Informaatiojärjestelmät, tietotulva ja tiedon louhinta

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

Informaatiotulva

- Vuoden 2000 aikana tuotettiin 3 exatavua dataa
  - kilotavu = 1024 tavua
  - megatavu = 1024*1024 tavua
  - ...giga, tera, peta...
  - exatavu = 1024^6 tavua ≈ 10^{18} tavua
- Datan määrä kaksinkertaistuu vuosittain
- Informaatiota esitetään moninaisissa muodoissa
  - relaatiotietokannat
  - teksti (Google-hakukone tuntee 3.3 miljardia sivua)
  - mittaus- ja lokitietokannat
  - biologiset aineistot (ihmisen dna: 3 miljardia emäsparia)
  - ...

Informaatiojärjestelmät

- Informaation hallinta
  - tiedon tallettaminen
  - tiedon esittäminen
  - tiedonhaku
  - tiedon analysointi
- TKTL:n "info"-linja

Infon opetus

- *Johdatus sovellussuunnitteluun*
- *Tietokantojen perusteet*
- *Tietokannan hallinta*
- Digitaalisen median tekniikat
- Tutkimustiedonhallinnan peruskurssi
- Tiedon louhinnan menetelmät
- XML-metakieli

} cum laude

- *Tietokannan mallinnus*
- *Tietokantarakenteet ja -algoritmit*
- Tiedon louhinnan erikoiskurssi
- Tietovarastot
- Tiedonhakumenetelmät
- Rakenteisten dokumenttien käsittely
- seminaareja
- ...

} laudatur

Infon tutkimus

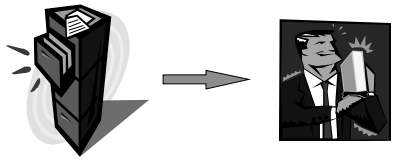
- Tiedon louhinta (Hannu Toivonen, Pirjo Moen, Marko Salmenkivi)
  - (tästä tarkemmin seuraavilla kalvoilla...)
- Dokumentit ja kieliteknologia (Helena Ahonen-Myka, Greger Linden)
  - rakenteisten dokumenttien hallinta, tiedonhaku, tiedon eristäminen tekstistä, tekstin louhinta
- Tietokannat (Seppo Sippu, Harri Laine, Jan Lindström, Juha Puustjärvi)
  - tietojen mallintaminen, samanaikaisen käytön valvonnan ongelmat, tietokantarakenteiden elvytys, tietokantasovellusten suunnitteluvälineet ja toteutusmenetelmät

Sisällysluettelo

- Informaatiotulva
- Informaatiojärjestelmien opetus ja tutkimus laitoksella
- Tiedon louhinta
- Esimerkkisovellus: sairausgeenien paikannus
- Esimerkkimenetelmä: assosiaatiosäännöt
- Tiedon louhinta tieteenalana
- Tiedon louhinnan tutkimus laitoksella
- Yhteenveto

Tiedon louhinta

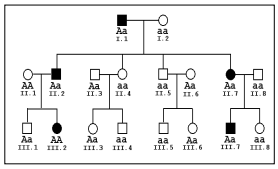
- Uuden ja hyödyllisen tiedon päätteleminen suurista datamassoista



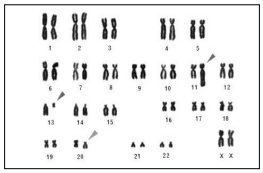
- "Moderni data-analyysi" tai "algoritminen tilastotiede"
- "Mitä data kertoisi, jos siltä osaisi kysyä oikeat kysymykset?"

