Tiedonhakumenetelmät

Helena Ahonen-Myka
Kevät 2004, osa 10
Likimääräinen täsmäytys: n-grammit
Tallennuksen ja haun tasoperiaate
Kyselyjen muokkaus

Tässä osassa

- likimääräinen täsmäytys
  - n-grammit
  - s-grammit
- tallennuksen ja haun tasoperiaate
  - käsitetaso
  - ilmaisutaso
  - merkkijonotaso
- kyselyjen muokkaus
  - laajentaminen ja kaventaminen

Likimääräinen täsmäytys

- kurssin edellisissä osissa (indeksointi, täsmäytys vektorimallissa, merkkijonohahmon etsintä tekstistä) on oletettu, että haetaan kokonaisia sanoja tai sanavartaloita
  - esim. hakusana täsmää indeksitermin kanssa, kun merkkijonot ovat täsmälleen samat
- on kuitenkin myös tilanteita, joissa sanan tai sen vartalon oikeasta kirjoitusasusta ei ole varmuutta

Likimääräinen täsmäytys

- tekstissä (hakusanoissa, indeksitermeissä) voi olla kirjoitusvirheitä
- vierasperäisten sanojen, esim. erisnimien, esiintymät voivat olla kirjoitusasultaan vaihtelevia
  - esim. Peking, Beijing
  - näille sanoille ei myöskään aina löydetä oikeaa perusmuotoa → jokainen taivutusmuoto on erilainen merkkijono
- kieltenvälisessä tiedonhaussa sanojen erikielisiä vastineita etsitään sanakirjasta
  - monet erisnimet ja tekniset termit eivät löydy sanakirjoista

Likimääräinen täsmäytys: n-grammit

- sanoja voidaan täsmäyttää myös likimääräisillä menetelmillä, mm. n-grammi –menetelmällä
- n-grammi: sanan n-kirjaiminen osamerkkijono
  - tyypillisesti n= 2 (digrammi) tai n=3 (trigrammi)
  - useimmiten peräkkäiset kirjaimet
- ”computer”
  - digrammit: *c, co, mp, pu, ut, te, er, r*
  - trigrammit: *co, com, omp, mpu, put, ute, ter, er*

Likimääräinen täsmäytys: n-grammit

- ”computer”
  - N1: *c, co, om, mp, pu, ut, te, er, r*
- ”compuetr”
  - N2: *c, co, om, mp, pu, ue, et, tr, r*
- samankaltaisuus voidaan laskea esim. kaavalla

$$SIM(N_1, N_2) = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}$$

- 6/12 = 0.5

Likimääräinen täsmäytys: n-grammit

- trigrammit:
  - computer: *co, com, omp, mpu, put, ute, ter, er*
  - compuetr: *co, com, omp, mpu, pue, uet, etr, tr*
- 4/12 = 0.3

7

s-grammit

- n-grammien merkkien ei välttämättä tarvitse olla vierekkäisiä
- Pirkola et al esittelevät s-grammit (skip grams)
  - tarkastellaan (lähinnä) digrammeja
  - oletetaan, että tarkastellaan sanaa w
  - CCI (Character combination index) kuvaa ylihypättävien merkkien määrää
  - CCI=(1,2) viittaa sanasta w muodostettuihin s-digrammeihin, joissa s-digrammin merkit ovat w:ssä yhden tai kahden merkin päässä toisistaan

8

Esimerkki s-digrammiluokista

Sana	CCI	s-digrammit
pharmacology	(0)	{ph, ha, ar, rm, ma, ac, co, ol, ...}
	(1)	{pa, hr, am, ra, mc, ao, cl, oo,...}
	(2)	{pr, hm, aa, rc, mo, al, co, og, ly}
farmakologian	(0)	{fa, ar,rm, ma, ak, ko, ol, lo,...}
	(1)	{fr, am, ra, mk, ao, kl, oo, lg,...}
	(2)	{fm, aa, rk, mo, al, ko, og, li, oa, gn}

9

s-digrammit

- CCI:n eli käytettävien digrammikombinaatioiden valinta on tärkeää täsmäytyksen onnistumisen kannalta
- jos otetaan mukaan kaikki digrammit (CCI = (0,1,2,...,m-2)), missä m sanan pituus, joukossa on todennäköisesti myös hyvin yleisiä digrammeja, jotka esiintyvät monissa sanoissa
- jos otetaan vain esim. peräkkäiset, monet samanlaisetkaan sanat eivät täsmää

