymien lukumäärä dokumentissa, suositaan saantia tarkkuuden kustannuksella
- ajatellaan, että termi ”omena” esiintyy joukossa dokumentteja riittävän monta (> K) kertaa
  - nämä dokumentit varmaan kertovat omenoista
  - jos käyttäjän haku sisältää termin ”omena”, löytyvät nämä dokumentit helpommin, jos ”omena” on valittu termiksi
- entä jos dokumenttikokoelma sisältää pelkästään omenoiden viljelyyn liittyviä dokumentteja?

18

Käänteinen dokumenttifrekvenssi

- tuloksen tarkkuus paranee, jos indeksissä on termejä, jotka esiintyvät vain pienessä osassa dokumentteja
  - nämä termit erottelevat tehokkaasti tämän pienen dokumenttijoukon muista dokumenteista
- olkoon dokumenttifrekvenssi df_j niiden dokumenttien lukumäärä, joissa termi T_j esiintyy (vähintään kerran)

19

Käänteinen dokumenttifrekvenssi

- termin erottelukykyä kuvaa käänteinen dokumenttifrekvenssi (inverse document frequency, idf)
- idf voidaan laskea monella tavalla
- yleinen tapa:

$$idf(T_j) = \log \frac{N}{df_j}$$

- missä N on dokumenttien kokonaismäärä

20

Termin paino dokumentissa

- sekä saantia että tarkkuutta voidaan parantaa, kun otetaan huomioon sekä termin frekvenssi dokumentin sisällä (tf) että termin esiintymien jakautuminen eri dokumenttien välillä (idf)
- dokumentin kuvaajaan kannattaa ottaa termi, joka esiintyy yleisesti tässä dokumentissa, mutta harvoin muissa dokumenteissa
- termin T_j paino w_{ij} dokumentissa D_i:

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j}$$

21

Termin paino dokumentissa

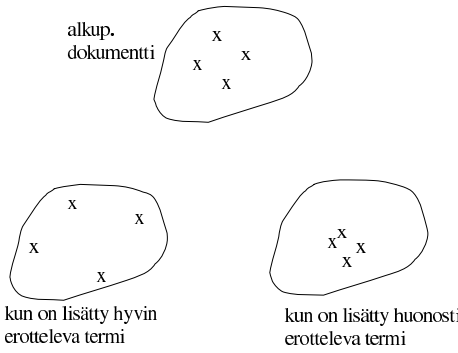
- paranneltu indeksointimenetelmä
  - Poista dokumenteista hukkasanalistan sanat
  - Laske (tf·idf)-paino w_{ij} kaikille termeille T_j kaikissa dokumenteissa D_i
  - Valitse dokumentin kuvaajaan termit, joiden paino w_{ij} ylittää kynnysarvon K´

22

Termin erottelukyky valintakriteerinä

- hyvä erottelukyky on siis toivottava ominaisuus termeille
- seuraavalla kalvolla kukin ‘x’ kuvaa yhtä dokumenttia ja ‘x’:ien etäisyys dokumenttien termien erilaisuutta
  - lähellä olevat ‘x’:t ovat samanlaisempia kuin kaukana olevat
- termin erottelukykyä voidaan mitata vertailemalla tilannetta, kun termi on valittu kuvaamaan dokumentteja, tilanteeseen, kun termiä ei ole valittu

23



24

Termin erottelukyky

- kun lisätään huonosti erotteleva termi, esim. hyvin yleinen sana, tulevat dokumentit samanlaisemmiksi
  - dokumenttien keskimääräinen etäisyys pienenee (kokoelman tiheys kasvaa)
- kun lisätään hyvin erotteleva, yleensä suhteellisen harvinainen termi, muuttuvat ne dokumentit, joissa termi esiintyy, erilaisemmiksi kuin muut
  - dokumenttien keskimääräinen etäisyys kasvaa (kokoelman tiheys vähenee)

