1. Johdanto

IR (Information Retrieval;
Information Storage and Retrieval)

Tiedonhaku =
prosessit, jotka liittyvät tiedon
· esittämiseen
· organisointiin
· tallentamiseen
· etsimiseen

Tiedonhaun kohde:
· relevantti tieto (tiedontarve, merkitys)
· yksikkönä
· dokumentti, 'teksti'
sisältö; rakenne, ulkoasu
tyyppi: myös kuva, ääni, ...

Esim. tieteellinen artikkeli, uutinen,
mail-viesti, WWW-sivu,
palvelu (aikataulu tms.)

Näkökulma 'dokumentti kohteena' on yksinkertaistettu:
- www-sivu tai -sivujoukko
- monimutkainen tiedontarve

BYRN: haetaan jonkin USAn yliopiston tennisjoukkueen sivua siten, että
· joukkue osallistuu säännöllisesti NCAA-turnauksiin
· sivu(i)lla on tiedot joukkueen sijoituksista ainakin kolmen vuoden ajalta
· sivu(i)lla on valmentajan yhteystiedot

- epämääräinen, huonosti tiedostettu tiedontarve

"jotain tiedonhaun menetelmistä"

Tieto (data) ... informaatio

data
dokumentissa esiintyvät avainsanat
datalla ja kyselyllä tarkka struktuuri
hakutuloksessa ei voi olla virheitä
vrt. tietokantahaku

informaatio
tiedontarpeen täyttävä informaatio
kysely voi olla luonnollista kieltä
hakutulos voi olla epätarkka
järjestelmä pyrkii tulkitsemaan hakutulosta, esimerkiksi järjestämällä dokumentit

=> tiedonhaun perusasetelma:
haetaan kaikki käyttäjälle relevantit, mutta mahdollisimman vähän muita!

Erilaisia hakutilanteita:

· aihehaku

esim. 'julkisen talouden supistamistoimet'
'lomakohteet, joissa golf, hiihto, ratsastus, uinti mahdollista'

· yksilöhaku

VN:n päätös nro 123/97
Esko Ukkosen väitöskirja

· faktahaku

HY:n rehtori v. 1956

- vaihtelua:
tuloksen koko, luonne
käyttäjän oletus (varmuus) tuloksen olemassaolosta ja löytymisestä

Erityisesti aihehakua halutaan usein tarkentaa:
- johdattelevat IR-artikkelit
- ajankohtaiset, uusimmat, …
- kaupalliset / tieteelliset tms. jakoja

- myös kognitiivisia tekijöitä:
  erilaisia hakustrategioita:
    kerralla, toistaen (,selaten)
  asioiden jäsentely:
    haku luokitukseen perustuen

Tiedonhaku ↔ tietokannan hallinta ?

tietokanta

· homogeeniset tietueet
· attribuutit, yksilön täsmällinen kuvaus
esim. relaatio 'henkilö(hetu, nimi, ...)'

· haun tulos yksikäsitteinen

dokumenttien joukko (dokumenttikanta)

· teksti
· asiasanoja, indeksitermejä
· luokituksia UDK, CR
· muita kuvaajia

· haun tulos epätarkka,
usein järjestetty (sopivimmat)

'dokumentit, joissa on tietoa kissan hoidosta'

Miksi epätarkkaa?

- dokumenttikannan kattavuus?
- sanastoerot
- epätäsmällisiä käsitteitä
  nuori, vaalea, räikeä, paljon, ...
- ehtojen ilmaisemisen vaikeus, likimääräinen tulkinta
  ilmestymisvuosi > 1985;
  entä 1985, 1984 ...

Tietokanta ↔ dokumenttikanta?

- myös dokumenttikantaa sanotaan usein tietokannaksi
- www-sivujen joukkoa voidaan sanoa dokumenttikannaksi (vaikka se on monessa mielessä epämääräinen)

Dokumenttikanta

= dokumenttien Di kokoelma, joukko

Di = (t1, t2, t3, ...) termien joukko

Di = (Kissa kulki aurinkoisena päivänä kuumalla katolla ja poltti käpälänsä)
Di = (kissa, katto, aurinkoinen, käpälä)

(*polttaa, kuuma?*)

Termijoukko = dokumentin **looginen näkymä** (logical view), joka saadaan useilla muunnoksilla:

· poistetaan turhat välimerkit yms.
· poistetaan täytesanat (stopwords)
· rajoitutaan substantiiveihin (ehkä)
· otetaan käyttöön sanojen vartalot ("stem")
· indeksoidaan dokumentti termien avulla
· kuvataan dokumentin struktuuri (mahdollisesti, lisäpiirre)

· indeksi = hakemisto,
esim. dokumentin sanoihin perustuva käänteishakemisto

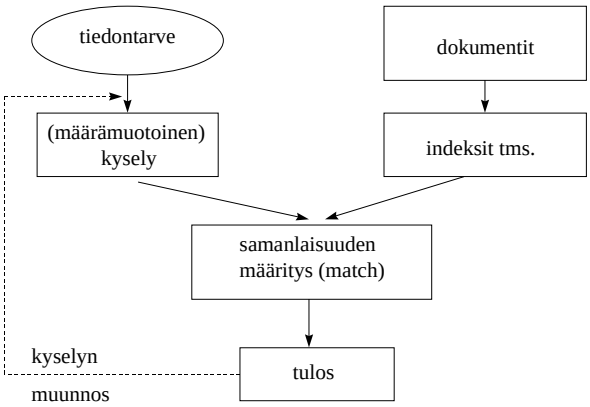
· indeksointi on kertaluontoinen tai harvoin toistuva, joka tapauksessa suurta kyselyjoukkoa palveleva (raskas) toiminto

Indeksin esikuvia:
- sisällysluettelo
- sanaluettelo (indeksi) kirjan lopussa (kattavuus vaihtelee …)

· manuaalinen / asiantuntijan indeksointi
asiantuntemus, tehokkuus

Kyselyn toteutus

· tietokanta: kysely (kyselykielellä)
evaluointi, toteutus
tulokset
· tiedonhaku: kysely
evaluointi (‘match’)
tulokset käyttäjälle
käyttäjä arvioi



· kyselyn muunnos voi olla ainakin osittain automaattinen: käyttäjän arvioon perustuva ‘relevance feedback’

· myös muut muunnokset mahdollisia: termien synonyymit ym. yhteydet voidaan ottaa huomioon

WWW:n erityispiirteitä:

· dokumentit kovin erilaisia
verrattuna esim. perinteiseen kirjastoon
- laatukriteerien puute
- erilaisia rakenteita
· puuttuu yhtenäinen tietomalli,
johon esim. haut voitaisiin perustaa
· selaaminen l. navigointi luontainen käyttötapa, laajasti hyvin tehoton
→ tarvetta tiedonhaun menetelmille
(kehittymässä, ei valmista)

(selaaminen <-> tiedonhaku?
eri menetelmät, sama yleinen tavoite)

Tiedonhaun evaluointi

= aihehaun ‘onnistumisen’ mittaaminen

Haun tavoite: tuloksessa
‘ALL and ONLY’ halutut (relevantit)

Relevanssi
· jakaa tulokseen kuuluvat dokumentit hyödyllisiin ja ei-hyödyllisiin

Erilaisia näkökulmia/määrityksiä:

· aiherelevanssi
yhteenkuuluva (related)
aiheenmukainen (topical)
vastaava, osuva, soveltuva
(responsive, pertinent)

· käyttäjärelevanssi
hyödyllinen (useful, beneficial)
käyttökelpoinen (utility ...)
tyydyttävä (user satisfaction)

objektiivinen / subjektiivinen käsite ?

Aiherelevanssin ja käyttäjärelevanssin erot?

Haun onnistuminen, perustunnusluvut:

saanti (recall)
tarkkuus (precision)
tulosjoukon koko
vasteaika

(+ muut?)

Haun tulos	**Relevanssiarvio**	
	relevantti	**ei relevantti**
Löydetyt		
a + b	**a**	**b**
Hylätyt		
c + d	**c**	**d**
Summa		
a+b+c+d	**a + c**	**b + d**

saanti = a / (a + c) (‘kattavuus’)

tarkkuus = a / (a + b)

· tarkkuus helpompi laskea kuin saanti

· ei-löydetyt (hylätyt) on arvioitava

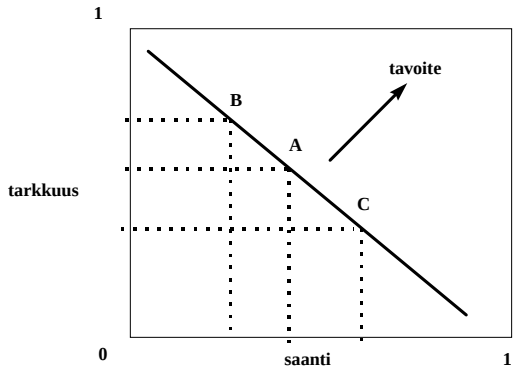
· ALL and ONLY - periaate:
saanti = 1, tarkkuus = 1

