# Tiedon louhinta ja tietotulva

Hannu Toivonen
Tietojenkäsittelytieteen laitos
Hannu.Toivonen@cs.helsinki.fi

1

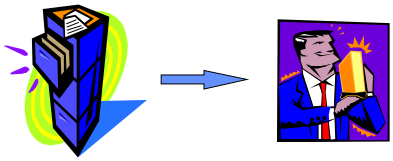
## Informaatiotulva

- Vuoden 2000 aikana tuotettiin 3 exatavua dataa
  - kilotavu = 1024 tavua
  - megatavu = 1024*1024 tavua
  - ...giga, tera, peta...
  - exatavu = $1024^6$ tavua $\approx 10^{18}$ tavua
- Datan määrä kaksinkertaistuu vuosittain
- Informaatiota esitetään moninaisissa muodoissa
  - relaatiotietokannat
  - teksti (Google-hakukone tuntee 4.3 miljardia sivua)
  - mittaus- ja lokitietokannat
  - geneettiset aineistot (ihmisen dna: 3 miljardia emäsparia)
  - ...

2

## Tiedon louhinta

- Uuden ja hyödyllisen tiedon päätteleminen suurista datamassoista

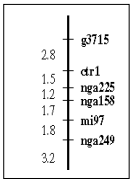


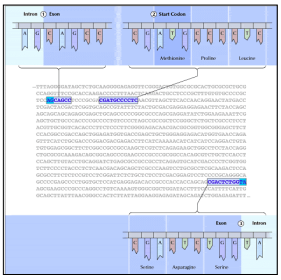
- "Moderni data-analyysi" tai "algoritminen tilastotiede"
- "Mitä data kertoisi, jos siltä osaisi kysyä oikeat kysymykset?"

3

## Sairausgeenien paikannus

Alttiusgeenin sijainti paikannetaan pienelle alueelle



Pieni alue tutkitaan tarkasti

- Tiedon louhinnalla voidaan havaita esim. yhteys kromosomin tietyn kohdan ja sairauden välillä

4

## Assosiaatiosäännöt

- Alkuperäinen ongelmatyyppi: mitä tavaroita ostetaan usein yhdessä?
  - Ostoskorianalyysi
  - `Jos vaippoja niin olutta (todennäköisyys 56 %, frekvenssi 12 %)`

5

## Assosiaatiosäännöt

- 1. yleistys: mitkä asiat esiintyvät usein yhdessä? Erilaisia sovelluskohteita:
  - kurssi-ilmottautumiset
  - `Jos tietoliikenne ja UNIX-ohjelmointi niin C-ohjelmointi (tod.näk. 72 %, frekv. 6 %)`
  - tekstidokumenttien analysointi
  - `Jos "www" ja "netscape" niin "browser" ja "internet" (tod.näk. 89 %, frekv. 0.12 %)`
  - geneettisten markkerit ja perinnöllinen sairaus
  - `Jos "marker9" ja "marker33" ja "tupakoi" niin "sairas" tod.näk. 34 %, frekv. 8 %)`

6

## Assosiaatiosäännöt

- Tavoitteena on **kuvailla** mahdollisesti mielenkiintoisia yksinkertaisia ilmiöitä
- Menetelmä tuottaa **kaikki** assosiaatiosäännöt, joilla frekvenssi > kynnysarvo
- Mahdollisia sääntöjä on valtavasti, läpikäynti käsin olisi mahdotonta
- Joukossa voi olla yllättäviäkin sääntöjä
- Tiedon louhintaprosessiin liittyvä ongelma: miten autetaan käyttäjää löytämään juuri häntä kiinnostavat säännöt?
- Menetelmä on sovellusriippumaton

7

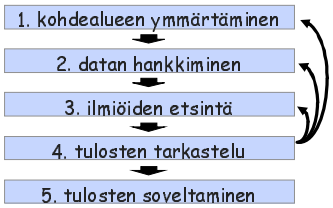
## Muita tiedon louhinnan muotoja

- Säännönmukaisuuksien etsintä
  - Millaiset hahmot ovat aineistossa tyypillisiä?
- Klusterointi
  - Millaisiin luonteviin ryhmiin aineiston voi jakaa?
- Luokittelu, ennustaminen
  - Miten havaintojen tietyn ominaisuuden voi ennustaa havainnon muista ominaisuuksista?
- Poikkeuksien etsiminen
  - Mitkä havainnot vaikuttavat poikkeuksellisilta?

8

## Tiedon louhintaprosessi

- Tiedon louhinnassa tutkitaan algoritmien lisäksi myös koko analyysiprosessia




9

## Tutkimusmenetelmä

- Esimerkkinä assosiaatiosäännöt
- Algoritmikehitys:
  assosiaatiosäännöt
  → episodisäännöt (assosiaatiot tapahtumajonoissa)
  → yleinen menetelmärunko
- Teoreettinen kehitys:
  konkreettinen ongelma (ostoskorianalyysi)
  → yleistetty ongelmatyyppi (toistuvat ilmiöt)
  → tehtävätyypin ja ratkaisuvaihtoehtojen analyysi

