7. Informaation suodatus

= filtering (extraction, categorization,
routing, retrieval)

(Comm. ACM 35,12 (1992):
special section, 26-81)

7.1 Yleistä

Suodatus = information valintaa ennalta
määriteltyjen kriteerien perusteella
= valikointia informaatiotulvan
kontrolloimiseksi

- uutisryhmät, WWW; myös radio, tv, ...
- henkilökohtaiset informaatiolähteet:
postituslistat, sähköposti yleensä

Kyselyt kertaluonteisia,
suodatus ‘jatkuvaa', taustalla

- samantapaisia kriteerejä
hakutermit = suodatusehdot

Suodatuksen piirteitä:

1° tieto vapaamuotoista tai
puolirakenteellista

(vrt. tietokanta: tietue, kenttä (rakenne))

Esim. sähköposti:

From: joku
Subject: suodatustekniikat
Message body: .....

- vast. bibliografiset tiedot:
tekijä, nimi, tiivistelmä, koko teksti (?)

- täysin vapaamuotoista:
tekstikentän sisältö

2° suodatuksen kohteena yleensä teksti
(text filtering, text categorization)

- myös multimediaa;
ominaisuuksien kuvailu vaikeaa
(tunnistus selaamalla helppoa;
mutta tiedämmekö, miten se tapahtuu?)

3° tiedon määrä todella suuri
MB ... GB
n viestiä /vrk

4° suodatuksen vaihe (milloin?)

tietojen vastaanotto
tietojen lähetys
(ohjaus halutuille tahoille, paikkoihin)
tietojen haku (agentit)

5° edellytys: suht. pysyvä tiedontarve

- kuvataan käyttäjän (haku)<u>profiilina</u>
≈ kysely: termejä, fraaseja,
kuvaavat käyttäjän kiinnostusta

- profiilia voidaan muuttaa vähitellen
≈ relevanssipalaute,
oppiva järjestelmä

6° vaihtoehtoiset näkökulmat:

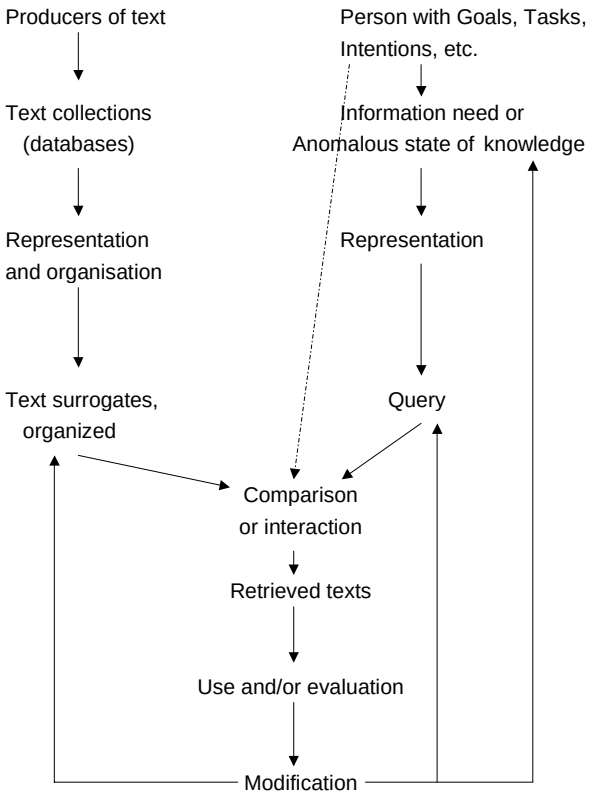
- tarpeellisten keräily
- tarpeettomien poisto, ohitus