25

Termin erottelukyky

- termin T_j erotteluarvo dv_j lasketaan dokumenttikokoelman tiheyden muutoksena kun termi lisätään dokumenttien kuvaajiin

$$dv_j = Q - Q_j$$

- Q: tiheys ennen termin lisäystä
- Q_j: tiheys termin lisäyksen jälkeen

26

Kokoelman tiheys

- kokoelman tiheys voidaan laskea seuraavalla kaavalla:

$$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} sim(D_i, D_k)$$

- lasketaan kaikkien dokumenttiparien keskimääräinen samanlaisuus
- sim() on jokin samanlaisuusfunktio, joka perustuu dokumenttien termeihin

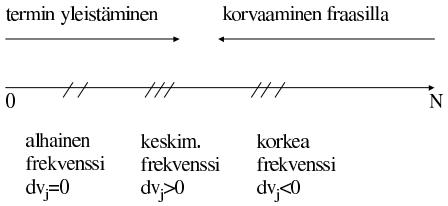
27

Termin erotteluarvo vs. termin dokumenttifrekvenssi

- termi, jolla on hyvä erottelykyky, saa positiivisen erotteluarvon
  - keskimääräinen dokumenttienvälinen samanlaisuus on pienempi, kun termi on lisätty
  - yleensä tällaisia ovat termit, jotka esiintyvät keskimääräisen usein
- usein esiintyvät termit saavat negatiivisen arvon
- hyvin harvinaiset termit eivät juurikaan vaikuta keskimääräiseen dokumenttienväliseen samanlaisuuteen → erotteluarvo lähenee nollaa

28

Termin erotteluarvo vs. termin dokumenttifrekvenssi



29

Termin erotteluarvo vs. idf

- termin idf pienenee, kun termin frekvenssi kasvaa
- termin erottelukyky kasvaa ensin, kun termin frekvenssi kasvaa (alhainen frekvenssi → keskimääräinen frekvenssi), mutta laskee, kun termin frekvenssi kasvaa edelleen (keskimääräinen → korkea)
- termin painotus erottelukyvyllä idf:n sijaan käyttäytyy eri tavalla

$$w_{ij} = tf_{ij} \cdot dv_j$$

30

Termien muokkaus erotteluarvon perusteella

- dokumenttikokoelman tiheyttä voidaan vähentää, jos termejä muokataan erotteluarvon perusteella
- harvinaiset sanat
  - korvataan yleisemmällä termillä (tesauruksesta)
- liian yleiset sanat
  - yhdistetään jonkun toisen sanan kanssa fraasiksi

31

Tiedonhakumenetelmät, harjoitustyö

- harjoitustyö tehdään 4-5 hengen ryhmissä
- ryhmä sopii (löyhästi) jostakin aihepiiristä
- kukin ryhmän jäsen kerää 10 aihepiiriin liittyvää dokumenttia esim. webistä
  - dokumenttien kannattaisi ehkä olla englanninkielisiä
- dokumentit talletetaan mg-hakukoneella, jolloin voidaan käyttää mg:n omaa kyselyliittymää

32

- kukin ryhmän jäsen keksii 2 hakutehtävää
- kustakin hakutehtävästä muotoillaan kaksi kyselyä:
  - Boolen lauseke (vastauksena dokumentit, jotka täyttävät lausekkeen kuvaamat ehdot)
  - joukko termejä (vastauksena järjestetty lista hakutuloksia)

33

- ryhmä arvioi kokoelman dokumenttien relevanssin hakujen suhteen
  - kullekin haulle vähintään 3 riippumatonta arviota
- ajetaan kyselyt mg-kyselyliittymällä
- kaikille hakutuloksille lasketaan saanti ja tarkkuus
  - kun tuloksena on järjestetty lista, lasketaan ja piirretään saanti-tarkkuus –käyrä (keskiarvokäyrä)

34

Raportti

- dokumenttikokoelman kuvaus, esim.
  - dokumenttien määrä ja aihepiiri
  - sanojen määrä yhteensä (saa mg:n tilastoista)
  - dokumenttien keskimääräinen pituus
- hakutehtävät ja kyselyt
- kokemukset relevanssiarvioiden antamisesta: olivatko arvioijat samaa mieltä?
- hakutulosten esittely (lukumäärät, saanti, tarkkuus…)
- luonnehdinta eroista tulosten käyttökelpoisuudessa eri kyselytyypeillä

35