Tiedonhakumenetelmät

Helena Ahonen-Myka
Kevät 2004, osa 4
Indeksointi (2/2)

Tässä osassa

- termin muuttaminen kapea-alaisemmaksi:
  - fraasien muodostaminen
- termin muuttaminen laaja-alaisemmaksi:
  - tesauruksen käyttö
- käänteistiedoston muodostaminen ja käyttö

Fraasien muodostaminen

- jos jonolla sanoja (= fraasilla) on jokin merkitys, on tämä merkitys aina kapea-alaisempi kuin jonon yksittäisten sanojen merkitys
  - "computer science" vs. "computer"
- jos dokumentin kuvaajaan lisätään fraaseja, onkin tarkoituksena yleensä kaventaa joidenkin liian laaja-alaisten termien merkitystä
  - tavoite: termit, joiden frekvenssi on korkea, muutetaan termeiksi, joiden frekvenssi on keskimääräinen
  - kahta harvoin esiintyvää termiä ei kannata yhdistää, koska fraasi olisi vielä harvinaisempi

Mahdollinen algoritmi

- valitaan fraasin pääsanaksi termi,
  - jonka dokumenttifrekvenssi ylittää tietyn kynnyksen (esim. df > 5) tai
  - jonka erotteluarvo on negatiivinen
- muiksi fraasin komponenteiksi valitaan harvoin tai keskimääräisen usein esiintyviä termejä, jotka esiintyvät pääsanan yhteydessä (esim. samassa lauseessa riittävän lähellä pääsanaa)
- hukkasanalistan sanoja ei (yleensä) kannata käyttää fraasin osana

Fraasin komponenttien valinta

- pääsanaan liitettävät muut termit voidaan valita monella tavalla
  - siis "esiintyy pääsanan yhteydessä" voidaan tulkita monella tavalla
- tarkastellaan esimerkkilausetta:
  - Effective retrieval systems are essential for people in need of information.

Fraasin komponenttien valinta

- termit "are", "for", "in", ja "of" ovat todennäköisesti hukkasanalistassa, eikä niitä oteta huomioon
- termit "systems", "people" ja "information" ovat todennäköisesti niin yleisiä, että ne sopivat fraasin pääsanaksi
- jos vaaditaan, että pääsana ja toinen komponentti esiintyvät peräkkäin:
  - retrieval systems, systems essential, essential people, people need, need information

Fraasin komponenttien valinta

- jos riittää, että termit ovat samassa lauseessa, saadaan lisäksi fraasit:
  - effective systems, systems need, effective people, retrieval people, effective information, retrieval information, essential information
- termien järjestys siis yleensä kannattaa säilyttää
- voidaan antaa myös lisärajoituksia pääsanalle ja komponenteille:
  - jos tiedetään sanojen sanaluokka, voidaan hyväksyä vain esim. adjektiivi-substantiivi tai substantiivi-substantiivi -pareja

7

Fraasin komponenttien valinta

- jos voidaan jäsentää tekstin syntaktista rakennetta, voidaan vaatia, että fraasin sanat esiintyvät samassa syntaktisessa komponentissa, esim. subjektilausekkeessa, verbilausekkeessa tai objektilausekkeessa
- esimerkkilauseen syntaktiset komponentit:
  - subjektilauseke: effective retrieval systems
  - verbilauseke: are essential
  - objektilauseke: people in need of information
- hyväksyttäisiin fraasit: effective systems, retrieval systems, people need, need information

8

Fraasin komponenttien valinta

- kun rajoitukset ovat tiukempia, tuotetaan luonnollisesti vähemmän fraaseja
- sekä löyhät että tiukat rajoitukset tuottavat sekä hyviä että huonoja fraaseja
- voitaisiin edelleen yrittää selvittää sanojen semanttisia suhteita
  - "high frequency transistor oscillator": high frequency ok, frequency transistor ei
  - voi olla vaikeaa ja työlästä eikä ehkä kuitenkaan paranna tuloksia merkittävästi
- "väärätkin" fraasit voivat auttaa kyselyiden ja dokumenttien vertailussa

