Tiedonhakumenetelmät

Helena Ahonen-Myka
Kevät 2004, osa 11
Tiedonhakustrategioita
Käyttöliittymä- ja visualisointiratkaisuja

Tässä osassa

- tiedonhakustrategioita
  - haku, selailu, navigointi, silmäily
  - suodatus ja reititys
- käyttöliittymä- ja visualisointiratkaisuja
- loppujärjestelyistä yms.

Tiedonhakuprosessi

- 1. tiedonhakijalla on tiedontarve
- 2. tiedonhakija muotoilee kyselyn
- 3. tiedonhakija lähettää kyselyn järjestelmälle
- 4. järjestelmä palauttaa tuloslistan
- 5. tiedonhakija silmäilee ja arvioi tuloksia
- 6. jos tiedonhakija on tyytyväinen, hän lopettaa
- 7. jos tiedonhakija ei ole tyytyväinen, hän muokkaa kyselyä ja palaa askeleeseen 3

Tiedonhakuprosessi

- taustaoletukset:
  - tiedonhakijan tiedontarve pysyy samana koko hakuprosessin aikana
  - prosessi on onnistunut, jos kyselyä iteroiden muokkaamalla lopputuloksena on dokumenttijoukko, jossa on kaikki relevantit dokumentit eikä epärelevantteja dokumentteja

Tiedonhakuprosessi

- käytännössä tiedonhakija oppii koko ajan uutta prosessin aikana
  - hakija silmäilee tuloslistan otsikoita, hakusanoja, tulosdokumentteja tai navigoi hyperlinkkejä pitkin
- "marjanpoimintamalli"
  - hakijan tiedontarve muuttuu hakuprosessin aikana
  - tiedontarve tyydyttyy hakuprosessin aikana (silmäilemällä, lukemalla) kerättyjen tiedonpalasten avulla
- haun lisäksi muita tiedonhakustrategioita ovat silmäily, selailu ja navigointi

Haku, selailu, navigointi ja silmäily

- haku (querying)
  - kohdedokumenttien eksplisiittinen kuvaaminen hakusanojen avulla
  - tuloksena kertaluonteisia dokumenttiryppäitä
- selailu (browsing)
  - hakija aloittaa jostakin potentiaalisesti mielenkiintoisesta alueesta ja selailee dokumentteja löytääkseen relevantteja
  - jos relevantteja ei löydy, siirrytään johonkin toiseen alueeseen
  - alkukohta esim. kyselyn avulla
  - oletus: saman aihepiirin dokumentit organisoitu yhteen

Haku, selailu, navigointi ja silmäily

- navigointi (navigating)
  - seurataan hyperlinkkejä kohti tunnettua päämäärää (esim. Juha Makkosen puhelinnumero TKTL:lla)
  - reitti oletetaan tunnetuksi (tai sen oletetaan helposti selviävän reittiä kuljettaessa)
- silmäily (scanning)
  - tiedonhakija silmäilee tuloslistan otsikoita, itse dokumentteja, hyperlinkkejä, metatietoja yms.
- valinta (selection)
  - apuoperaatio: esim. selailussa silmäillään sivuja ja valitaan hyperlinkki, jota seurataan

7

Sisältöperustainen suodatus ja reititys

- suodatus
  - tavoitteena on valita henkilölle tai organisaatiolle tulevasta dokumenttivirrasta (esim. päivän uutisdokumenteista, sähköposteista) kiinnostavat tai karsia ei-toivotut (esim. roskaposti)
- reititys
  - dokumenttivirran dokumentti ohjataan henkilölle, jota dokumentti kiinnostaa tai jonka toimialaan se kuuluu (esim. asiakkailta tulevien kysymysten ohjaus eri asiantuntijoille)

8

Sisältöperustainen suodatus ja reititys

- suodatus ja reititys perustuvat suotimiin (profiileihin)
- tavallisessa tiedonhakujärjestelmässä dokumenttikokoelma on yleensä varsin staattinen, kun taas kyselyt vaihtelevat
- suodatuksen ja reitityksen tapauksessa dokumenttikokoelma muuttuu jatkuvasti, kun taas suotimet ovat pitkään käytettäviä ja hitaasti muuttuvia kyselyjä