7 TKTL_S2003.PPT

Sairausgeenien paikannus



1. Sukupuita tarkastelemalla havaitaan, että alttius sairastua tutkittavaan tautiin periytyy

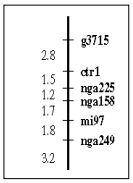


2. Perimästä paikannetaan kromosomi tai karkea alue, jolla altistava geeni on

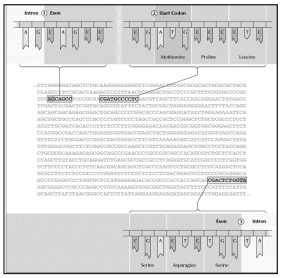
8 TKTL_S2003.PPT

Sairausgeenien paikannus

3. Alttiusgeenin sijainti paikannetaan pienelle alueelle



4. Pieni alue tutkitaan tarkasti



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

9 TKTL_S2003.PPT

Assosiaatiosäännöt

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

10 TKTL_S2003.PPT

Assosiaatiosäännöt

- 1. yleistys: mitkä asiat esiintyvät usein yhdessä?
  - 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 %)

11 TKTL_S2003.PPT

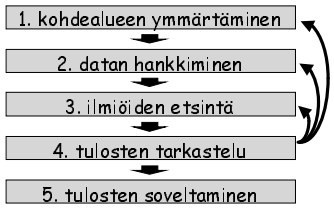
Assosiaatiosäännöt

- 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

12 TKTL_S2003.PPT

Tiedon louhintaprosessi

- Tiedon louhinnassa tutkitaan algoritmien lisäksi myös koko analyysiprosessia



13 TKTL_S2003.PPT

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*

14 TKTL_S2003.PPT

Tutkimusmenetelmä

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

15 TKTL_S2003.PPT

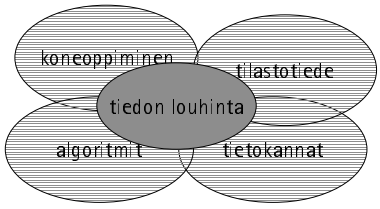
Tiedon louhinta tieteenalana

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?

16 TKTL_S2003.PPT

Tiedon louhinnan lähinaapurit



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

17 TKTL_S2003.PPT

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

18 TKTL_S2003.PPT

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

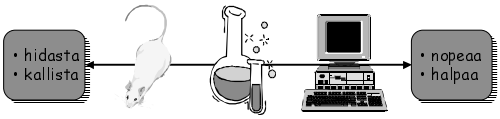
19 TKTL_S2003.PPT

Tiedon louhinta ja TKTL

- Informaatiojärjestelmien linja
  - geenikartoitusmenetelmät, geneettisen datan analyysi
  - ekologiset data-analyysiongelmat (mm. ilmaston rekonstruointi)
  - hahmokielet, algoritmikehitys
  - tekstien ja dokumenttirakenteiden louhinta
  - Hannu Toivonen, Helena Ahonen-Myka, Pirjo Moen, Marko Salmenkivi, ...
- mukana laitoksella toimivissa "virtuaaliorganisaatioissa"
  - **FDK-huippuyksikkö** (From Data to Knowledge)
    - tiedon louhinnan ja hahmonsovituksen "kattoprojekti"
  - **HIIT/BRU**
    - data-analyysi, proaktiivinen laskenta
- muita laitoksen ryhmiä
  - koneoppiminen (Jyrki Kivinen, Tapio Elomaa)
  - informaatioteoreettiset ja bayesilaiset menetelmät (Henry Tirri, Petri Myllymäki)

20 TKTL_S2003.PPT

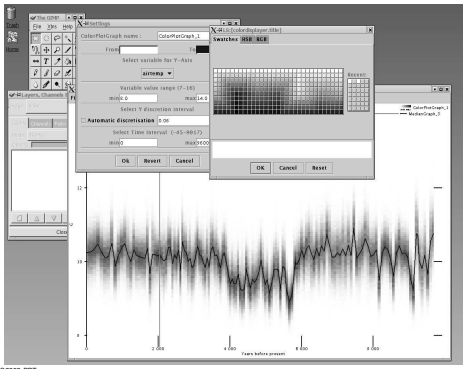
Syövälle altistavien yhdisteiden tunnistaminen



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

21 TKTL_S2003.PPT

Ilmaston rekonstruointi



22 TKTL_S2003.PPT

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

23 TKTL_S2003.PPT