4. Dokumenttien indeksointi

= sisällön kuvaus kyselyjä varten
- content descriptor, profile (surrogate)
- myös dokumenttien suhteet selvenevät

Tähän asti:
- käänteistiedosto voi olla täydellinen indeksi termin tai jopa termin esiintymä(kohda)n tarkkuudella
- kaikki termit eivät ole tiedonhaun kannalta yhtä hyviä, yhtä kuvaavia
termin paino: esiintymiskertojen lkm (likiarvo ...)

Millainen olisi hyvä indeksointi?
- semantiikka, toteutus

Indeksoinnin päävaihtoehdot:

- manuaalinen
  - asiantuntija tekee
    aihe (semantiikka vaikuttaa)
    indeksoinnin tekniikka
  - kirjoittaja itse, informaatikko
  - työläs
  - yhtenäinen vain, jos kiinteä asiasanasto (jos silloinkaan)

- automaattinen
  - tilastolliset menetelmät
  - lingvistiset menetelmät

Nykyiset volyymit 'vaativat' automatisointia.
- 'riittävän hyvä' tulos (hakutulokset!)
- semantiikan approksimointi esim. sanojen frekvensseillä (~)

Yleisiä näkökohtia / tavoitteita:

- kyselyjen tuki, ei itsetarkoitus

- objektiiviset kuvaukset
  - tekijä, julkaisija, aika, sarja, media, …
  - sanojen esiintymiin perustuvat

- ei-objektiiviset kuvaukset
  - sisältö: teksti, informaatio

- yksittäinen termi / termi jossain kontekstissa
  - todellinen yhteys usein monimutkainen
    esim. sisällöllinen fraasi /
    (muuten) peräkkäin esiintyminen

- rajoitettu / vapaa sanasto
  vrt. ACM:n asiasanat

- tietokannan luonne: onko koko teksti vai vain kuvaavat metatiedot (viitetietokanta, abstraktitietokanta)

ACM Classification system:

General terms: algorithms, design, documentation, economics, experimentation, human factors, languages, legal aspects, management, measurement, performance, reliability, security, standardization, theory, verification.

- yleisiä; käytetäänkö kyselyissä ?

Subject descriptors alaluokasta '*Contents analysis and indexing*': abstracting methods, dictionaries, indexing methods, linguistic processing, thesauruses

- aihekohtaisia, kuvaavia

(Additional) keywords and phrases eli kirjoittajan itse valitsemat lienevät kuvaavimpia. Näiden valinta on epätriviaali ongelma: yksilöllisyys / yhteinen 'kieli'!
(vrt. otsikon valinta ...)

4.1 Indeksoinnin semantiikka

1° termien riippuvuudet

synonyymit: täydellinen
osittainen riippuvuus [0,1]
– teoriassa, määrittely vaikeaa

esim. lemmikki(eläin), koira, kissa
vuodenaika, kesä, heinäkuu, 12.7.2000

- muut yhteydet: antonyymit (~vastakohdat), homonyymit (hankalia)
is-a - ja part-of -suhteet jne,
yleisesti käsiteverkot (semanttiset verkot)

- yhteyksien perinteinen ilmaisumuoto:
thesaurus (asiasanaluettelo)

perustermi + siihen yhdistyvät termit
(+ termin käyttötapoja …)

NT narrower term
BT broader term
RT related term
TT top term
UF used for
CL classification
L-SP Spanish
L-FR French …

Esim.
temperature
BT climate
RT pressure
NT Celsius, Kelvin, absolute zero
CL physics

absolute temperature scale (fraasi)
BT temperature scale
RT Celsius temperature scale
Kelvin
temperature (vai BT, TT?)
NT absolute zero point

Muita apuneuvoja:

· selityssanasto (sanoille, lyhenteille)

Esim. *ATM* *Automated Teller Machine*
Asychronous Transfer Mode
Anti Tank Missile

- ei yhteistä semantiikkaa
- tarpeen erityisesti homonyymeille

2° Sananmuotojen huomiointi

· lauserakenteet
substantiivit vs. muut

· muotojen yhtenäistäminen (stemming)

1° päätteiden poisto
-ing, -ed, -ly

engineering, engineered → engineer

2° yleinen 'vartalon' muodostus

- kieliopillinen vartalo
- pituuden huomiointi
päätteet pois, mahdollinen katkaisu
- erilaisia menetelmiä
- kielisidonnaista, erityisesti suomi (?)

user, users, used, using, usage → us
vyö, vöitä → ?

marina, marinade, marine, marital → mari ??
equipment, equivalence, equilateral → ?