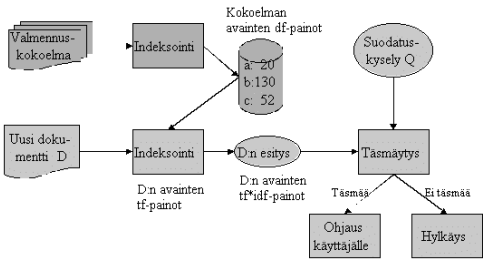
9

Sisältöperustainen suodatus ja reititys

- suotimet voivat perustua dokumenttien metatietoon (esim. sähköpostin lähettäjätietoon) tai sisältöön
- täystäsmäytys: suotimia sovelletaan Boolen kyselyinä kuhunkin saapuvaan dokumenttiin
- osittaistäsmäytys: dokumentin relevanssi kunkin suotimen suhteen → hyväksyntä/hylkäys tai valitaan paras vastaanottaja
  - ongelma: kokoelmaa ei oikeastaan ole olemassa → mistä saadaan df-arvot termien painotukseen?
  - ratkaisu: termien df-arvot voidaan oppia samankaltaisesta opetusaineistosta

10

Sisältöperustainen suodatus ja reititys



11

Käyttöliittymä- ja visualisointiratkaisuja

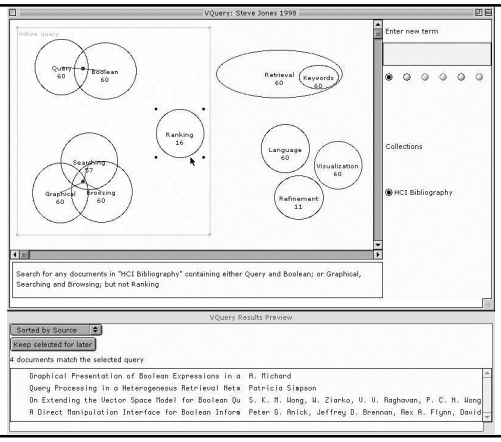
- kyselyn määrittely
- hakutulosten ja niiden kontekstin visualisointi

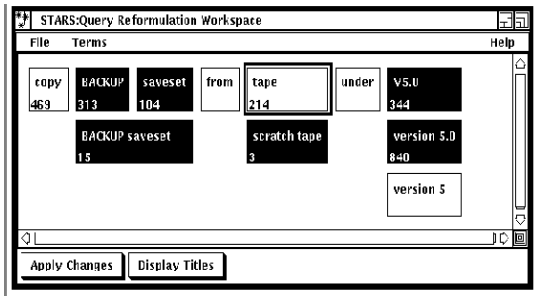
12

Kyselyn määrittely

- Boolen kyselyjen tekeminen on käyttäjälle usein vaikeaa
  - mm. AND ja OR poikkeavat merkitykseltään yleiskielisistä vastineista
- ratkaisuja
  - Venn-diagrammien käyttö (Hearst: kuva 10.10)
  - lohkovisualisointi (rajaavat ja rinnakkaiset käsitteet) (Hearst: kuva 10.12)

13



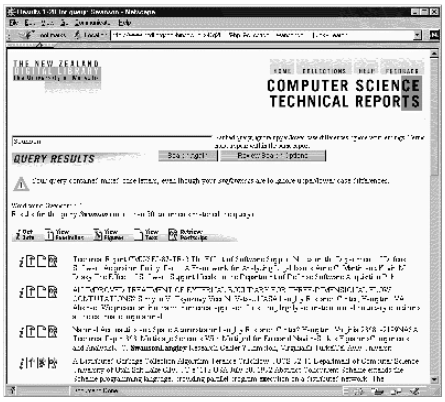


15

Hakutulosten ja niiden kontekstin visualisointi

- tiedonhakijan on helpompi hahmottaa tulosdokumenttien joukko, jos esim. hakusanojen esiintymistä dokumenteissa visualisoidaan jollakin tavalla
- tavallinen tapa: dokumenttisurrogaatit
  - dokumenttien otsikot, katkelma tekstin alusta, linkki tiivistelmään, luokkakoodi,… (Hearst: kuva 10.14)
- KWIC (keyword-in-context)
  - lauseet, joissa hakusanat ovat esiintyneet

