Tiedonhakumenetelmät

Helena Ahonen-Myka

Kevät 2004, osa 7

Ryvästäminen

Tässä osassa

- dokumenttien ryvästäminen (klusterointi)
  - dokumentti-dokumentti –samanlaisuusmatriisiin perustuvat menetelmät
  - heuristiset menetelmät
- ryvästyksen käytöstä tiedonhaussa

Ryvästäminen

- ryvästyshypoteesi (clustering hypothesis): keskenään samanlaiset dokumentit ovat yleensä relevantteja samojen hakutehtävien kannalta
- → jos samanlaiset dokumentit ryhmitellään, voidaan tehostaa hakua
  - selaavan tiedonhaun (browsing) tuki
  - samanlaiset dokumentit voidaan tallettaa fyysisesti lähekkäin → nopeampi hakea
- voidaan tarjota rinnakkainen tapa dokumenttien saavuttamiseen (käänteistiedoston lisäksi)

Ryvästäminen

- kaksi osaongelmaa tiedonhaun kannalta
  - ryvästyksen muodostaminen
    - yleensä harvoin tehtävä, joten voi olla raskaampi operaatio
    - kun kokoelma päivittyy, täytyy ryvästäminen yleensä tehdä uudelleen
  - ryvästyksen hyödyntäminen hakuvaiheessa
    - täytyy olla nopeaa

Ryvästämismenetelmiltä toivottavia ominaisuuksia

- tehokkuus, lähinnä ajan suhteen
- muutamia ryvästystulokseen liittyviä ominaisuuksia
  - ryvästyksen ei pitäisi muuttua huomattavasti, kun lisätään uusia dokumentteja
  - jos dokumenttien kuvaajissa on pieniä virheitä, on näiden virheiden vaikutus ryvästykseen myös pieni
  - ryvästystuloksen tulisi olla riippumaton dokumenttien käsittelyjärjestyksestä

Ryvästysmenetelmät

- menetelmät, jotka perustuvat dokumentti-dokumentti –matriisiin
  - täyttävät ryvästystulokseen liittyvät ehdot
  - hitaita: $O(n^2)$
- heuristiset menetelmät
  - eivät täytä em. ehtoja, mutta tulos on usein kohtuullisen hyvä
  - nopeita: O(n log n)

Ryvästysmenetelmät

- menetelmät, jotka perustuvat dokumentti-dokumentti –matriisiin
  - hierarkkiset (hierarchical) menetelmät
- heuristiset menetelmät
  - yhden läpikäynnin (one pass) menetelmä
  - k-means -menetelmä

7

dokumentti-termi –matriisi (kertaus)

dok.			ter	mit				
vekt.	a	b	c	d	e	f	g	h
D1	1	0	1	1	0	1	0	1
D2	0	1	1	0	0	1	1	0
D3	0	0	0	1	1	0	1	0
D4	1	1	0	0	1	0	0	1

8

dokumentti-dokumentti –matriisi (kertaus)

doku		doku	mentit	
mentit	D1	D2	D3	D4
D1	1	0.45	0.26	0.45
D2		1	0.29	0.22
D3			1	0.26
D4				1

9

Yksinkertainen menetelmä

- kahden dokumentin d_i ja d_j välistä samanlaisuutta mitataan kosinikaavalla $\cos(d_i, d_j)$
- valitaan jokin kynnysarvo
- jos dokumenttien d_i ja d_j samanlaisuusarvo ylittää kynnysarvon, d_i ja d_j kuuluvat samaan ryppääseen
- esim. edellisen kalvon esimerkissä
  - kynnysarvo 0.4: ryppäät {D1, D2, D4} ja {D3}
  - kynnysarvo 0.5: jokainen dokumentti muodostaa oman ryppäänsä

10

Hierarkkiset menetelmät

- tiedonhaun kannalta on toivottavaa tuottaa hierarkinen ryvästys
- em. yksinkertaista menetelmää voitaisiin toistaa eri kynnysarvoilla
  - kynnysarvojen sopiva valinta voi olla vaikeaa?
- on myös erityisiä menetelmiä, jotka perustuvat ryvästyshierarkian muodostamiseen
  - voidaan käyttää koko hierarkiaa hyväksi tai määritellä kynnysarvo, josta hierarkia leikataan poikki → saadaan useita erillisiä ryppäitä

