Tiedon louhinta

Hannu Toivonen
15.10.2009

http://www.cs.helsinki.fi/hannu.toivonen

A	1	4	8	2	2	1	2	6	2
A	2	4	3	7	3	2	8	4	2
A	4	5	2	4	5	5	2	6	4
A	7	2	3	7	5	4	5	2	2
A	5	2	4	6	2	4	2	6	1
A	3	4	3	7	3	1	3	3	4
A	1	2	1	5	2	5	2	6	2
A	5	3	3	7	3	2	1	4	3
B	2	4	7	1	3	4	1	4	8
B	7	3	7	7	5	7	8	6	6
B	3	4	3	2	5	3	2	3	2
B	2	5	2	4	3	1	3	6	2
B	3	3	1	2	4	2	1	4	2
B	1	6	4	5	5	5	9	1	3
B	4	2	8	4	2	3	5	2	5
B	2	2	4	9	5	4	4	2	4

A	1	4	8	2	2	1	2	6	2
A	2	4	3	7	3	2	8	4	2
A	4	5	2	4	5	5	2	6	4
A	7	2	3	7	5	4	5	2	2
A	5	2	4	6	2	4	2	6	1
A	3	4	3	7	3	1	3	3	4
A	1	2	1	5	2	5	2	6	2
A	5	3	3	7	3	2	1	4	3
B	2	4	7	1	3	4	1	4	8
B	7	3	7	7	5	7	8	6	6
B	3	4	3	2	5	3	2	3	2
B	2	5	2	4	3	1	3	6	2
B	3	3	1	2	4	2	1	4	2
B	1	6	4	5	5	5	9	1	3
B	4	2	8	4	2	3	5	2	5
B	2	2	4	9	5	4	4	2	4

yksilön tyyppi	osa yksilön perimästä				luettua dataa				
sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sairas	4	5	2	4	5	5	2	6	4
sairas	7	2	3	7	5	4	5	2	2
sairas	5	2	4	6	2	4	2	6	1
sairas	3	4	3	7	3	1	3	3	4
sairas	1	2	1	5	2	5	2	6	2
sairas	5	3	3	7	3	2	1	4	3
verrokki	2	4	7	1	3	4	1	4	8
verrokki	7	3	7	7	5	7	8	6	6
verrokki	3	4	3	2	5	3	2	3	2
verrokki	2	5	2	4	3	1	3	6	2
verrokki	3	3	1	2	4	2	1	4	2
verrokki	1	6	4	5	5	5	9	1	3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4

Luennon sisältö

1. Yleiskuvaa tiedon louhinnasta

2. Toistuvien hahmojen etsintä
   - klassinen tiedonlouhintaongelma

Tiedon louhinta

- Uuden ja hyödyllisen tiedon päättelemistä datasta
- “Mitä data kertoisi, jos siltä osaisi kysyä oikeat kysymykset?”

- Tyypillisiä tiedonlouhintatehtäviä
  - ennustaminen (luokittelu, regressio)
  - klusterointi eli ryvästys
  - poikkeusten havaitseminen
  - toistuvien hahmojen etsintä

Ennustaminen (prediction)

sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sairas	4	5	2	4	5	5	2	6	4
sairas	7	2	3	7	5	4	5	2	2
sairas	5	2	4	6	2	4	2	6	1
sairas	3	4	3	7	3	1	3	3	4
sairas	1	2	1	5	2	5	2	6	2
sairas	5	3	3	7	3	2	1	4	3
verrokki	2	4	7	1	3	4	1	4	8
verrokki	7	3	7	7	5	7	8	6	6
verrokki	3	4	3	2	5	3	2	3	2
verrokki	2	5	2	4	3	1	3	6	2
verrokki	3	3	1	2	4	2	1	4	2
verrokki	1	6	4	5	5	5	9	1	3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4

3 7 → sairas

(Ks. luento 4, koneoppiminen)

