10. WWW-tiedonhaku

= yhteenvedon luonteista WWW-tiedonhausta, kehityksestä

Lähteitä: BYRN, Ch. 13
IEEE Data Engineering Bulletin 23,3 (2000)
(special issue on Next-Generation Web Search)

10.1 WWW-tiedonhaun erityispiirteitä

- hajautettu järjestelmä: tiedot todella monessa pisteessä
  (miten otetaan huomioon?)
- tiedot hyvin muuttuvia
  haun tuloksessa usein vanhentuneita linkkejä
  entä sivujen sisältö?
  vrt. perinteisen kirjaston kokoelmat ...
- suuret, jatkuvasti kasvavat volyymit
- kokoelman struktuuri epäselvä, toisteista tietoa (jopa 30%)
  käsitemallina hyperteksti, todellisuus 'vapaampi'
  myös sivujen rakenne ja laatu vaihtelevat
- kokoelman luotettavuus huono
  'kaikkia mahdollisia' virheitä:
  vääriä tietoja, vanhaa, tyyliltään huonoa,
  viimeistelemätöntä
  (esim. joka kolmas nimi väärin kirjoitettu (?))

- heterogeenistä: eri muotoja, medioita, kieliä,
  jopa aakkostoja

WWW-dokumenttien tekstiominaisuudet:
sanasto suhteessa laajempi kuin paremmin kontrolloiduille teksteille,
sanojen frekvenssit vaihtelevat enemmän
(onko silti merkitystä?)

- käyttäjän ongelmat: kyselyn muoto, tuloksen tulkinta ja käsittely (suuri määrä, yksittäinen dokumentti voi olla suuri, ilman metatietoja?)

Perusasetelma silti sama kuin muille kokoelmille ...

<u>Hakukoneiden arkkitehtuureja</u>

1° Keskitetty järjestelmä
indeksointi yhdessä pisteessä
robotit hakevat tietoa verkon eri paikoista
käyttö yhden pisteen (indeksin) kautta
mirror-pisteitä (USA, Eurooppa, ...)

- riittääkö keskitetyn järjestelmän teho enää kauan?

- arviot hakukoneiden indeksien kattavuudesta (1998):
  15-50 % kaikista sivuista (yli 300 milj.)
  (nyt ainakin pari miljardia?)

  Indeksien koot (4/2001) 200..500..700 milj. sivua

  (Google palauttaa linkkiyhteyksien takia
  kaksinkertaisen määrän eli noin 1.3 mrd. sivua.)

  - eräiden indeksien kasvu 2000->2001 noin 100%

2° Hajautettu arkkitehtuuri

- monet hakukoneet käyttävät nyt jo yhteistä indeksiä
  (vähentää palvelinten kuormitusta, tietoliikennettä)

- usean keräilijän (gatherer) käyttö vähentää
  liikennettä; voi olla useita välittäjiä, joihin käyttäjät
  ovat yhteydessä (Harvest-järjestelmä)
  - toistopalvelimia esim. maantieteellisin syin
  - keräilijät toimivat järjestelmällisesti (periodein)

Käyttöliittymät

- kyselyn esitys, tuloksen esitys
- yleisesti yksinkertainen ja laajempi kyselymuoto

kehittyneempiä piirteitä (HotBot, NorthernLight)
and ja or, vaaditut, toivotut, kielletyt sanat, fraasit
otsikko, henkilö
useita kieliä
aikarajoja, joustavia aikailmaisuja ('viime viikko'
maantieteellisiä alueita, domain-alueita
valitut luokat
Arts, Computing, Education, Medicine, ...
sivusto- tai sivutyypit
Commercial, Educational, Military sites
learning, questions/answers, press releases
sivun sisältöön kuuluvia objektityyppejä
kuvamuotoja, audio, Java, ...
sanojen katkaisu, vartalo?
sivun asema rakenteessa (syvyys)
domain-kohtainen määrärajoitus

tuloksen esitysmuodolla muutama (tuttu) vaihtoehto,
tulosten järjestäminen ja ryhmittely
relevanssi, joskus aika

Sivujen haku verkosta (crawling)

- linkit syvyys- tai leveyssuuntaisesti
- rajoituksia syvyydelle
(eri tapojen edullisuus?)
- aikamääreitä: haku periodeittain
- myös sivun muutostiheyttä voidaan seurata
- sivuun osoittavien linkkien huomiointi
- lähetetyt sivut (seurauksena myös vaara indeksien manipuloinnista)

Hakurobottien vaikutus palvelimiin

- kuormittavat; tarpeen minimoida
- jotkut haluavat estää indeksoinnin kokonaan

Robotin toteuttajan 'etiketti' (Koster, 1994)

- tarvitsetko oman robotin?
- tunnistettavuus
- ei turhia hakuja

Robotin estoprotokolla

- palvelimella tiedosto .../robots.txt
voi sisältää kiellon tietyille roboteille tiettyyn hakemistohierarkian osaan tai tiedostoon

Esim. User-agent: *
Disallow: /

User-agent: *
Disallow:

User-agent: *
Disallow: /cgi-bin/
Disallow: /minun_omat/
Disallow: /minun/tama_vain.html

User-agent: WebCrawler
Disallow:
User-agent: *
Disallow: /

HTML:n META-tag:
<meta name="robots" content="noindex, nofollow">
- kielletään indeksointi, linkin seuraaminen

Indeksit

- käänteishakemistoja, erilaisia piirteitä

- tulokseen tiedot hakemistosta, siis esim. ensimmäiset rivit ym. tulokseen liittyvät indeksiin (+ koko, pvm, otsikko jne)

- indeksin koko esim. luokkaa 150 GB / 100 milj. sivua
(kuvaustiedot noin 500 tavua)

- lähekkäisyyskyselyt ja fraasit vaativat sanaesiintymien paikat
- hakemisto järjestyksessä: binäärihaku tyypillinen
- monta hakusanaa: hakutulosten yhdistäminen (vrt. tietokantatekniikka)
- koko tuloksen muodostus kerralla yleistä (usein käytetään vain ensimmäisiä sivuja!), myös 'lazy evaluation ...
- monimutkaiset haut (likimääräiset jne) vaativat helposti koko hakemiston selausta; liian hidas WWW:lle

Selaustyyppinen haku

- sellaisenaan liian hidasta (paitsi pienessä sivustossa, muutaman linkin verran)

- luokitteluun perustuvat hakemistot (esim. Yahoo)
- haun ja (hakemisto)selauksen yhdistäminen trendinä

- luokkahakemisto voi olla laaja
  (kymmeniä tuhansia luokkia)

- aiheenmukaisia luokkia, paikkaa tai aikaa ilmaisevia jne

Haun ja selauksen yhdistäminen:

- käyttäjä yksinkertaisesti vuorottelee ...

esim. WebGlimpse: search box (+ naapurisolmut jne)

- erinäisiä apuvälineitä:
  sivuston havainnollistaminen visuaalisesti,
  standardien puuttuessa ei globaalissa käytössä

Metahakukoneet

- kysely lähetetään monelle hakukoneelle (5..25) ,
  tulokset yhdistetään (esim. MetaCrawler)

- yksi käyttöliittymä
- pääsy eri indekseihin
  (eri hakukoneiden indeksien päällekkäisyys yllättävän vähäistä)
- tuloksen muodostamisessa erilaisia tapoja
  dokumenttien järjestys monen tekijän mukaan
  painotus: esim. monen koneen löytämät ensin

Yhteenveto

”finding the needle in the haystack”

Käyttäjän ongelmia:

- kyselyn sisältö ja muoto (vrt. käyttöliittymät edellä)
  - erillisten sanojen merkitys
  - and/or-vaikeudet (→ operaatioita käytetään vähän)
  - kirjoitusasun vaikeudet
    - isot/pienet kirjaimet, erisnimet, vieraskieliset sanat
  - synonyymi/homonyymi-ongelmat
    - Esim. pelit Shogi, Go:
      - Go vaikea löytää (merkitykset, Go/go)
  - haun tarkoitus vaikea kuvata
    - Esim. ’jaguar speed’
- tuloksen hahmottaminen
  - koko, järjestys eri tavoin epäselvä
  - yleensä ei voi tallentaa ja kokeilla joustavasti
    - (uusi selainikkuna ...)

Kyselyjen tekemistä on tutkittu:

- hakukoneen valinta:
  helppokäyttöisyys, nopeus, indeksin kattavuus,
  kokemus yleensä, kokemus tulosten relevanssista

- käyttötarkoituksia:
  tutkimus, vapaa-aika, kaupalliset, opiskelu

- kyselyt:
  25 %: yksi termi
  keskimäärin 2.5 termiä (max 393! – agentit vääristävät)
  kyselyä muunnetaan harvoin (20 %)
  64 % kyselyistä uniikkeja (?)

  tuloksen käsittely:
  85 % tyytyy tuloksen ensimmäiseen näyttöön (10)

Esim. (BYRN) Aihekohtainen haku (esim. tutkimusaihe):
- valitaan tiedossa oleva relevantti artikkeli
- haetaan sen tekijöiden nimillä (yleiset / harvinaiset!)
- saadaan uudempia viiteosumien kautta, ja
  muita mahdollisesti relevantteja
- saadaan tekijöiden kotisivut

- iteroidaan ...
  kovin heuristista!

Entä hakemistot?