9

Fraasien rakenne

- dokumenttien ja kyselyjen vertailun kannalta pitäisi samaa tarkoittavat mutta rakenteeltaan erilaiset fraasit saada yhdistettyä
  - "information retrieval" vs. "retrieval of information"
  - synonyymit, erilaiset sanajärjestykset, täytesanat
- voidaan yrittää normalisoida fraasit johonkin kanoniseen muotoon
- tai muodostaa jokaisesti fraasista myös vaihtoehtoisia muotoja
- molemmat vaihtoehdot ovat käytännössä hankalia

10

Esimerkki

- perusmuoto: text analysis system
- muunnelmia:
  - System analyses the text
  - Text is analyzed by the system
  - System carries out text analysis
  - Text is subjected to analysis by the system
  - Text is subjected to system analysis
- synonyymejä, jotka voisivat korvata termejä:
  - text → documents, information items
  - analysis → processing, transformation, manipulation
  - system → program, process

11

Tesauruksen käyttö

- termi, joka on liian harvinainen, voidaan korvata yleisemmällä termillä
  - yleisempi termi löydetään käsitemallista, tesauruksesta (thesaurus)
- tesaurus ryhmittelee kapea-alaisia termejä luokaksi
  - luokan jäsenten yhteinen esiintymisfrekvenssi on keskimääräistä tasoa
  - esim. refusal, declining, noncompliance, rejection, denial voivat olla samassa luokassa
  - luokan jäsenten esiintymät dokumenteissa voidaan korvata luokan tunnuksella tai yhdellä jäsenellä (esim. refusal)

12

Tesauruksen muodostaminen

- tesaurus voi olla manuaalisesti tai automaattisesti tuotettu
- manuaalisia, esim.
  - WordNet: yleinen englanninkielinen tesaurus
  - alakohtaiset tesaurukset
- manuaalista työtä voidaan tukea automaattisilla menetelmillä, esim. voidaan listata
  - kaikki sanan esiintymät kokoelmassa → sanalla ehkä useita merkityksiä
  - eri termit, jotka esiintyvät samanlaisessa ympäristössä -→ samaan luokkaan sopivat termit

13

Tesauruksen muodostaminen

- automaattisia menetelmiä
  - verrataan termien yhteisesiintymien määrää
  - hyödynnetään joukkoa hakutehtäviä ja niihin liittyviä relevanssiarvioita

14

Dokumentti-termi -matriisi

	T1	T2	...	Tt
D1	w_{11}	w_{12}	...	w_{1t}
D2	w_{21}	w_{22}	...	w_{2t}
.	.			.
.	.			.
.	.			.
Dn	w_{n1}	w_{n2}	...	w_{nt}

15

Termien yhteisesiintymien tarkastelu

- tarkoituksena on löytää termejä, jotka esiintyvät usein yhdessä
- kahden termin läheisyyttä voidaan kuvata esim. seuraavalla samanlaisuusmitalla

$$sim(T_j, T_k) = \sum_{i=1}^{n} w_{ij} \cdot w_{ik}$$

- missä n on dokumenttien määrä
- kun pareittaiset samanlaisuusarvot on laskettu, voidaan klusteroida samaan ryhmään ne termit, jotka ovat keskenään samanlaisia

16

Termien klusterointi

- kaksi tapaa klusteroida
  - termi liitetään klusteriin, jos termin samanlaisuusarvo vähintään yhden klusterin jäsenen kanssa ylittää annetun kynnysarvon
    - tuottaa yleensä vähemmän ja suurempi klustereita
  - termi liitetään klusteriin, jos termin samanlaisuusarvo kaikkien klusterin jäsenten kanssa ylittää annetun kynnysarvon
    - tuottaa yleensä paljon pieniä klustereita