Ennustaminen

sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sairas	4	5	2	4	5	5	2	6	4
sairas	7	2	3	7	5	4	5	2	2
sairas	5	2	4	6	2	4	2	6	1
sairas	3	4	3	7	3	1	3	3	4
sairas	1	2	1	5	2	5	2	6	2
sairas	5	3	3	7	3	2	1	4	3
verrokki	2	4	7	1	3	4	1	4	8
verrokki	7	3	7	7	5	7	8	6	6
verrokki	3	4	3	2	5	3	2	3	2
verrokki	2	5	2	4	3	1	3	6	2
verrokki	3	3	1	2	4	2	1	4	2
verrokki	1	6	4	5	5	5	9	1	3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4

3 7 tai 2 6 → sairas

Ennustaminen

sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sai									4
sai									2
sai									1
sai									4
sai									2
sai									3
ver									8
ver									6
ver									2
ver									2
ver									2
ver									3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4

Kiinnostavaa tiedon louhinnassa on usein se, mihin ennustus perustuu -- ei välttämättä itse ennustus.

Taudille (tilastollisesti) altistavien kohtien löytäminen perimästä auttaa paikantamaan tautiin vaikuttavia geenejä (geenikartoitus).

3 7 → sairas

Klusterointi (clustering)

1	4	8	2	2	1	2	6	2
2	4	3	7	3	2	8	4	2
4	5	2	4	5	5	2	6	4
7	2	3	7	5	4	5	2	2
5	2	4	6	2	4	2	6	1
3	4	3	7	3	1	3	3	4
1	2	1	5	2	5	2	6	2
5	3	3	7	3	2	1	4	3
2	4	7	1	3	4	1	4	8
7	3	7	7	5	7	8	6	6
3	4	3	2	5	3	2	3	2
2	5	2	4	3	1	3	6	2
3	3	1	2	4	2	1	4	2
1	6	4	5	5	5	9	1	3
4	2	8	4	2	3	5	2	5
2	2	4	9	5	4	4	2	4