10

s-digrammiluokat

- täsmäytystä voidaan tarkentaa s-digrammiluokkien avulla
- luokittelemattomat digrammit
  - CCI muotoa (i, i+1,…,i+j), i, j ≥0
  - kaikkia sanan w1 CCI:n mukaisia digrammeja verrataan sanan w2 CCI:n mukaisiin digrammeihin
- luokkien hyödyntäminen täsmäytyksessä
  - luokat määritellään CCI:ssä, esim. ([0], [1,2])
    - luokka [0] ja luokka [1,2]
  - täsmäytykset tehdään vain saman luokan digrammien kesken

11

Esimerkkejä luokitelluista ja luokittelemattomista s-digrammeista

Sana	CCI	s-digrammit
abcde	(0)	{ab,bc,cd,de}
	(0,1)	{ab,ac,bc,bd,cd,ce,de}
	(0,1,2)	{ab,ac,ad,bc,bd,be,cd,ce,de}
	([0],[1])	{ab,bc,cd,de} {ac,bd,ce}
	([0],[1,2])	{ab,bc,cd,de} {ac,ad,bd,be,ce}
abce	(0,1)	{ab,ac,bc,be,ce}
	([0],[1,2])	{ab,bc,ce}{ac,ae,be}

12

Tuloksia

- tutkimuksessa verrattiin englannin-, saksan- ja ruotsinkielisiä lääketieteellisiä termejä ja paikannimiä suomenkielisiin
- luokiteltujen s-digrammien käyttö antoi paremman tuloksen kuin n-digrammien käyttö (jossa ei sallita merkkien ylihyppäyksiä)
- myös luokittelemattomat s-digrammit toimivat paremmin kuin n-digrammit
- erityisesti s-grammit parantavat tuloksia, kun sanat ovat lyhyitä
  - kun sanat ovat hyvin lyhyitä, mikään menetelmä ei toimi hyvin

13

Tallennuksen ja haun tasoperiaate

- hakutehtävät ja dokumentit voidaan esittää kolmella tasolla
  - käsitetaso (conceptual level)
  - ilmaisutaso (expression level)
  - merkkijonotaso (occurrence level)

14

Käsitetaso

- käsitetasolla tarkastellaan hakutehtävän ja dokumentin käsitteitä ja niiden suhteita
- otettava huomioon
  - hakutehtävän tarkoitus
  - käytettävän tietokannan sisältämien relevanttien dokumenttien tekstien edustama käsitteistö
    - myös relevanttien dokumenttien käsitteistö voi poiketa toisistaan
- "parhaimmillaan" ihmisen suorittama analyysi, mutta nykyisin usein käsitetaso tietoisena vaiheena ohitetaan sekä indeksoinnissa että hakuvaiheessa

15

Ilmaisutaso

- ilmaisutasolla tarkastellaan käsitteiden ilmaisutapoja luonnollisessa kielessä (tai jossain erikoiskielessä)
- dokumentin ilmaisutason esitys on teksti itse: dokumentin sanat + mahdollisesti muita termejä
- hakutehtävä: etsittävä vaihtoehtoisia ilmauksia käsitteille

16

Ilmaisutaso

- käsitetasolta ilmaisutasolle:
  - synonyymit, sanaliitot
  - yleiskielen nimitykset, kaupalliset nimitykset, tieteelliset nimitykset
  - nykyiset, suositeltavat, kartettavat (vanhentuneet) nimitykset
  - kirjoitusasuvaihtoehdot
  - lyhenteet ja täydelliset nimitykset
  - ositetut ja yhdistetyt nimitysvaihtoehdot ("tiedon tarve", "tiedontarve")
  - kapea-alaisemmat ja laaja-alaisemmat nimitykset