11

Hierarkkiset menetelmät

- voidaan lähteä liikkeelle tilanteesta, jossa kaikki dokumentit muodostavat oman ryppäänsä
  - ryppäitä yhdistellään, kunnes jäljellä on vain yksi rypäs → agglomeratiivinen ryvästys
- tai alkutilanteessa kaikki dokumentit voivat olla yhdessä ryppäässä
  - ryppäitä jaetaan, kunnes kaikki dokumentit muodostavat oman ryppäänsä → divisiivinen ryvästys

12

Hierarkkinen, agglomeratiivinen ryvästäminen

- 1. muodosta dokumentti-dokumentti –matriisi
- 2. sijoita kukin dokumentti omaan ryppääseensä
- 3.
  - muodosta uusi rypäs yhdistämällä ryppäät i ja j, joiden samanlaisuusarvo on suurin
  - päivitä samanlaisuusmatriisi poistamalla ryppäitä i ja j vastaavat rivit ja sarakkeet
  - lisää uusi rivi ij ja laske uuden ryppään samanlaisuus muihin ryppäisiin nähden → rivin alkiot
- toista askelta 3, kunnes jäljellä on vain yksi rypäs

13

Hierarkkinen, agglomeratiivinen ryvästäminen

- kahden ryppään välinen samanlaisuus
  - jos kummassakin ryppäässä on vain yksi dokumentti, ryppäiden välinen samanlaisuusarvo on dokumenttien välinen samanlaisuusarvo
  - jos ryppäissä on enemmän dokumentteja, samanlaisuusarvo täytyy määritellä erikseen: useita vaihtoehtoja
    - yhden linkin (single link) kriteeri
    - täydellisen linkityksen (complete link) kriteeri
    - ryhmän keskiarvo (group average) -kriteeri

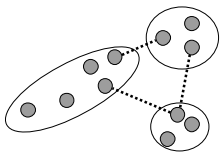
14

Yhden linkin kriteeri

- kahden ryppään välinen samanlaisuusarvo on niiden kahden, eri ryppäissä olevan, dokumentin välinen samanlaisuusarvo, joiden samanlaisuusarvo on suurin
- jokainen (uuden) ryppään dokumentti on lähempänä jotakin muuta saman ryppään dokumenttia kuin mitä tahansa dokumenttia muissa ryppäissä

15

Yhden linkin kriteeri



Yhdistetään ne kaksi ryvästä, joiden lyhin etäisyys on pienin.

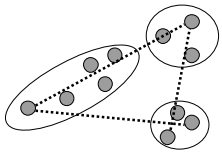
16

Täydellisen linkityksen kriteeri

- kahden ryppään samanlaisuusarvo on niiden kahden, eri ryppäissä olevan, dokumentin samanlaisuusarvo, joiden samanlaisuusarvo on pienin
- kukin (uuden) ryppään jäsen on samanlaisempi kaikkien ryppään dokumenttien kanssa kuin minkä tahansa muun ryppään kaikkein erilaisimman dokumentin kanssa

17

Täydellisen linkityksen kriteeri



Yhdistetään ne kaksi ryvästä, joiden pisin etäisyys on pienin.

18

Ryhmän keskiarvo -kriteeri

- kompromissiratkaisu
- kullakin ryppään dokumentilla on suurempi keskimääräinen samanlaisuusarvo muihin oman ryppäänsä dokumentteihin kuin minkä tahansa muun ryppään kaikkiin dokumentteihin

19

Heuristiset menetelmät

- heuristiset menetelmät perustuvat dokumenttivektorien vertailuun vain tarvittaessa
  - dokumentti-dokumentti –matriisia ei tarvitse muodostaa
- usein tarvitaan kokeellisesti määritettyjä parametreja, (joita ei aina ole helppo määrätä) esim.
  - haluttu ryppäiden lukumäärä
  - kunkin ryppään minimi- ja maksimikoko (dokumenttien lukumäärä)
  - dokumentin ja ryppään välisen samanlaisuuden kynnysarvo; jos samanlaisuus pienempi, dokumenttia ei liitetä ryppääseen.
  - kuinka paljon ryppäät saavat mennä päällekkäin

