Tiedonhakumenetelmät

Helena Ahonen-Myka
kevät 2004, osa 1
Johdanto

1. Tässä osassa...

- määritelmä
- kurssin järjestelyistä
- johdanto kurssin aihepiiriin
- kurssin sisällöstä

Määritelmä

- Tiedonhaku
- Information retrieval; information storage and retrieval (IR)
- ”IR is concerned with the processes involved in the representation, storage, searching and finding of information which is relevant to a requirement for information desired by a human user.” (Ingwersen, 1992)

Kurssin järjestelyistä

- laudatur-kurssi, 3 ov
- luennot: ti 12-14 A217, to 10-12 A414
  - poikkeus: to 25.3. B450
- harjoitukset (Juha Makkonen) alkaen 25.3.
  - to 12-14 A319
  - pe 10-12 A216
- ei luentoja 8.-15.4.
- ei harjoituksia 8.-9.4., 22.-23.4.
  - huom. 15.-16.4. on harjoitukset, vaikka tuolla viikolla ei ole luentoja!

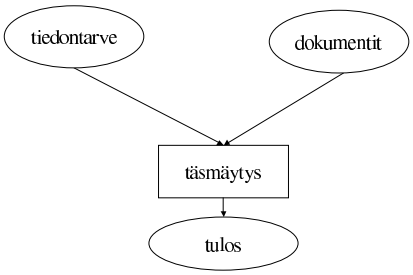
Kurssin järjestelyistä

- harjoitustyö
  - työn määrittely annetaan ti 23.3.
  - tehdään 4-5 hengen ryhmissä, ryhmät kootaan 1. harjoituksissa (to 25.3. ja pe 26.3.)
  - työstä palautetaan raportti
  - viimeisillä harjoituskerroilla 6.-7.5. ryhmät esittelevät töitään
- koe: ma 10.5. klo 16-20 päärakennus, Sali 1
- materiaali
  - kalvot kurssin sivustolla
  - kuhunkin luentoon liittyvä keskeinen muu materiaali kerrotaan myös kurssin sivustolla

Johdanto

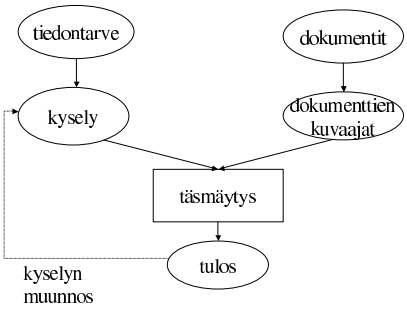
- tiedonhaun tavoitteena on tiedontarpeiden tyydyttäminen
- tiedonhaussa pyritään löytämään tiedontarpeiden tyydyttämistä mahdollisimman hyvin palveleva dokumentti tai dokumenttijoukko

Tiedonhakuprosessi



7

Tiedonhakuprosessi



8

Erilaisia tiedontarpeita

- aihehaku
  - "julkisen talouden supistamistoimet"
- yksilöhaku
  - "Esko Ukkosen väitöskirja"
- faktahaku
  - "HY:n kansleri v. 2002"

9

Erityyppisiä dokumentteja

- tiukasti rakenteiset (relaatiotietokanta)
  - esim. kirjastotietokanta, aikataulut,…
- puolirakenteiset
  - esim. viitetietokanta: julkaisun metatiedot ja tiivistelmä
  - XML-dokumentit
- rakenteettomat tekstidokumentit
  - sisältävät silti tekstin luontaista rakennetta

10

Tiedontarpeet vs. erilaiset dokumenttityypit

- kyselyt voivat olla joko täsmällisiä tai sumeita
  - täsmällinen kysely: vastausdokumentit voidaan kuvailla joidenkin ominaisuuksien mukaan yksikäsitteisellä tavalla
  - sumea kysely: vastausdokumentteja ei voida kuvailla joko tiettyjen ominaisuuksien avulla tai yksikäsitteisellä tavalla

11

Tiedontarpeet vs. erilaiset dokumenttityypit

- täsmällinen tietokantakysely
  - "vuonna 2001 aloittaneet tietojenkäsittelytieteen pääaineopiskelijat"
    - attribuutit: aloitusvuosi, koulutusohjelma
  - vastaus aina oikein (jos tietokannassa ei ole virheitä)
- tietokantakysely voi olla myös sumea
  - voitaisiin palauttaa vuosina 2000-2002 aloittaneet (esim. jos v. 2001 ei olisi aloittaneita)

