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 $\approx 10^{18}$ tavua
- Datan määrä kaksinkertaistuu vuosittain
- Informaatiota esitetään moninaisissa muodoissa
  - relaatiotietokannat
  - teksti (Google-hakukone tuntee 2.5 miljardia sivua)
  - mittaus- ja lokitietokannat
  - geneettiset aineistot (ihmisen dna: 300 teratavua)
  - ...

Informaatiojärjestelmät

- Informaation hallinta
  - tiedon tallettaminen
  - tiedon esittäminen
  - tiedonhaku
  - tiedon analysointi
  - käyttöliittymät
- TKTL:n “info”-linja

Infon opetus

- *Johdatus sovellussuunnitteluun*
- *Tietokantojen perusteet*
- *Tietokannan hallinta*
- Käyttöliittymät
- Digitaalisen median tekniikat
- Tutkimustiedonhallinnan peruskurssi
- XML-metakieli

cum laude

- *Tietokannan mallinnus*
- *Tietokantarakenteet ja -algoritmit*
- Tietämyksen muodostaminen
- Tietovarastot
- Käyttöliittymät II
- Tiedonhakumenetelmät
- Rakenteisten dokumenttien käsittely
- seminaareja

laudatur

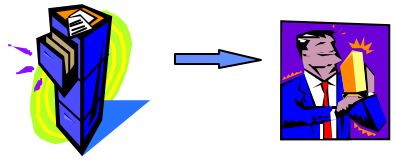
- ...

Infon tutkimus

- Tiedon louhinta (Hannu Toivonen)
  - (tästä tarkemmin seuraavilla kalvoilla...)
- Dokumentit ja kieliteknologia (Helena Ahonen)
  - rakenteisten dokumenttien hallinta, tiedonhaku, tiedon eristäminen tekstistä
- Tietokannat (Seppo Sippu, Harri Laine)
  - tietojen mallintaminen, samanaikaisen käytön valvonnan ongelmat, tietokantarakenteiden elvytys, tietokantasovellusten suunnitteluvälineet ja toteutusmenetelmät
- Käyttöliittymät (Sari Laakso)
  - graafiset käyttöliittymätekniikat, suunnittelumallit käyttöliittymien toteuttamisessa

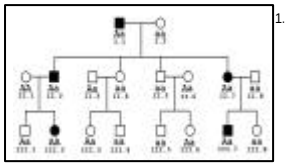
Tiedon louhinta

- Uuden ja hyödyllisen tiedon päätteleminen suurista datamassoista

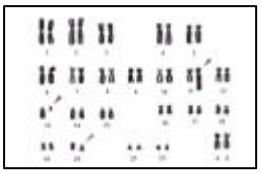


- "Moderni data-analyysi" tai "algoritminen tilastotiede"
- "Mitä data kertoisi, jos siltä osaisi kysyä oikeat kysymykset?"

Sairausgeenien paikannus



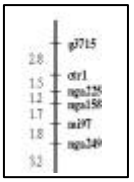
1. Survey pedigree data for the role and type of genetic component



2. Scan the genome for a rough location

Sairausgeenien paikannus

3. Map the gene to a narrow region



4. Analyze this area in more detail

- Tiedon louhinnalla voidaan havaita esim.yhteyksiä geneettisten markkereiden ja sairauden välillä

Assosiaatiosäännöt

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

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

Assosiaatiosäännöt

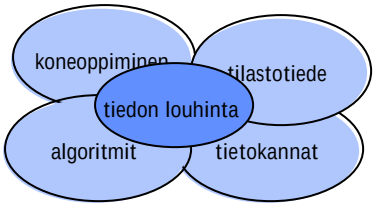
- 2. yleistys: mitkä hahmot esiintyvät aineistossa usein?
  - Syöte
    - *r:* tietokanta
    - *P:* suuri joukko hahmoja tai hahmojen "kieli"
    - *k:* yleisyyden kynnysarvo
  - Tulos
    - kaikki joukon *P* hahmot, joiden yleisyys ylittää kynnysarvon *k* tietokannassa *r*
- Algoritmikehitys:
  assosiaatiosäännöt
  → episodisäännöt (assosiaatiot tapahtumajonoissa)
  → yleinen menetelmärunko
- Teoreettinen kehitys:
  konkreettinen ongelma (ostoskorianalyysi)
  → yleistetty ongelmatyyppi (toistuvat ilmiöt)
  → tehtävätyypin ja ratkaisuvaihtoehtojen analyysi

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?

Tiedon louhinnan lähinaapurit



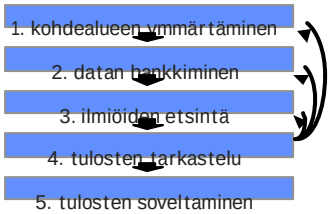
Tiedon louhinta:
- automatisoitu analyysi, algoritmit
- hahmojen ("hypoteesien") löytäminen
- suurten datamassojen käsittely
- tavoitteena ymmärryksen lisääminen

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

Tiedon louhintaprosessi

- Tiedon louhinnassa tutkitaan algoritmien lisäksi myös koko analyysiprosessia



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

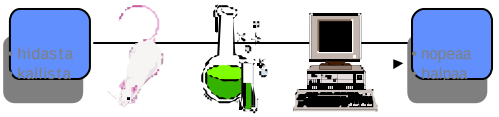
Tiedon louhinta ja TKTL

- Informaatiojärjestelmien linja
  - geenikartoitusmenetelmät
  - ekologiset data-analyysiongelmat (mm. ilmaston rekonstruointi)
  - hahmokielet, algoritmikehitys
  - tekstien ja dokumenttirakenteiden louhinta
  - Hannu Toivonen, Helena Ahonen-Myka, Pirjo Ronkainen, Mika Klemettinen, Marko Salmenkivi, Oskari Heinonen, ... (+ Heikki Mannila)

+ laitoksella toimivat "virtuaaliorganisaatiot"
  - FDK-huippuyksikkö (From Data to Knowledge)
    - tiedon louhinnan ja hahmonsovituksen "kattoprojekti"
    - Esko Ukkonen
  - HIIT/BRU
    - proaktiivinen laskenta
    - data-analyysi
    - Heikki Mannila

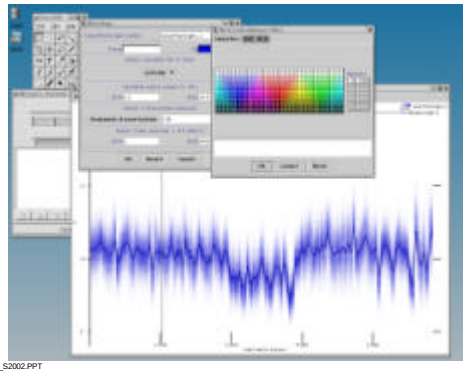
+ muita laitoksen ryhmiä
  - koneoppiminen ja robotiikka (Tapio Elomaa)
  - informaatioteoreettiset ja bayesilaiset menetelmät (Henry Tirri)

Syövälle altistavien yhdisteiden tunnistaminen



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

Ilmaston rekonstruointi



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

Laitoksella kansainvälisesti korkealaatuista tutkimusta

Runsaasti tieteellisiä yhteistyöprojekteja