16

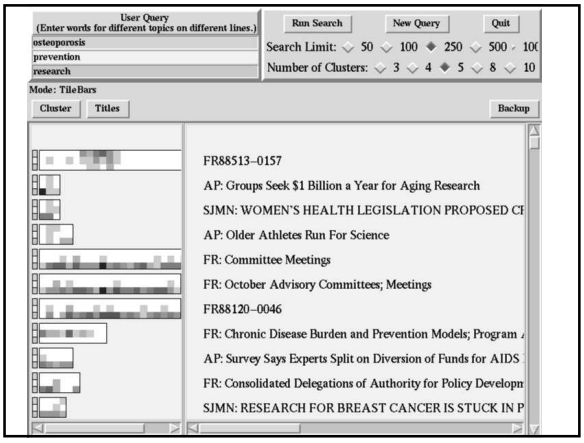


17

Hakutulosten ja niiden kontekstin visualisointi

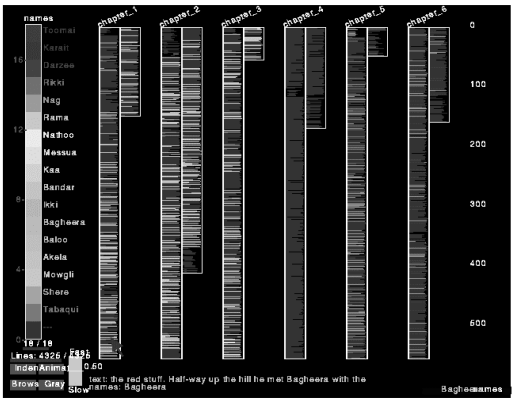
- TileBars
  - tiedonhakija erittelee hakusanat eri aspekteihin
  - tuloslistassa näytetään kutakin aspektia vastaavien hakusanojen esiintyminen kussakin dokumentin osassa (esim. tekstikappaleessa)
    - tiedonhakija voi nähdä, missä osissa kaikki aspektit ovat läsnä
    - (Hearst: kuva 10.15, parempi kuva PostScript- ja PDF-versioissa kuin HTML-versiossa)

18



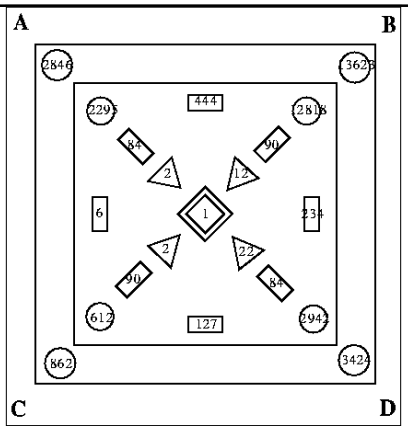
Hakutulosten ja niiden kontekstin visualisointi

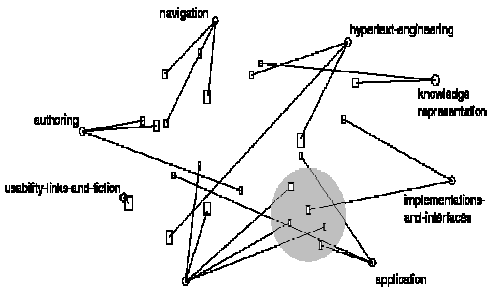
- SeeSoft
  - hakusanojen esiintymien visualisointi
  - kukin sarake kuvaa yhtä (kirjan) lukua
  - kukin väri kuvaa tietyn henkilöhahmon esiintymistä tekstissä
  - (Hearst: kuva 10.16)



Hakutulosten ja niiden kontekstin visualisointi

- InfoCrystal
  - jokaiselle hakusanojen kombinaatiolle: kuinka monessa dokumentissa kombinaatio esiintyy
  - (Hearst: kuva 10.17)
- Vibe
  - hakusanat sijoitetaan käsiteavaruuteen
  - kuvataan, miten dokumentit asettuvat tähän avaruuteen
  - (Hearst: kuva 10.18)