12

Tiedontarpeet vs. erilaiset dokumenttityypit

- vapaaseen tekstiin kohdistuvat kyselyt ovat yleensä aina sumeita: esim. ”julkisen talouden supistamistoimet”
  - vaikea tietää, millä sanoilla kyseistä aihetta on eri dokumenteissa käsitelty: supistaa, karsia, alasajo, hevoskuuri, kurittaa, keventää,...
  - monia muitakin asioita on voitu käsitellä samoilla sanoilla
  - → vastaukset ovat usein epätäydellisiä
  - → myös virheellisiä dokumentteja saadaan tuloksena

13

Tiedontarpeet vs. erilaiset dokumenttityypit

- puolirakenteisiin dokumentteihin kohdistuvat kyselyt yhdistelevät tarkkoja ja sumeita kyselyjä
  - ”kirjat, joiden kirjoittaja on Aleksis Kivi, ja joissa esiintyy henkilö Juhani”

14

Tiedontarpeet vs. erilaiset dokumenttityypit

- kyselyn tulos voi olla suora tai epäsuora
  - suora: saadaan suoraan haettu tieto
  - epäsuora: tulokset ovat osoitteita varsinaisesti haetun tiedon lähteille, esim. kirjallisuusviitteitä dokumentteihin tai osoitteita yrityksiin

15

Tiedontarpeet vs. erilaiset dokumenttityypit

- aikaisemmin tiedonhakujärjestelmät (tekstitietokannat) ja tietokannanhallintajärjestelmät (“tavalliset” tietokannat”) erotettiin tutkimuksessa ja tuotekehittelyssä toisistaan
- nykyisin järjestelmät ovat usein integroituja järjestelmiä, joilla pystytään hallitsemaan sekä rakenteista että vapaamuotoista tietoa (myös kuvia ja multimedia yms.)
  - vrt. myös XML: rajaa ei voi vetää

16

Kyselyjen ja dokumenttien esitysmuodot

- vapaamuotoisten, luonnollisellä kielellä esitettyjen kyselyjen ja tekstidokumenttien vertailu on hankalaa
- → sekä kyselyjen että dokumenttien esitysmuoto on muokattava sopivammaksi
  - usein joukko termejä
  - termi = semanttinen ilmaisun yksikkö, esim. sana, fraasi, sanan vartalo

17

Kyselyjen esitysmuodot

- joukko hakutermejä
- lauseke, jossa hakutermejä on yhdistetty Boolen operaattoreilla
  - “Aleksis and Kivi”
  - “(text or image) and retrieval”
- läheisyysoperaattorilla yhdistettyjä termejä
  - “Aleksis near Kivi”

18

Kyselyjen esitysmuodot

- myös luonnollisen kielen lauseita
  - esim. ns. kysymysvastausjärjestelmissä käyttäjä voi antaa kysymyksiä
    - "Kuka oli HY:n kansleri v. 2002?"
  - kyselynä voi olla myös aikaisemmin löydetty dokumentti
    - etsitään dokumentteja, jotka ovat samanlaisia kuin tämä
  - lauseet esiprosessoidaan: perusmuotoon palautus, liian yleisten sanojen poistaminen
  - voidaan myös analysoida tarkemmin: mikä on kysymyssana, mitä odotetaan vastaukseksi

19

Dokumenttien esitysmuodot

- dokumentti voidaan kuvata
  - automaattisesti siitä tilastollisin perustein valittujen termien avulla
  - automaattisesti siitä lingvistisin perustein valittujen ja muokattujen termien avulla
  - ihmisen valitsemien termien avulla
- vaihtoehdot eivät ole toisiaan poissulkevia

20

Dokumenttikokoelman esitystapa

- haun vastauksen täytyy tulla nopeasti
- dokumenttikokoelma voi olla hyvin suuri
- termejä voi olla hyvin suuri määrä
  - (~ 10 000 – 100 000)
- → ei voida käyttää tekstiä selaavaa merkkijonohakua (esim. UNIXin grep)
- → ei voida rakentaa erikseen indeksejä jokaiselle termille tai termiyhdistelmälle (vrt. tietokantojen indeksit attribuuteille)

21

Dokumenttikokoelman esitystapa

- rakennetaan käänteishakemisto, jonka kautta kunkin termin esiintymät dokumenttikokoelmassa löydetään nopeasti
- indeksointi = termien valinta + käänteishakemiston muodostaminen
  - raskas operaatio, joka tehdään harvoin
- seuraavalla kalvolla
  - yläosassa kunkin dokumentin sisältämät termit
  - alaosassa kunkin termin esiintymät eri dokumenteissa → käänteishakemisto