17

Merkkijonotaso

- konkreettinen tiedonhaku tapahtuu aina merkkijonotasolla
  - perusoperaatioita: merkkijonojen samanlaisuus tai samanlaisuuden aste, merkkijonojen esiintymien keskinäinen sijainti, esiintymien lukumäärä yms
- kyselyn muotoilussa ilmaisutason hakuavaimet käännetään merkkijonotason merkkijonoiksi
  - esim. yhdistely operaattoreilla, katkaisu, säännölliset lausekkeet, n-grammit
- tiedonhakija voi ohittaa käsitetason ja ilmaisutason ja antaa suoraan kyselyn

18

Hakutehtävän käsiteanalyysi

- käsiteanalyysin tarkoitus on tunnistaa hakuaiheen keskeiset käsitteet ja käsitteiden suhteet
- tuloksena on käsitteellinen hakusuunnitelma, jonka pohjalta toteutetaan kysely
  - käsitteellinen hakusuunnitelma: mitä tietoa haetaan
  - kysely: miten tietoa haetaan
- käsiteanalyysi yksinkertaistaa tarpeellisten käsitteellisten suhteiden tunnistamista
  - muutaman käsitteen tarkastelu on paremmin hallittavissa kuin suuren hakuavainjoukon tarkastelu

19

Hakutehtävän käsiteanalyysi

- käsitteet jäsennetään rinnakkaisiin ja rajaaviin suhteisiin
  - rinnakkaiset suhteet: käsitteet edustavat hakuaiheen samaa aspektia (näkökulmaa)
  - rajaavat suhteet: käsitteet edustavat eri aspekteja
- suhteet usein assosiaatiosuhteita
  - tekijä – teko – tekeminen – väline
  - teko – kohde – tulos
  - syy – seuraus

20

Hakutehtävän käsiteanalyysi

- assosiaatiosuhteen olemassaolo ei riitä ratkaisemaan, kuuluvatko kaksi käsitettä samaan aspektiin
- ratkaisu riippuu siitä, miten tiedontarvitsija liittää käsitteitä yhteen
  - "metsä- ja karjatalouden tuottavuus ja tukimuodot"
    - "tuottavuus" ja "tukimuodot" edustavat samaa aspektia
  - "metsä- ja karjatalouden tukimuotojen vaikutus tuottavuuteen 1980-luvulla"
    - "tuottavuus" ja "tukimuodot" edustavat eri aspekteja

21

Hakutehtävän käsiteanalyysi

- rinnakkaiset suhteet johtavat Boolen kyselyssä disjunktion (OR) käyttöön käsitteitä edustavien ilmaisujen välillä
- rajaavat suhteet johtavat konjunktion (AND) tai negaation (NOT) käyttöön
- "kalojen tuottamat myrkyt"
  - myrkyt AND kalat NOT (kalatalous AND vesiensuojelu)

22

Hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus

- tyhjentävyys
  - kuinka moni hakuaihetta jäsentävistä aspekteista on mukana hakusuunnitelmassa
  - aspekti on edustettuna hakusuunnitelmassa, jos suunnitelma sisältää jonkin aspektiin kuuluvan käsitteen
  - tyhjentävyys liittyy aina hakuaiheen aspektien välisiin rajaaviin suhteisiin

23

Hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus

- tarkkuus
  - kuinka täsmällisillä käsitteillä hakusuunnitelman eri aspekteja suunnitelmassa kuvataan
  - hakusuunnitelma on täydellisen tarkka, jos aspektit esitetään täsmälleen hakuaiheen tarkkuustasolla, ja muulloin epätarkka
  - tarkkuus liittyy aina käsitteiden hierarkkisiin suhteisiin

24

Hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus

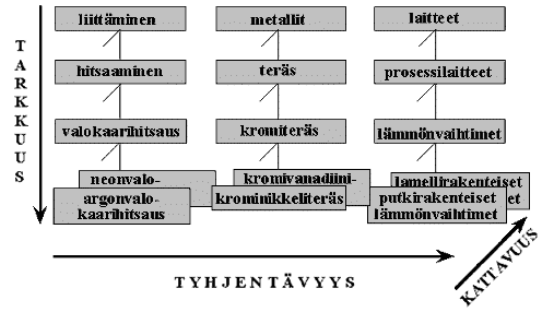
- kattavuus
  - kuinka monella tiedontarpeeseen sisältyvällä käsitteellä hakusuunnitelman eri aspekteja suunnitelmassa kuvataan
  - hakusuunnitelma on kattava, jos se sisältää kaikki kunkin aspektin ulottuvuudet ja muulloin epäkattava
  - kattavuus liittyy aina hakuaiheen aspektien sisäisiin rinnakkaisiin suhteisiin