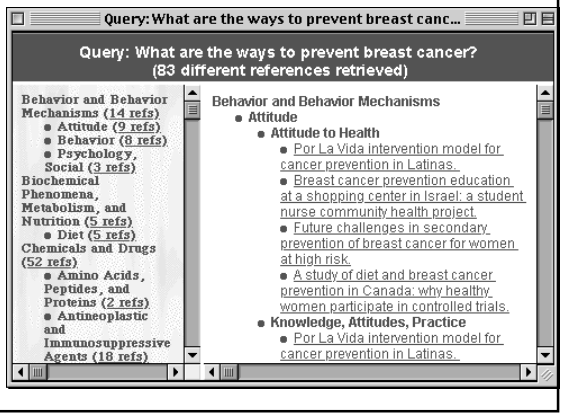




Hakutulosten ja niiden kontekstin visualisointi

- DynaCat
  - tulosdokumentit esitetään luokittelujärjestelmän mukaan järjestettynä
  - mukaan ei valita kaikkia luokkia, vaan vain ne, jotka ovat ennaltamääritellyn kyselytyypin kannalta relevantteja
    - esim. tyyppi: ”hoidon vastavaikutukset”
    - tyyppiin kuuluva kysely: ”Mitkä ovat aspiriinin sivuvaikutukset?”
  - (Hearst: kuva 10.20)

25

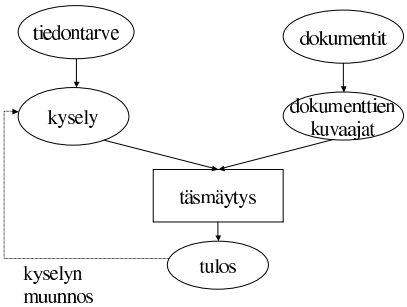


Käyttöliittymä- ja visualisointiratkaisuja

- tiedonhakujärjestelmien käyttöliittymien suunnittelussa on luonnollisesti paljon muitakin osa-alueita
  - relevanssipalaute: mitä automatisoidaan, mitä annetaan tiedonhakijan kontrolloida
  - hakuprosessin tukeminen: mm. miten hakuhistoria säilytetään
  - pitkäaikaisten tiedonhakuprosessien tukeminen (esim. kilpailevien yritysten jatkuva seuranta)
- + kokonaisuus…

27

Lopuksi: kurssin sisältö



28

Lopuksi: kurssin sisältö

- relevanssin käsite, tiedonhakujärjestelmien evaluointi
- dokumenttikokoelmien indeksointi
- kyselyjen käsittely, täsmäytysmenetelmät
- dokumenttien ryvästäminen
- merkkijonohahmon etsintä selaamalla
- luonnollisen kielen käsittely tiedonhakua varten, monikielinen tiedonhaku
- tiedonhakustrategioita, käyttöliittymäratkaisuja

29

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
- seminaari syksyllä: Modernit tiedonhakumenetelmät (Juha Puustjärvi)

30

Loppujärjestelyistä

- harjoitustöiden esittely ensi viikon harjoituksissa (6.-7.5.)
  - jokaisen ryhmäläisen tulisi olla paikalla
  - harjoitustilaisuudessa ryhmät sekoitetaan uusiksi ryhmiksi, joissa kukin selostaa muille oman ryhmänsä työtä ja tuloksia
  - lopuksi valitaan yksi alkuperäinen ryhmä, joka esittelee työtään kaikille, jolloin voidaan vielä keskustella yhdessä

31

Loppujärjestelyistä

- harjoitustyöraportti on palautettava pe 14.5. klo 24:00 mennessä, myöhästyminen verottaa pisteitä 2p/päivä
- ohjeet palauttamisesta tulevat ensi viikon alussa
- koe ma 10.5. klo 16-20, Päärakennus, Sali 1

32

Kiitos!!!

- osallistumisesta!
- hyvistä kommenteista ja keskustelusta!

33