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 4.3 miljardia sivua)
  - mittaus- ja lokitietokannat
  - geneettiset aineistot (ihmisen dna: 3 miljardia emäsparia)
  - ...

Informaatiojärjestelmät

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

Infon opetus

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

} cum laude

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

} laudatur

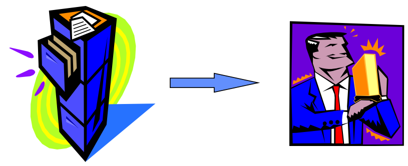
Infon tutkimus

- Tiedon louhinta (Hannu Toivonen, Helena Ahonen-Myka, 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)
  - tietojen mallintaminen, samanaikaisen käytön valvonnan ongelmat, tietokantarakenteiden elvytys, tietokantasovellusten suunnitteluvälineet ja toteutusmenetelmät

Sisällysluettelo

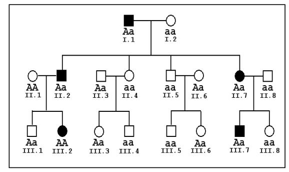
Tiedon louhinta

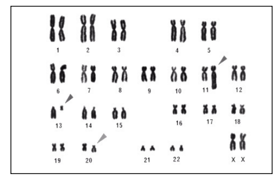
- Uuden ja hyödyllisen tiedon päätteleminen suurista datamassoista



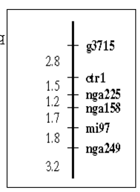
- "Moderni data-analyysi" tai "algoritminen tilastotiede"
- "Mitä data kertoisi, jos siltä osaisi kysyä oikeat kysymykset?"

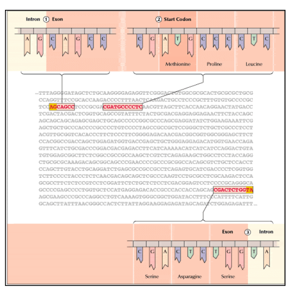
Sairausgeenien paikannus





Sairausgeenien paikannus





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

Assosiaatiosäännöt

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

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 %)

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

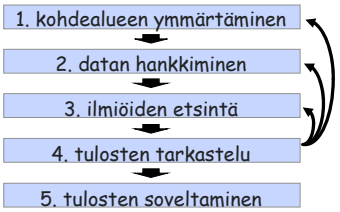
Muita tiedon louhinnan muotoja

- 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?

13 TKTL_S2004.PPT

Tiedon louhintaprosessi

- Tiedon louhinnassa tutkitaan algoritmien lisäksi myös koko analyysiprosessia



14 TKTL_S2004.PPT

Tutkimusmenetelmä

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

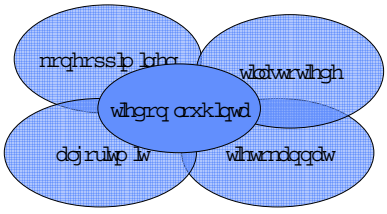
15 TKTL_S2004.PPT

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*
  - Analyysi
    - ongelman teoreettisista ominaisuuksista seuraa, että tietty yksinkertainen algoritmi on tehtävään optimaalinen (tietyillä oletuksilla)

16 TKTL_S2004.PPT

Tiedon louhinnan lähinaapurit



Wlhgrq ocxklqwd=
- dxwrp dwlvr lwx dqdd||vl/#dojrulwp lw
- kdlkp rmlnq +#k|sr whhvlhq%,#5´|wlip lqhq
- vxxuwhq gdwdp dvvrmlnq nfivlwwhd|
- wdyr lwwhhqd |p p fiuu|nvhq dvfifip lqhq

18 TKTL_S2004.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

19 TKTL_S2004.PPT

Tiedon louhinta ja TKTL

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

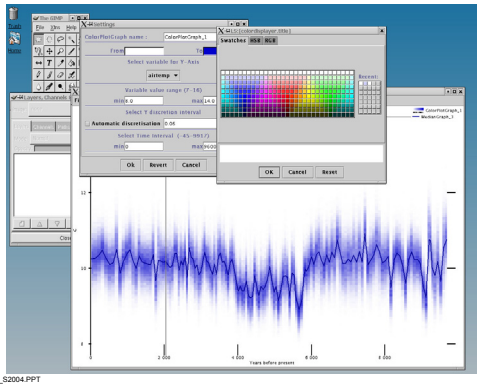
21 TKTL_S2004.PPT

Syövälle altistavien yhdisteiden tunnistaminen

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

22 TKTL_S2004.PPT

Ilmaston rekonstruointi



23 TKTL_S2004.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

24 TKTL_S2004.PPT