17

Hakujen ja relevanssiarvioiden hyödyntäminen

- oletetaan, että käytettävissä on kokoelma + joukko hakutehtäviä + relevanssiarviot
- oletetaan, että termi T_j esiintyy kyselyssä Q ja toinen termi T_k esiintyy dokumentissa D, joka on relevantti kyselyn Q kannalta
- jos T_j ja T_k sijoitetaan samaan tesauruksen luokkaan, lisääntyy Q:n ja D:n samanlaisuus
  - T_j ja T_k korvataan samalla luokan tunnisteella
- voidaan myös varmistaa, ettei tesauruksen luokkiin tule kahta termiä, joista toinen esiintyy haussa ja toinen epärelevantissa dokumentissa

18

Automaattinen tesauruksen muodostaminen

- automaattisesti jonkin tekstikokoelman perusteella tuotettua tesaurusta voidaan käyttää termien korvaamiseen vain samantyyppisiä tekstejä indeksoitaessa
  - tai tesauruksen tuottamiseen on käytettävä hyvin monenlaisia tekstiaineistoja, jotta saadut luokat ovat riittävän yleisiä
- jos käytetään hyväksi hakuja ja relevanssiarvioita, myös hakujen täytyy kattaa kokoelman aihepiirejä hyvin laajasti

19

Dokumenttien kuvaajien muodostaminen

- kerää jokaisesta dokumentista kaikki dokumentissa esiintyvät sanat
- poista hukkasanalistan sanat
- muokkaa tarvittaessa jäljelle jääviä sanoja
- laske termeille paino kussakin dokumentissa käyttäen tf·idf-funktiota
- kuvaa kukin dokumentti joukolla termejä ja niiden painoja

$$D_i = (T_1, w_{i1}; T_2, w_{i2}; ...; T_t, w_{it})$$

20

Vaihtoehto

- kerää jokaisesta dokumentista kaikki dokumentissa esiintyvät sanat
- poista hukkasanalistan sanat
- muokkaa tarvittaessa jäljelle jääviä sanoja
- laske erotteluarvo kaikille termeille
- korvaa tesauruksen avulla yleisemmillä käsitteillä kaikki termit, joiden erotteluarvo on lähellä nollaa
- korvaa fraaseilla termit, joiden erotteluarvo on negatiivinen (= hyvin yleiset termit)

21

Vaihtoehto, jatkuu

- laske painot yksittäisille termeille, fraaseille ja tesauruksen käsitteille
  - fraasin paino, esim., komponenttien painojen keskiarvo
- kuvaa jokainen dokumentti joukolla yksittäisiä termejä, tesauruksen käsitteitä ja fraaseja sekä niihin liittyvillä painoilla
- kummassakin vaihtoehdossa voidaan ajatella, että dokumenttikokoelmassa on yhteensä t termiä ja jokainen dokumentti kuvataan näillä t:llä termillä
  - jos termi ei esiinny dokumentissa, termin paino on 0

22

Kyselyiden kuvaajat

- jos ajatellaan, että kyselyt annetaan luonnollisen kielen lauseina, kyselyiden kuvaajat muodostetaan kuten dokumenttien kuvaajat
  - termit + painot
- koska kyselyt ovat yleensä lyhyitä, ei termifrekvenssillä (tf) ole merkitystä
  - painona käytetään pelkkää käänteistä dokumenttifrekvenssiä (idf)

23

Hakemiston muodostaminen

- termien valinnan ja painotuksen jälkeen jokaista dokumenttia kohti on talletettuna joukko termejä (tavalliseen peräkkäistiedostoon)
- kysely sisältää joukon termejä
  - hakutilanteessa on löydettävä nopeasti dokumentit, joissa termit esiintyvät
- muodostetaan käänteistiedosto (inverted file), jossa jokaista termiä kohti on talletettuna dokumentit, joissa termi esiintyy
- lisäksi yleensä muodostetaan sanakirjatiedosto (dictionary file) käänteistiedoston hakemistoksi

24

sanakirjatiedosto:

termit	a	b	c	d	e	f	g	h	j	k	l	...
df	3	2	3	3	1	2	3	1	3	2	2	

käänteistiedosto:

termit	a	b	c	d	e	f	g	h	j	k	l	...
dokumentit	136	37	246	689	8	48	137	2	157	59	24	

perustiedosto:

dokumentit	1	2	3	4	5	6	7	8	9	...		...
termit	agj	cht	abg	cft	jk	acd	bgj	def	dk	...		