25

Hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus

- seuraavan kalvon kuvassa esitetään kolme aspektia, joiden käsitehierarkioiden ylimpinä käsitteinä ovat ”liittäminen”, ”metallit” ja ”laitteet”
- oletetaan, että hakuaiheen täydellinen kuvaus koostuu täsmälleen käsitteistä ”argonvalokaarihitsaus”, ”neonvalokaarihitsaus”, ”kromimikkeliteräs”, ”putkirakenteiset lämmönvaihtimet” ja ”lamellirakenteiset lämmönvaihtimet”
  - ”argonvalokaarihitsaus” ja ”neonvalokaarihitsaus” ovat rinnakkaisia käsitteitä
  - ”putkirakenteiset lämmönvaihtimet” ja ”lamellirakenteiset lämmönvaihtimet” ovat rinnakkaisia

26



27

Hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus

- hakusuunnitelma on täydellisen tyhjentävä, jos se sisältää jokaisen tiedontarpeen aspektin
  - esim. “valokaarihitsaus”, “teräs” ja “lämmönvaihtimet”
- jos hakusuunnitelmassa käytetään käsitteitä “argonvalokaarihitsaus” ja “lamellirakenteiset lämmönvaihtimet”, on suunnitelma näiden aspektien suhteen täydellisen tarkka
  - jos käytetään käsitteitä “liittäminen” ja “laitteet”, on suunnitelma epätarkka (näiden aspektien suhteen)
- suunnitelma on laite-aspektin suhteen kattava, jos se sisältää käsitteet “putkirakenteiset lämmönvaihtimet” ja “lamellirakenteiset lämmönvaihtimet”

28

Käsitteiden muuntelu täydellisessä täsmäytyksessä

- hakusuunnitelman tyhjentävyyden lisääminen kaventaa hakua
  - tarkoittaa uusien rajaavien käsitteiden lisäämistä
  - tuloksen saanti kärsii, tarkkuus paranee ja koko vähenee
- tarkkuuden lisääminen kaventaa hakua
  - tuloksen saanti kärsii, tarkkuus paranee ja koko vähenee
- kattavuuden lisääminen laajentaa hakua
  - tarkoittaa ko. aspektiin liittyvien rinnakkaisten käsitteiden lisäämistä
  - tuloksen saanti paranee, tarkkuus kärsii ja koko vähenee

29

Käsitteiden muuntelu osittaistäsmäytyksessä

- osittaistäsmäytystä käytettäessä ei käsitteillä ole rinnakkaisia tai rajaavia suhteita
- oleellisen käsitteen pois jättäminen hakusuunnitelmasta → dokumentit eivät saa ”lisäpisteitä” ko. käsitettä edustavista hakuavaimista
  - oleellinen dokumentti voi jäädä hakutuloksen kynnysarvon alapuolelle

30

Käsitteiden muuntelu osittaistäsmäytyksessä

- kyselyyn voidaan liittää kaikkia tiedontarpeen ulottuvuuksia kuvaavia käsitteitä (rajaavia, rinnakkaisia, eri hierarkiatasoilta)
  - hakusuunnitelmasta ei silti tule liian tyhjentävä
- dokumentit saavat lisäpisteitä kaikista niistä käsitteistä, joita edustavia hakuavaimia niistä löytyy
  - relevantit hakuavainkombinaatiot voivat vaihdella relevantista dokumentista toiseen
  - käyttäjän ei tarvitse huolehtia sopivista kombinaatioista
- monien osittaistäsmäyttävien järjestelmien kyselykielissä on Boolen operaatioiden vastineita
  - mm. "kaikkien hakuavainten pitää esiintyä"

31

Kyselyjen muotoilu

- ensimmäistä kyselyä on useimmiten muokattava, jotta päästäisiin hyvään hakutulokseen
- hakuavaimia voidaan lisätä, poistaa tai vaihtaa
- hakuavainten välisiä operaattoreita tai painotusta voidaan muuttaa
- kyselyn laajentaminen yleisintä