Suodatus vs. tiedonhaku

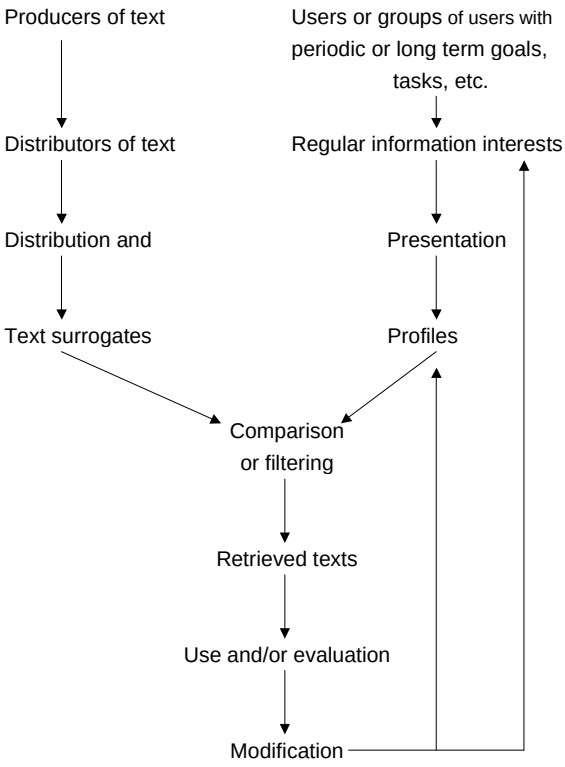
“two sides of the same coin?”

profiili	kysely
- käytössä jatkuvasti	- kertaluontoinen
- muuttuu hitaasti, oletetaan tarkaksi	- muodostamisen ongelmat tiedossa
- dokumenttien jakelu tärkeää	- dokumenttien organisointi, indeksointi
- ‘kokoelma' dynaaminen, informaatiovuo	- kokoelma staattinen, tietokanta
- valmisteltua, pitkäkestoista	- erillinen toiminta, mahdollisesti vuorovaikutteista

Tiedonhaun yleinen malli



Informaation suodatuksen yleinen malli



(Suodatus vs. tiedonhaku)

- tiedon ajantasaisuus tärkeää — vähemmän tärkeää (?)
- 'haetaan' uutta — - haetaan myös vanhaa
- käyttö vaihtelevaa, tarpeet epäselviä — - käyttäjät asiantuntijoita, motivoituneita
- haetaan 'varmuuden vuoksi', myöhemmin — - tulos tarvitaan yleensä heti

tutkittavaksi

- mallit (ks. kuvat)

Suodatus vs. selaus ?

- kiinnostus ilmaistaan etukäteen muodollisesti — vähitellen vapaamuotoisesti (valitsemalla)

7.2 Suodattimien muodosta, laadinnasta

- kyselyehdot
- ehdot ja toiminnot (muitakin kuin valitse)

Esim.

help(msg) → not_read(msg)

about(SE,msg) → not_read(msg)

about(printer,msg) AND not_have(printer)
→ not_read(msg)

about(LC,msg) AND not_know(LC)
AND short(msg) → read(msg

about(LC,msg) AND not_know(LC)
AND long(msg) →save(LC_folder,msg)

- yleisesti:
  tiedon laji, tyyppi, muu ominaisuus
  monimutkaisuus, koko
  lähettäjä
  käyttäjän tilanne (tavoite)

Suodattimien henkilökohtaisuus?

- henkilökohtaiset profiilit
- stereotyypit
  riittävät ehkä jatkuvassa käytössä?
  lähtökohta uusille käyttäjille

Suodatuksen tiedonhakumalli?

1° Boolen-tyyppinen ‘exact match'
  - periaatteessa yksinkertainen

Esim.
  “tänä vuonna jossakin USA:n yliopistossa julkaistut raportit, joissa käsitellään virtuaali-
  todellisuutta”

- samat ongelmat kuin kyselyissä (tiedonhaussa):

  tuloksessa ei järjestystä
  tuloksen koko vaihtelee rajusti

2° ‘best match' : järjestetty tulos

  esim. “15 parasta”

  - ongelma: päätös olisi tehtävä tietovirtaa käsiteltäessä
    - mihin verrataan?
    jonkin käsittelyerän sisällä
    profiiliin, johonkin normiin
    (suhteellinen järjestys ok, valinta ?)