25

Käänteistiedoston muodostaminen

- käänteistiedosto voidaan muodostaa useilla eri menetelmillä, esimerkiksi näin:
- perustiedostoa luetaan dokumentti kerrallaan
  - muodostetaan lista (termi, dokumentti) –pareista
    - (a,1), (g,1), (j,1), (c,2), (h,2), (t,2), (a,3), (b,3), (g,3), (c,4), (f,4), (t,4), (j,5), (k,5), (a,6), (c,6), (d,6),…
- lista järjestetään termien mukaan nousevaan järjestykseen (jos sama termi, dokumenttinumeron mukaan nousevaan):
  - (a,1), (a,3), (a,6), (b,3), (c,2), (c,4), (c,6), (d,6), (f,4), (g,1), (g,3), (h,2), (j,1), (j,5), (k,5), (t,2), (t,4),…

26

Käänteistiedoston muodostaminen

- lopuksi yhdistetään saman avaimen sisältävät parit: avainarvoon liitetään kaikki eri dokumenttinumerot lajiteltuna listana
  - (a,<1,3,5>), (b,<3>), (c,<2,4,6>), (d, <6>), (f,<4>), (g, <1,3>), (h, <2>), (j,<1,5>), (k, <5>), (t,<2,4>),…
  - tästä muodosta voidaan muodostaa myös sanakirjatiedosto

27

Toteutuksesta

- edellä on jätetty pois termien painot dokumenteissa, mutta ne ajatellaan olevan myös perustiedostossa
  - poimitaan siis kolmikkoja (termi, dokumentti, termin paino)
  - jos käytetään tf · idf-painoja, perustiedostoon riittää tallettaa tf, koska idf on termille sama kaikissa dokumenteissa

28

Toteutuksesta

- tässä menetelmässä (termi,dokumentti)-parien järjestäminen on raskain operaatio
- kun dokumenttikokoelma on kohtuullisen suuri, järjestämistä ei voi tehdä keskusmuistissa
  - järjestäminen voidaan tehdä lomitusjärjestämismenetelmällä (external merge sort)

29

Järjestäminen

- oletetaan, että
  - (dokumentti-termi)-parit on talletettu levylle
  - keskusmuistissa on tilaa k kpl (dokumentti,termi)-parille
- luetaan k kpl (dokumentti, termi)-pareja keskusmuistiin ja järjestetään ne esim. pikajärjestämistä (quicksort) käyttäen
- kirjoitetaan järjestetty lista pareja takaisin levylle
- toistetaan, kunnes kaikki parit on käsitelty kertaalleen

30

Järjestäminen

- lomitusvaihe:
  - luetaan 2 ensimmäistä järjestettyä listaa levyltä, lomitetaan ne yhdeksi listaksi ja kirjoitetaan lista takaisin levylle
  - luetaan seuraavat 2 listaa jne. kunnes kaikki k:n mittaiset listat on käyty läpi
  - luetaan sitten 2 ensimmäistä 2k:n mittaista listaa ja lomitetaan ne
  - jatketaan samalla periaatteella, kunnes jäljellä on vain yksi lista

31

Tehokkuudesta

- mitä suurempi osa pareista mahtuu kerrallaan keskusmuistiin, sitä nopeampi indeksointimenetelmä on
- menetelmä vaatii paljon levytilaa
  - järjestämisen aikana tarvitaan kaksi kopiota parit sisältävästä tiedostosta
- hyvin suuret dokumenttikokoelmat täytyy järjestää muilla menetelmillä