20

Heuristiset menetelmät: yhden läpikäynnin menetelmä

- esim. yhden läpikäynnin (one pass) menetelmä
  - jokainen dokumentti käsitellään yhden kerran
  - ensimmäinen dokumentti muodostaa oman ryppään
  - myöhemmin käsiteltävän dokumentin vektoria verrataan olemassaoleviin ryppäisiin (esim. keskusalkioon)
  - jos samanlaisuusarvo ylittää annetun kynnysarvon, liitetään dokumentti yhteen tai useampaan ryppääseen (ja päivitetään keskusalkioita)
  - muuten dokumentti muodostaa yksin uuden ryppään

21

Heuristiset menetelmät: yhden läpikäynnin menetelmä

- yhden läpikäynnin menetelmä muodostaa usein hyvin erikokoisia ryppäitä
- ryvästyksen tasapainoisuutta täytyy tarkkailla ja korjata ryppäiden kokoa, ryppäiden lukumäärää ja ryppäiden välisen päällekkäisyyden määrää (= yhteisten dokumenttien lukumäärä)
- esim. voidaan antaa keskimääräinen ryppään koko
  - jos rypäs kasvaa suuremmaksi, se jaetaan kahtia

22

Heuristiset menetelmät: k-means -menetelmä

- valitaan ryppäiden lukumäärä k
- jaetaan dokumentit k ryppääseen jollakin tavalla
- kutakin dokumenttia verrataan vuorollaan ryppäiden keskusalkioihin ja dokumentti liitetään siihen ryppääseen, jonka keskusalkio on samanlaisin
- läpikäynnin jälkeen lasketaan keskusalkiot uudelleen ja toistetaan läpikäynti, kunnes ryvästyksessä ei tapahdu enää merkittäviä muutoksia

23

Hybridimenetelmät

- esim. I
  - muodostetaan ensin karkea ryvästys käyttäen jotain heuristista menetelmää
  - käytetään sitten dokumentti-dokumentti –matriisiin perustuvaa menetelmää ryppäiden jakamiseksi aliryppäiksi
- esim. II
  - poimitaan jokin dokumenttien (kuvaava) osajoukko ja ryvästetään tämä otos dokumentti-dokumentti –matriisiin perustuvalla menetelmällä
  - loput dokumenteista ryvästetään syntyneisiin ryppäisiin heuristista menetelmää käyttäen

24

Lisää ryvästyksen käytöstä tiedonhaussa

- edellä on ajateltu, että dokumenttikokoelma ryvästetään kokonaan
  - päivitysten jälkeen ryvästys on tehtävä uudelleen
  - hakutilanteessa ryppäät ovat siis valmiiksi olemassa
  - käyttäjän kyselyä verrataan ryppäisiin ja tuloksena palautetaan parhaat ryppäät
- voitaisiin myös rajoittua ryvästämään dynaamisesti vain hakutulokset
  - dokumenttiavaruus on pienempi eikä muutu
  - ryvästyksen täytyy olla nopeaa
  - ryvästyshypoteesi toimii paremmin, koska dokumentit ovat vastauksia tiettyyn kyselyyn

25

Hakutulosten ryvästäminen: Scatter/Gather-menetelmä

- 1. tehdään ensimmäinen ryvästys
- 2. käyttäjälle näytetään esimerkinomaiset tekstilyhennelmät (esim. otsikoita) kustakin ryppäästä
- 3. käyttäjä valitsee kiinnostavat ryppäät
- 4. valittujen ryppäiden dokumentit yhdistetään yhteen joukkoon, joka ryvästetään
- 5. jatketaan askeleesta 3, kunnes käyttäjä ei halua jatkaa

26

Hakutulosten ryvästäminen: Scatter/Gather-menetelmä

- käyttäjä voi selailla dokumentteja millä tasolla tahansa
- ryppäät käsittelevät aina vain pienempiä aiheita, mutta eivät välttämättä aluksi valittujen aiheiden aliaiheita
  - käyttäjä voi myös vaihtaa kiinnostuksen kohdettaan

27

Tässä osassa

- dokumentti-dokumentti –samanlaisuusmatriisiin perustuvia ryvästysmenetelmiä
  - yksinkertainen menetelmä
  - hierarkkinen, agglomeratiivinen ryvästäminen
- heuristisia menetelmiä
  - yhden läpikäynnin menetelmä
  - k-means -menetelmä
- hybridimenetelmiä
- ryvästyksen käyttö tiedonhaussa: hakutulosten ryvästäminen

28