10

## Assosiaatiosäännöt

- 2. yleistys: mitkä hahmot esiintyvät aineistossa usein?
  - Syöte
    - *r:* tietokanta
    - *P:* suuri joukko hahmoja tai hahmojen "kieli"
    - *k:* yleisyyden kynnysarvo
  - Tulos
    - kaikki joukon *P* hahmot, joiden yleisyys ylittää kynnysarvon *k* tietokannassa *r*
  - Analyysi
    - ongelman teoreettisista ominaisuuksista seuraa, että tietty yksinkertainen algoritmi on tehtävään optimaalinen (tietyillä oletuksilla)

11

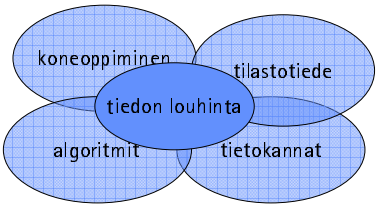
## Tiedon louhinta tieteenalana

Tutkimuskohteita:

- Louhinta-algoritmien suunnittelu ja analyysi
  - miten annettu data-analyysitehtävä ratkaistaan?
- Tiedon louhinnan teoria
  - millaisia tehtävätyyppejä ja millä edellytyksillä tietyllä algoritmilla voidaan ratkaista?
  - millaisia ominaisuuksia eri tehtävätyypeillä on?
  - miten tulosten laatua voidaan arvioida?
- Tehtävätyyppien muotoilu, "hyvät kysymykset"
  - millaiset data-analyysitehtävät ovat yleiskäyttöisiä?
- Tiedon louhintaprosessi
  - mitkä ovat ne toimintatavat, joilla uudelle ongelmalle löydetään hyvät kysymykset ja niille hyvät ratkaisut?

12

## Tiedon louhinnan lähinaapurit




Tiedon louhinta:
- automatisoitu analyysi, algoritmit
- hahmojen ("hypoteesien") löytäminen
- suurten datamassojen käsittely
- tavoitteena ymmärryksen lisääminen

13

## Millaisista taidoista on hyötyä

- algoritmiikka
- todennäköisyyslaskenta
- tilastotiede
- tietokannat (??)
- koneoppiminen
- sovellusalueen tuntemus
  - poikkitieteellisillä taidoilla iso tutkimuspotentiaali
- tiedon louhinta ei ole helppoa: jokainen ongelma vaatii luovuutta ratkaisujen kehittämisessä ja soveltamisessa

14

## Miksi? Miksi juuri nyt?

- Soveltajat: dataa on, samoin taloudellisia tarpeita
  - Datan kerääminen ja tallettaminen on helppoa
  - Ensisijaiset tarpeet on tyydytetty (tapahtumankäsittely, talletus, yhteenvedot)
  - Tietovarastot (data warehouse)
  - Usein louhitaankin muita tarkoituksia varten kerättyä tietoa
  - Tieteelliset ja taloudelliset mahdollisuudet
- Tietojenkäsittelijät: uusia mielenkiintoisia ongelmia
  - Uusia ongelmatyyppejä...
  - ...joitten teoria on vasta muodostumassa
  - Hyvät lähtökohdat lähitieteistä
  - Poikkitieteellisen yhteistyön mahdollisuus

15

## Tiedon louhinta ja TKTL

- Tiedonhallinnan (ent. informaatiojärjestelmien) linja
  - geenikartoitusmenetelmät, geneettisen datan analyysi
  - ekologiset data-analyysiongelmat (mm. ilmaston rekonstruointi)
  - hahmokielet, algoritmikehitys
  - tekstien ja dokumenttirakenteiden louhinta
  - mobiililaitteiden tilannetietoisuus
- mukana laitoksella toimivissa "virtuaaliorganisaatioissa"
  - FDK-huippuyksikkö (From Data to Knowledge)
    - tiedon louhinnan ja hahmonsovituksen "kattoprojekti"
  - HIIT/BRU
    - data-analyysi, proaktiivinen laskenta

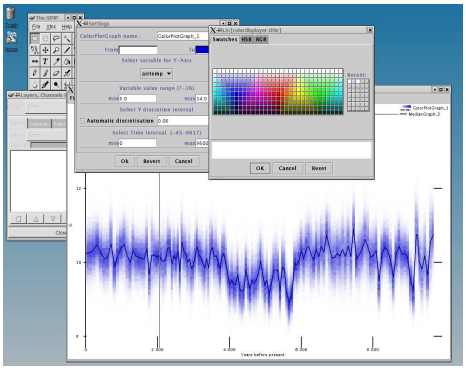
16

## Syövälle altistavien yhdisteiden tunnistaminen


- Kansainvälinen "haastekilpailu" tiedon louhijoille
  - järjestäjänä mm. NIH Yhdysvalloista
  - todellinen sokkotesti
- Mallinnus- ja ennustusongelma
- Mallien ja tulosten testaaminen ja arviointi

17

## Ilmaston rekonstruointi



18

## Yhteenveto tiedon louhinnasta

Tiedon louhinta tieteenalana

- tuottaa ja tunnistaa erilaisia datan automaattiseen analysointiin ja kuvailemiseen liittyviä tehtävätyyppejä tai lähestymistapoja
- analysoi ja kategorisoi niitä
- kehittää niihin tehokkaita ratkaisuja
- myös: tietosuoja ja etiikka tiedon louhinnassa

Laitoksella kansainvälisesti korkealaatuista tutkimusta

Runsaasti tieteellisiä yhteistyöprojekteja

19