Porterin algoritmi lienee tunnetuin

(periaate esim. BYRN, Appendix:
http://www.sims.berkeley.edu/~hearst/irbook/porter.html)

3° Fraasit

- yksinkertainen lähtökohta: termi = sana

- kieli > sanojen joukko!

· merkitys ilmaistaan usein monen sanan (termin) muodostamalla fraasilla

Esim. *computer science*
computer performance
computer programming
desktop computer

Fraasien automaattinen muodostus:
- kielen analysointi
- termien ryhmittely
- riippuvuuksien analysointi (tilastollisesti)

Kielen analysointi

- syntaktisia sääntöjä
esim. (adjektiivi, substantiivi)
(substantiivi, substantiivi)
kielestä riippuvia
vierekkäisyys / lähekkäisyys

Esim. *information retrieval =*
retrieval of information
- retrieval of most important information

- automaattiset menetelmät tuottavat myös virheellisiä fraaseja

Esim. *high frequence transistor oscillator*
- mitkä kombinaatiot mielekkäitä fraaseja?

Esim. käytännön tilanteesta:

fraasi *text analysis system*

asian esiintymiä:

system analyzes the text
text is analyzed by the system
system carries out text analysis
text is subjected to analysis by the system

- entä synonyymit:

text: document, information item
analysis: processing, transformation
system: program, process

(lisää fraaseista erotteluarvon yhteydessä)

4.2 Indeksoinnin prosesseista

Perusalgoritmi:

1. etsi kaikki eri sanat ja laske niiden frekvenssit

2. poista hukkasanat (stoplist)

3. poista sanoista päätteet (stemming)

4. laske sanojen (termien) T_j frekvenssit tf_{ij}
   jokaisessa dokumentissa D_i

5. valitse indeksitermeiksi kynnysarvon ylittävät sanat
   $tf_{ij} > k$

Hukkasanat?

- liian yleiset, merkityksettömät
- kontekstisidonnainen:
  yleinen teksti, aihekohtaiset
- lkm vaihtelee: 8, 200, 350, …

Hukkasanalistat

ORBIT Search Service:
- vain 8 sanaa !
  an and by from of the with

Brown corpus:
- "from a broad range of literature in English"
- 425 sanaa:

a	about	above	across
after	again	against	all
almost	alone	along	already
also	although	always	among
an	and	another	any
anybody	. . .	area …	case ...
face	fact …	goods	
group…	interest…	member…	work

- täytesanoja ei aina poisteta ollenkaan (erikseen)
  (suuri frekvenssi → jäävät pois sen takia)

Termien painotus

- termillä dokumentissa suuri frekvenssi
  → oikeita dokumentteja tulee mukaan (saanti kasvaa)
- termin frekvenssi suuri useissa dokumenteissa
  → tarkkuus vähenee

  Esim. *computer* tietokonealan dokumenteissa / yleensä

Hyvä indeksitermi:
- esiintyy usein 'hyvissä' dokumenteissa
- esiintyy harvoin muualla

=> tf_{ij}:n tilalle tulee löytää parempi valintakriteeri:

lasketaan frekvenssi kokoelmassa (df_j) ja
yksittäisissä dokumenteissa (tf_{ij})

df_j = termin T_j dokumenttifrekvenssi
N dokumentin joukossa ("kuinka monessa esiintyy")

idf = inverse document frequency
= $\log(N/df_j)$

w_{ij} = termin T_j paino dokumentille D_i
= $tf_{ij} * \log(N/df_j)$

algoritmi: kynnysarvo w_{ij}:lle ...

Toinen perusta termin painolle:
termin erotteluarvo dv_j (discrimination value)

- dokumenttijoukon tiheys

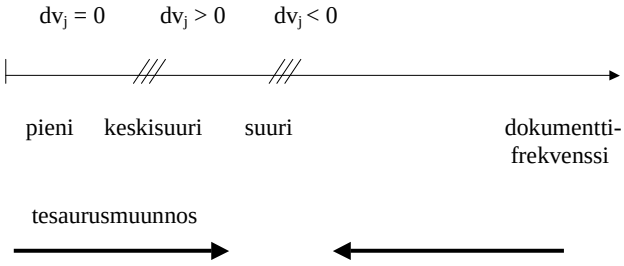
pieni <=> dokumentit kaukana toisistaan
suuri <=> dokumentit lähellä toisiaan

- termin paino sen mukaan, lisääkö se dokumenttien etäisyyttä vai ei ("erotteleeko")

$w_{ij} = tf_{ij} * dv_j$

Merkitään: Q = dokumenttijoukon tiheys
Q_j = joukon tiheys, kun on lisätty termi T_j

Tällöin termin T_j erotteluarvo $dv_j = Q - Q_j$



Tiheyden määritys?