22

	T1	T2	T3	T4
D1	1	1	0	1
D2	0	1	1	1
D3	1	0	1	1
D4	0	0	1	1

	D1	D2	D3	D4
T1	1	0	1	0
T2	1	1	0	0
T3	0	1	1	1
T4	1	1	1	1

23

Sanasto

- indeksoitujen (= käänteishakemistoon valittujen) termien luettelo
  - kertoo myös kunkin termin frekvenssin ja antaa linkin käänteishakemistoon

T1	2	→
T2	2	→
T3	3	→
T4	4	→

24

Kyselyn täsmäytys dokumentteihin

- jos kyselyssä on vain yksi hakutermi
  - etsitään termi sanastosta
  - linkin perusteella etsitään oikea kohta käänteishakemistosta
  - palautetaan käänteislistan dokumentit
- jos kyselyssä on useita hakutermejä, on lomitettava käänteislistoja
  - jos kunkin termin käänteislistat on tallennettu järjestyksessä dokumentin numeron mukaan, voidaan kahden listan lomitus tehdä yhdellä läpikäynnillä

25

Boolen kyselyjen käsittely

- A or B
  - A:n ja B:n käänteislistojen liitos
- A and B
  - A:n ja B:n käänteislistojen leikkaus
- A not B
  - A:n ja B:n käänteislistojen erotus
  - yleensä ei ole mahdollista käyttää not-operaatiota yksinään: vastaus olisi liian suuri

26

kysely: ((T1 or T2) and T3)

- käänteishakemistosta:
  - T1: <D1, D3>
  - T2: <D1, D2>
  - T3: <D2, D3, D4>
- T1 or T2: <D1, D2, D3>
- (T1 or T2) and T3: <D2, D3>

27

kysely: T1 and T3 and T4

- voidaan optimoida termien frekvenssin perusteella (nähdään suoraan sanastosta)
  - T1: 2 dokumenttia <D1, D2>
  - T3: 3 dokumenttia <D2, D3, D4>
  - T4: 4 dokumenttia <D1, D2, D3, D4>
- T1 and T3: <D3>
- (T1 and T3) and T4: <D3>
- vrt. jos laskettaisiin ensin T3 and T4

28

Läheisyysoperaattori

- indeksoinnin raekoko voi vaihdella
  - dokumentin numeron lisäksi voitaisiin käänteishakemistoon lisätä kappaleen numero, lauseen numero, sanan tarkka sijainti
- kyselyyn "A near B" voidaan löytää vastaus suoraan hakemiston avulla
  - veisi paljon tilaa
- jos riittävän tarkkaa sijaintitietoa ei ole talletettu hakemistoon, haetaan ensin esim. "A and B" ja tarkistetaan sitten läheisyys selaamalla tulosdokumentteja

29

Merkkijonohahmon etsintä selaamalla

- perusongelma: etsi hahmon P esiintymät merkkijonosta T
  - etsi kaikki sanan 'relevanssi' esiintymät dokumentista
- find-operaatio selaimessa tai editorissa, grep-komento Unixissa
- tarkempi haku hakutuloksista jälkiprosessointina
  - esim. läheisyysoperaattorin toteutus
- jos dokumenttikokoelma mahtuu keskusmuistiin, myös muuten

30

Kurssin sisältö

- relevanssin käsite, tiedonhakujärjestelmien evaluointi
- dokumenttikokoelmien indeksointi
- kyselyjen käsittely, täsmäytysmenetelmät
- dokumenttien klusterointi, tiedon suodatus ja reititys, muita sovelluksia
- merkkijonohahmon etsintä selaamalla
- luonnollisen kielen käsittely tiedonhakua varten, monikielinen tiedonhaku
- rakenteen hyödyntäminen (XML, linkit)

31

Kurssin jälkeen...

- Tampereen yliopiston informaatiotutkimuksen laitoksen kurssit (KIT-verkoston kautta)
  - Tiedonhaun kieliteknologia (verkkokurssi)
  - Kieltenvälisen tiedonhaun menetelmät (verkkokurssi)
  - Johdanto tiedonhaun tutkimukseen (voi suorittaa kirjatenttinä)
- Laajojen dokumenttiaineistojen käsittely K05
- mahdollisesti myös muuta aihepiiriin liittyvää

32