Klusterointi

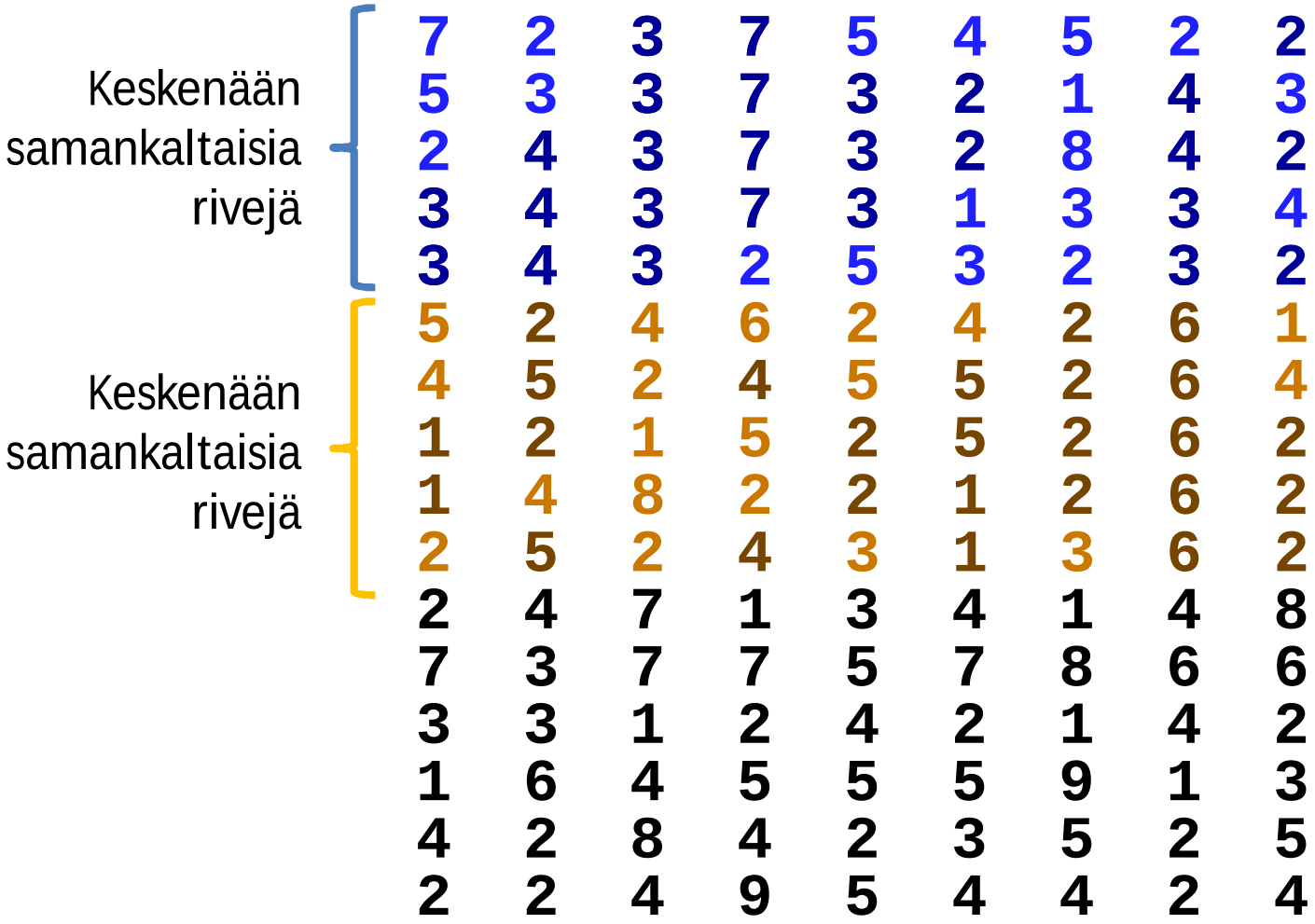
1	4	8	2	2	1	2	6	2
2	4	3	7	3	2	8	4	2
4	5	2	4	5	5	2	6	4
7	2	3	7	5	4	5	2	2
5	2	4	6	2	4	2	6	1
3	4	3	7	3	1	3	3	4
1	2	1	5	2	5	2	6	2
5	3	3	7	3	2	1	4	3
2	4	7	1	3	4	1	4	8
7	3	7	7	5	7	8	6	6
3	4	3	2	5	3	2	3	2
2	5	2	4	3	1	3	6	2
3	3	1	2	4	2	1	4	2
1	6	4	5	5	5	9	1	3
4	2	8	4	2	3	5	2	5
2	2	4	9	5	4	4	2	4

Klusterointi

Keskenään samankaltaisia rivejä

7	2	3	7	5	4	5	2	2
5	3	3	7	3	2	1	4	3
2	4	3	7	3	2	8	4	2
3	4	3	7	3	1	3	3	4
3	4	3	2	5	3	2	3	2
5	2	4	6	2	4	2	6	1
4	5	2	4	5	5	2	6	4
1	2	1	5	2	5	2	6	2
1	4	8	2	2	1	2	6	2
2	5	2	4	3	1	3	6	2
2	4	7	1	3	4	1	4	8
7	3	7	7	5	7	8	6	6
3	3	1	2	4	2	1	4	2
1	6	4	5	5	5	9	1	3
4	2	8	4	2	3	5	2	5
2	2	4	9	5	4	4	2	4

Klusterointi



Toistuvat hahmot (frequent patterns)

sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sairas	4	5	2	4	5	5	2	6	4
sairas	7	2	3	7	5	4	5	2	2
sairas	5	2	4	6	2	4	2	6	1
sairas	3	4	3	7	3	1	3	3	4
sairas	1	2	1	5	2	5	2	6	2
sairas	5	3	3	7	3	2	1	4	3
verrokki	2	4	7	1	3	4	1	4	8
verrokki	7	3	7	7	5	7	8	6	6
verrokki	3	4	3	2	5	3	2	3	2
verrokki	2	5	2	4	3	1	3	6	2
verrokki	3	3	1	2	4	2	1	4	2
verrokki	1	6	4	5	5	5	9	1	3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4