- yleensä liian ylimalkaisia varsinaisen tuloksen löytämiseksi
  'not enough depth to find the needle'
  ('search engines return too much hay with the needle')

BYRNin yleisohje kyselytarpeen mukaan:

- tarkka haku (esim. termin määritelmä, selitys)
  käytä tietosanakirjaa
  (eli valmiiksi jäsennettyä tietolähdettä)
- väljä hakukohde (laaja tai tuntematon asia)
  käytä hakemistoa aihepiiriin pääsemiseen
  (sitten laajuudesta riippuen selausta tai tarkennettua hakua)
- epäselvä hakukohde
  käytä hakukonetta, varaudu muuntamaan kyselyä eli iteroimaan

10.2 Tulevaisuus

IEEE-artikkelin teemat:
- linkki-informaation käyttö: PageRank, HITS ym.
- sivun aiheen määritys epäsuorasti viittausten perusteella
- aihe-keskeiset haut
- kontekstin käyttö haun tukena
- hajautettu tiedonhaku: 'monitietokantamalli'

- Hearst: Next generation web search: setting our sites

1. Johdanto

Käyttäjien tyytyväisyys on aikaisempaa parempi (jopa 80%).
- vaikka kyselyt minimaalisen lyhyitä: yleensä 2 sanaa

Selitys?
- haku on tarkoitettu vain lähtökohdaksi;
  löytynyt sivusto tutkitaan selaamalla
  tai paikallisella haulla
  hakua tarkennetaan (refine)

Muita kehityspiirteitä:
- hakukoneet ilmaisevat tuloksen yhteyden johonkin luokitteluun

  Esim. 'horseradish'

  Recreation > Food
  Business > Industries > Food Products > Wegetables

- linkkiyhteyksien huomiointi: osoitetut sivut (sivustot) ovat hyviä lähtökohtia jatkaa

2. Sivustohaku, tehtävän merkitys

- näkökulma yksittäisen sivun sijasta sivustoon ja tehtävään

- käyttäjä selaa mielellään, jos sivusto tukee sitä
  sivulla on näkyvissä, mihin se liittyy
  seuraava/edellinen jne
  liikkuminen on teknisesti joustavaa

- sivuston suunnittelu (selaavaa hakua varten) on tärkeää;
  vaikeaa: käyttötapoja, -tilanteita paljon

  Kokeista: käyttäjä luulee toimineensa nopeasti, jos hän on voinut edetä mielekkäästi
  (vaikka sivut latautuisivat hitaasti - (?))

- metadatan merkitys: luokittelun luonteiset metatiedot auttavat erottamaan erityyppisiä dokumentteja
  tieteelliset artikkelit, reseptit, arkkitehtuurikuvat, ...
  metadatan haku on varsinaista dataa deterministisempää
  "kaikki luokkaan 'tekoäly' kuuluvat artikkelit" vs.
  "kaikki artikkelit, joissa jotain tekoälystä"

  luokittelun yhteydessä voidaan ilmoittaa luokkaan kuuluvien dokumenttien määrä (vaikuttaa selaamiseen)

  jos luokkia/dokumentteja on paljon search-tyyppisen paikallisen haun merkitys korostuu

  dynaamiset kyselyt, useita dimensioita
  esim. alue, aika (muiden tietojen ohella)

'related'-tyyppiset yhteydet
voivat olla hankaliakin: esim. ravintolaa etsittäessä
voidaan olla kiinnostuneita muista iltamenoista tai ei

- metadata voi olla 'dynaamista': luokat muodostetaan haun jälkeen (eräänlaista ryvästämistä)
- kaikki aihepiirit eivät ole hyvin jäsentyneitä

- haun tuloksen muodostus voisi olla dynaamista:
  alkuvaiheessa overview-dokumentteja
  "syvemmällä" vain erikoistuneempia

- tuloksen esitysjärjestyksiä:
  tekijöittäin, aihepiireittäin, populaarisuuteen perustuen (muut käyttäjät) jne

[yhteys Flamenco-projektiin ...]

3. Muita sivuhaun näkökulmia

- sivuston järjestäminen organisaation mukaisesti
  (periaatteessa aika yleistä ...),
  organisaation käyttö tuloksen esittämiseen

- Berkeleyn yliopiston sivut: esim. haku 'earthquake'
  tuloksen osat järjestetään osastokohtaisiin ryhmiin
  (sopii hyvin joskus, ei aina ...)

- Citeseer: tutkimusjulkaisut
  tulokseen liitetään viittaus-, related- yms. tietoja

- luonnollisen kielen käyttö, question/answering
  (alakohtaisena hyvä, yleisenä ?)

Hearstin yhteenveto:
- hakukoneet tulisi saada ottamaan käyttäjän tehtävä huomioon:
  ei kokoelmakeskeistä, vaan tehtäväkeskeistä

BYRN: trendejä, tutkimuskohteita:

- tiedonhaun mallintaminen (erityisesti WWW:lle)
  linkkirakenteet
  kyselyjen/suodatuksen yhteys

- kyselyt
  rakenteen ja sisällön käytön yhdistäminen
  visuaaliset piirteet
  luonnollinen kieli

- hajautetut arkkitehtuurit
  välttämättömyys volyymin kasvaessa
  palvelimet vai tiedonsiirto rajoitteena?

- tuloksen järjestelyn parantaminen
  relevanssi, muu erottelu

- indeksointi
  tiivistysmenetelmät

- dynaamisten sivujen haku
  'hidden Web' käyttöön

- toistuvan tiedon eliminointi
- multimediahaut
- käyttöliittymien kehittäminen
- selaustyyppinen haku
  erikseen, yhdessä kyselyjen kanssa