1° dokumenttiparien similariteettien
(samanlaisuusarvojen) keskiarvo:

$Q_j = (1/N(N-1)) \Sigma_i \Sigma_{k \neq i} sim(D_i, D_k)$

- samanlaisuusarvo?
dokumenttien vektoriesityksestä:

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

d_{ij} = 1 tai 0 (termin esiintyminen)

- yleistys: termien painotus

2° dokumentin etäisyys joukon
keskipisteestä (sentroidista)

$C = (c_1, c_2, \ldots, c_t)$

$c_j = (1/N) \Sigma_k d_{kj}$

$Q = (1/N) \Sigma sim(C, D_k)$

Erotteluarvon suhde frekvenssiin?

pieni frekvenssi: $dv_j \approx 0$
suuri frekvenssi: $dv_j < 0$
keskimääräinen frekvenssi: $dv_j > 0$

Erotteluarvo & indeksointi:

- tesaurusyhteydet lisäävät harvinaisen termin dokumenttifrekvenssiä
- fraasi vähentää liian yleisen termin dokumenttifrekvenssiä
- tavoitteena hyvä erotteluarvo

3° termien ryhmittely

- yhdistetään useimmin yhdessä esiintyvät termit (co-occurrence)

termi/dokumentti - matriisi

	T_1	T_2	...	T_3
D_1	d_{11}	d_{12}	...	d_{1t}
D_2	d_{21}	d_{22}	...	d_{2t}
:	:	:		:
D_n	d_{n1}	d_{n2}	...	d_{nt}

d_{ij} = termin T_j merkitys dokumentille D_i

1) sarakkeiden vertailu
=> yhdistetään läheiset termit

2) rivien vertailu
=> yhdistetään dokumentit ryhmiksi,
sitten ryhmää kuvaavat termit

= termien ryvästäminen

single link / complete link: heikkoja / vahvoja

Ongelmia:
- co-occurrence ≠ merkitysyhteys
- kohta 2: yhteyden pätevyys dokumenttiryhmän ulkopuolella ?

4.3 Fraasien muodostaminen

- tekstin perusteella, erotteluarvo

- termin frekvenssi suuri → dv_j pieni
- fraasin dv_j suurempi

- automaattisesti:
1° valitaan fraasin kahva (phrase head)
- riittävä frekvenssi

2° liitetään kahvaan komponentteja
- pieni frekvenssi
- esiintyvät kahvan lähellä

(3° hukkasanat jätetään pois)

Esim.
"Effective retrieval systems are essential for people in need of information."

kahvoja: systems, people, information

vierekkäisyys (2°):

*retrieval systems**	*people need*
systems essential	*need information**
essential people	

yhdessä esiintyminen (lisäksi):

*effective systems effective information**
*systems need retrieval information**
*effective people essential information**
retrieval people
(fraasien laatu?)

- täydennetään menetelmää:

4° syntaktiset muodot, esim.
(adj, subst) (subst, subst)

*retrieval systems**
people need
*need information**

5° lauserakenne
subjektifraasi, verbifraasi, objektifraasi

effective retrieval systems
are essential
people in need of information

4.4 Tesaurusten muodostaminen

- 'oikeat' tesaurukset vaativat termien merkityksen tunnistamisen

apuvälineitä; esim. konkordanssilistan muodostus
- erilaiset esiintymät voivat paljastaa termin käytön eri merkityksissä

- termien yhteyksiä voidaan päätellä myös niiden dokumenttiesiintymistä; tämä tieto voidaan kirjata tesauruksen muotoon

· termien t_j ja t_k samanlaisuutta voidaan mitata termi/dokumenttimatriisista sisätulolla:

$$\text{sim}(t_j, t_k) = \sum_{i=1}^{n} d_{ij} \cdot d_{ik}$$

(vastaavasti esim. kosinikerroin)

· termit ryvästetään termiparien samanlaisuuksiin perustuen
- single link / complete link: heikko / vahva yhteys

Huom. tesauruksen käyttä 'vain' tiedonhaussa, ei merkitystä ilmaisemassa
- pätevyys riippuu silti vahvasti dokumenttijoukosta
- menetelmän yhdistäminen olemassa oleviin termiluokkiin mahdollista

- pseudoluokittelu: käytetään hyväksi tunnettujen dokumenttien ja kyselyjen (otosten) relevanssitietoja

kyselyssä ja siihen nähden relevanteissa dokumenteissa esiintyminen lisää termien samanlaisuutta ja vie ne samaan tesaurusluokkaan

- otosten tulee olla edustavia