  - relevanssipalautteen käyttö:
    relevantin dokumentin kanssa läheiset valitaan
    - tässäkin tuloksen koko ongelma
  - läheisyys:
    vektorimalli
    relevanssin todennäköisyyteen perustuvat menetelmät

  - LSI: Latent Semantic Indexing
    - otetaan huomioon termien piilevät semanttiset yhteydet

7.3 Suodatuksen käytäntö, erikoistapauksia

- esimerkkijärjestelmiä Usenet-käyttöön

SIFT (Stanford)
  - käyttäjällä voi olla monta profiilia
  - vektorimalli termipainoin ja Boolen malli
  - käyttäjä määrittää samanlaisuus-kynnyksen
  - valitut artikkelit käyttäjälle by mail (tehokkaammin: käyttäjien ryhmittely, lähetys palvelimelle ...)

NewsClip
  - merkinnät .newsrc-tiedostoon (pois-valitut luetuiksi)
  → lukeminen normaaliohjelmilla

Browse
  - suodatus viestin alkuosan (300 sanaa) perusteella
  - toteutus neuroverkkotekniikalla
  - käyttäjä lisää huonoja/hyviä termejä

7.4 Sosiaalinen suodatus

  - yhteistyöhön perustuva, ‘kollaboratiivinen'

Miksi?
- hyvien profiilien rakentaminen vaikeaa
- artikkelit eritasoisia; ‘match' ei kerro mitään artikkelien laadusta

Muiden lukijoiden valintojen hyväksikäyttö?

Oletuksia:

- käyttäjien profiileissa yhteisiä piirteitä
- käyttäjät valmiita jakamaan kokemuksia
- käyttäjät (ehkä) valmiita jopa arvioimaan artikkeleita eksplisiittisesti

⇒ suodatus voidaan perustaa sille, mitä muut ovat valinneet tai arvioineet

  - arviointi → suosittelujärjestelmät (recommendation systems)

Mitä tietoa voidaan käyttää hyväksi (implisiittisesti)?

- lukemiseen käytetty aika
- dokumentin tallennus, tulostus; poisto
- viittaus, sitaatti
- vastaaminen, kommentin kirjoitus
- toistuva valinta (esilleotto)
- merkintä kiinnostavaksi
- dokumenttien välinen (epäsuora) yhteys

Eksplisiittiset arviot:
- arvo, hyöty tms. (esim. 1-5)
- sanalliset tms. annotaatiot

Henkilösidonnaisuus?
- anonyymisuudella hyvät puolensa
- tietoisuus arvioijasta lisää arvion (tai muun tekijän) merkitystä
- asiantuntijuuden hyväksikäyttö ($\Rightarrow$ vahvaa yhteistyötä ...)

- henkilöaspekti: myös lisäongelmia
  käsitysten pysyvyys? (oppiminen ...)
  subjektiivisuus

Esim. GroupLens (1994)

- kohteena Usenet News
- lukijat arvioivat artikkelit 1-5
- lukuaika (optionaalinen)

- arviot yhdistetään se. lukijoiden (arvioiden) läheisyys vaikuttaa

Esim.

Artikkeli	Ken	Lee	Meg	Nan
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	?

Ken(6) $\leftarrow$ 4.6 Nan(6) $\leftarrow$ 3.7

- tulos voi vaikuttaa suoraan valintaan
- arviot voidaan näyttää; käyttäjä valitsee (selaus eikä haku ...)

PHOAKS
"People Helping One Another Know Stuff"
(CACM 1997)

- WWW-resurssien valinta Usenet-artikkelien avulla

- URL:n maininta uutisartikkelissa on hyvä suositus, jos
  - artikkeli on riittävän uniikki (ei cross-posted)
  - URL esiintyy tekstissä (ei vain lähettäjän viitteenä)
  - URL ei esiinny vanhan artikkelin kopiossa
  - URL:n ympäristö tekstissä viittaa sisältösuositukseen (eikä esim. mainokseen

- vain yksi maininta/henkilö hyväksytään (ei manipulointia)