32

Indeksoinnin raekoko

- edellä on hakemistoon talletettu tieto termin sijainnista dokumentin tarkkuudella
- jos halutaan tukea läheisyyskyselyjä tai dokumenttia pienempien tekstifragmenttien palauttamista hakutuloksena, voidaan termin sijainti merkitä indeksiin tarkemmin
- toisaalta voitaisiin määritellä niin, että dokumentti on tekstikappale, virke tai sana
  - tällöin hukataan tieto dokumentin osien hierarkiasta

33

Indeksoinnin raekoko

- ”normaali” tapaus:
  - information: <D345, D348,D350,…>
  - retrieval: <D123, D128, D345,…>
- lisätään tieto, missä lauseessa termi esiintyy:
  - information: <D345,25; D345,37; D348, 10; D350,8;…>
  - retrieval: <D123,5; D128,25; D345,37; D345,40;…>
  - voidaan vastata nopeasti kyselyyn: “information samassa lauseessa kuin retrieval” → D345
  - tilaa kuluu enemmän kahdesta syystä: lauseen numero + dokumentin sisällä useassa lauseessa esiintyvä termi tuottaa listaan useita alkioita (aikaisemmin vain yhden)

34

Indeksoinnin raekoko

- lisätään tieto, monentenako sanana termi esiintyy
  - information: <D345,25,4; D345,37,3; D348, …>
  - retrieval: <D123,5,2; D128,25,4; D345,37,4;…>
  - voidaan vastata kyselyihin
    - “information ja retrieval vierekkäin”
    - “information ja viiden sanan sisällä retrieval”
- jos dokumenteista on talletettu myös metatietoja (tekijä, otsikko, kustantaja,…), voidaan samoin kuin yllä lisätä hakemistotietueeseen tieto sijainnista jossakin metatietokentässä
  - voidaan vastata kyselyihin kuten “tekijä on Aleksis Kivi”

35

Indeksoinnin raekoko

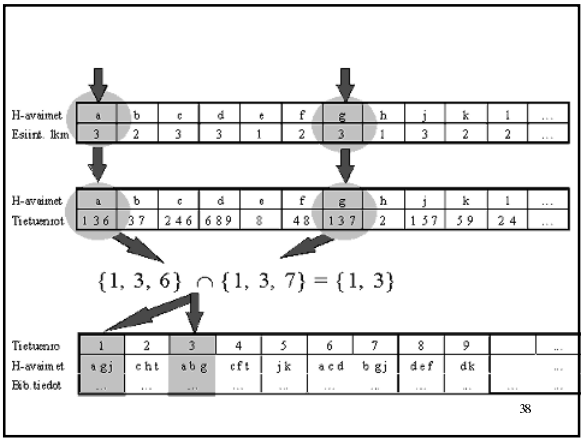
- jos ei ole odotettavissa erittäin paljon läheisyysoperaattoreita käyttäviä kyselyjä, indeksointi dokumentin tasolla on riittävä
  - läheisyysrajoitukset voidaan tarkistaa hakutuloksista jälkiprosessointina

36

Käänteistiedoston käyttäminen

- haetaan sanakirjasta kyselyn hakutermejä vastaavat tietueet
  - tietueessa on osoitin termiä vastaavaan tietueeseen käänteistiedostossa
- haetaan termejä vastaavat tietueet käänteistiedostosta
  - saadaan kullekin termille lista dokumenteista, joissa se esiintyy
- dokumenttilistoja tutkitaan haun muotoilua vastaavalla tavalla
  - esim. "a and g": etsitään listojen yhteisiä dokumentteja
- poimitaan dokumentit numeroidensa perusteella perustiedostosta
- jos kaikki ehtoja ei voitu ratkaista käänteistiedoston perusteella, tutkitaan loput ehdot poimituista dokumenteista

37



38

Tässä osassa

- fraasien muodostaminen yleisen termin ympärille
- termin korvaaminen yleisemmällä tesauruksesta löydettävällä käsitteellä
- tesauruksen muodostaminen automaattisesti
- käänteistiedoston muodostaminen järjestämismenetelmällä
- indeksoinnin raekoko
- käänteistiedoston käyttäminen

39