Käytännössä:

arvot lähempänä 0.5 kuin 0.9 ...
parempi saanti <=> huonompi tarkkuus
haun kaventaminen (narrowing)
haun laajentaminen (broadening)

Mikä vaikuttaa?

· indeksoinnin johdonmukaisuus
· dokumenttien pää- ja sivuteemat
· kyselyn termivalinnat



Muita laskennallisia mittareita:

häly (noise) = b / (a + b)

unohdetut (fallout) = c / (a + c)

oikein hylättyjen osuus = d / (a+b+c+d)

Muita onnistumisen / laadun kriteerejä:

· dokumenttikanta

kattavuus (eri tasoilla, osissa,...)
ajantasaisuus
tietojen laatu
suodatuskyky indeksointi jne
käyttöominaisuudet

· hakujärjestelmä

kyselyjen muoto
muut käyttöominaisuudet
suorituskyky (vasteaika)
tulosten käytettävyys (muoto jne)

Määritelmiä:

- termi = semanttinen ilmaisun yksikkö,
  esim. sana, fraasi, sanan vartalo

- dokumentti = haun kohde,
  loogisesti termien jono

- kysely = termien jono
  ilmaisee tiedon tarpeen
  (vertailuun sopivassa muodossa)

- relevantti dokumentti:
  (käyttäjän kannalta) kyselyyn sopiva,
  merkityksellinen dokumentti

- samanlaisuusarvo (similarity, score):
  dokumentin relevanssin mittaluku (arvio)

- tiedonhakujärjestelmä (IR system):
  valitsee ‘kyselyyn sopivat’ dokumentit
  usein relevanssijärjestyksessä

Menetelmiä:

- Boolen lausekkeina annetut kyselyt
  yksinkertainen termivertailu

- vektorimalliin perustuvat
  lasketaan lähekkäisyys termiavaruudessa

- todennäköisyysmallit

Aputekniikkoja:

- ‘stemming’: sanat perusmuotoon

- merkkijonojen vertailu: sanat, n-grammit

- semanttiset yhteydet
  synonyymit
  thesaurukset

- esiintymäyhteydet
  termien lähekkäisyys
  lauserakenteet
  dokumentin rakenteen käyttö (osat jne)

Tiedonhaun tietokantatyyppejä

· lähdetietokannat vs. korviketietokannat
(source) (surrogate)

Lähdetietokantoja
· tekstitietokanta full text, free text
· faktatietokanta
· kuvatietokanta
· ohjelmistotietokanta
· hypermediatietokanta
...

Korviketietokantoja
· viitetietokanta
· hakemistotietokanta

- perinteinen IR: kohteena
  (bibliografinen) viitetietokanta

- nykyisin tekstitieto tärkein

Tekstitietokanta

esim. uutiset, tieteelliset artikkelit,
lakitekstit
· yleensä rakenteetonta
· mukana bibliografisia kuvailutietoja
· käyttö yleensä helppoa,
vaikka kyselyt sumeita

Faktatietokanta

· tilastot, muut
· mittaustuloksia, tosiasioita
· kyselyt yleensä täsmällisiä
(mutta ei triviaaleja)

esim. suomalaisten televisioiden
markkkinaosuus Euroopan
eri maissa v. 1993?

BKT henkeä kohti (virallisempi)

Kuvatietokanta

· digitoidut kuvat + kuvailutiedot
· haku kuvailutietojen kautta t.
kuva-aiheiden perusteella

Ohjelmistotietokanta

· fakta- ja tekstikannan piirteitä

Hypermediatietokanta

· rakenteellinen erikoistapaus
tärkeä: esim. WWW
(WWW-sivuilla myös omat,
sisältöön liittyvät tyyppinsä!)
· haku kyselyjä, selaamista

Viitetietokanta

· tieto osittain rakenteista
myös tekstiä
· vain ‘välillinen’ merkitys

Hakemistotietokanta

· tieto pääasiassa rakenteista
· välillinen merkitys