32

Kyselyn laajentaminen

- voidaan soveltaa sekä täys- että osittaistäsmäytyksessä
- suoritustapa:
  - tiedonhakija laajentaa itse
  - järjestelmä laajentaa automaattisesti
  - laajennus vuorovaikutteisesti: järjestelmä ehdottaa hakuavaimia, tiedonhakija valitsee
- laajennosten lähteenä voivat olla hakutulokset tai rakenteiset kokoelmat (esim. tesaurukset)

33

Kyselyjen muotoilu täystäsmäytyksessä

- kyselyn suorituksessa voi tulla vastaan kaksi tilannetta
  - kysely tuottaa liian vähän dokumentteja
  - kysely tuottaa liikaa dokumentteja
- → kyselyä pitää laajentaa tai kaventaa
- tilanteisiin voidaan valmistautua jo käsitteellistä hakusuunnitelmaa tehtäessä
  - laajentamista ja kaventamista voidaan tehdä muuttamalla tyhjentävyyttä, tarkkuutta ja kattavuutta

34

Kyselyjen kaventaminen täystäsmäytyksessä: keinoja

- parannetaan kyselyn tyhjentävyyttä lisäämällä rajaavia käsitteitä edustavia hakuavaimia konjunktion tai läheisyysoperaattorin avulla
- vähennetään kyselyn käsitteitä edustavia vaihtoehtoisia hakuavaimia (= vähennetään disjunktioita)
- parannetaan tarkkuutta käyttämällä tarkkoja hakusanoja (tai luokitustermejä)
- tarkastetaan, ettei hakuavainten joukossa ole monimerkityksisiä sanoja tai lyhenteitä
- muutetaan konjunktioita läheisyysoperaattoreiksi

35

Kyselyjen laajentaminen täystäsmäytyksessä: keinoja

- lisätään kyselyyn rinnakkaisia käsitteitä lisäämällä uusia käsitteitä edustavia hakuavaimia disjunktion avulla
- vähennetään kyselyn tyhjentävyyttä poistamalla rajaavia käsitteitä edustavia hakuavaimia
- lisätään kyselyyn vaihtoehtoisia hakuavaimia (ilmaisutasolla)
- tarkastetaan, ettei negaatio ole epäselvä tai monitulkintainen

36

Kyselyjen laajentaminen täystäsmäytyksessä: keinoja

- vähennetään tarkkuutta vähentämällä tarkkoja luokituskoodeja tai hakusanoja (ja korvaamalla nämä yleisemmillä)
- korvataan hakusanoja säännöllisillä lausekkeilla, jotka täsmäävät useampiin sanoihin
- muutetaan läheisyysoperaattoreita konjunktioiksi

37

Kyselyjen muotoilu osittaistäsmäytyksessä

- täystäsmäytyksessä käytettävät kyselyjen muotoilukeinot eivät toimi osittaistäsmäytyksessä
  - hakusanan lisääminen kyselyyn ei pienennä tulosjoukkoa, mutta saattaa parantaa tuloksen tarkkuutta
- tyypillisiä muotoilukeinoja: hakuavainten lisääminen, poisto tai uudelleenpainotus relevanssipalautteen perusteella

38

Kyselyjen muotoilu osittaistäsmäytyksessä

- kyselyn automaattinen laajentaminen voi perustua
  - kokoelmasta riippumattomiin sanastoihin (esim. tesauruksiin), tai
  - kokoelmasta koostettuihin assosiaatiosanastoihin
- laajennoslähteiden tehokkuudesta on saatu ristiriitaisia tutkimustuloksia
  - pitemmät haut tuottavat parempia tuloksia kuin lyhyet → laajentaminen kannattavaa
  - lisättävien hakuavainten on esiinnyttävä kokoelmassa → ulkopuoliset kokoelmat eivät välttämättä tuota hyviä tuloksia

39

Tässä osassa

- likimääräinen täsmäytys: n-grammit, s-grammit
- tallennuksen ja haun tasoperiaate: käsitetaso, ilmaisutaso, merkkijonotaso
  - hakutehtävän käsiteanalyysi
  - hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus
  - käsitteiden muotoilu täys- ja osittaistäsmäytyksessä
  - kyselyjen muotoilu täys- ja osittaistäsmäytyksessä

40