Toistuvat hahmot

sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sairas	4	5	2	4	5	5	2	6	4
sairas	7	2	3	7	5	4	5	2	2
sairas	5	2	4	6	2	4	2	6	1
sairas	3	4	3	7	3	1	3	3	4
sairas	1	2	1	5	2	5	2	6	2
sairas	5	3	3	7	3	2	1	4	3
verrokki	2	4	7	1	3	4	1	4	8
verrokki	7	3	7	7	5	7	8	6	6
verrokki	3	4	3	2	5	3	2	3	2
verrokki	2	5	2	4	3	1	3	6	2
verrokki	3	3	1	2	4	2	1	4	2
verrokki	1	6	4	5	5	5	9	1	3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4
-hahmo 1:	(3)	(4)	3	(7)	(3)	(2)			

Toistuvat hahmot

sairas	1	4	8	2	2	1	2	6	2
sairas	2	4	3	7	3	2	8	4	2
sairas	4	5	2	4	5	5	2	6	4
sairas	7	2	3	7	5	4	5	2	2
sairas	5	2	4	6	2	4	2	6	1
sairas	3	4	3	7	3	1	3	3	4
sairas	1	2	1	5	2	5	2	6	2
sairas	5	3	3	7	3	2	1	4	3
verrokki	2	4	7	1	3	4	1	4	8
verrokki	7	3	7	7	5	7	8	6	6
verrokki	3	4	3	2	5	3	2	3	2
verrokki	2	5	2	4	3	1	3	6	2
verrokki	3	3	1	2	4	2	1	4	2
verrokki	1	6	4	5	5	5	9	1	3
verrokki	4	2	8	4	2	3	5	2	5
verrokki	2	2	4	9	5	4	4	2	4
-hahmo 1:	(3)	(4)	3	(7)	(3)	(2)			
-hahmo 2:						(5)	(2)	6	(2)

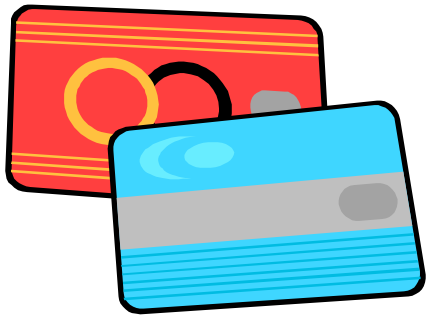
Tavoitteita tiedonlouhintamenetelmille

- automaattisuus
  - laajan etsintäavaruuden läpikäynti
- korkean tason kuvaus
  - ymmärrettävät tulokset
- hyödylliset tulokset
  - uutta/yllättävää/arvokasta
- tulosteen tarkkuus
  - hyvä ja tarvittaessa arvioitavissa
- tehokkuus
  - analyysiajat kohtuullisia ja ennakoitavissa

Esimerkkejä sovelluksista

- Miten menetelmiä voisi soveltaa laitoksen kurssi-ilmoittautumistietoihin?
  - klusterointi: etsi erilaisia opiskelijaprofiileja tai ryhmittele kursseja niiden kävijöiden mukaan
  - luokittelu: ennusta tietylle kurssille osallistumista tai etsi valmistumista ennustavia tekijöitä
  - toistuvat hahmot: etsi tyypillisiä kurssiyhdistelmiä
  - muutosten tai poikkeusten havaitseminen
- HUOM: yksityisyyden suoja henkilötietojen käsittelyssä!

Harjoitus: luottokorttipetosten tunnistaminen



- Motivaatio: identiteettivarkauksista ja muista luottokorttipetoksista miljardiluokan menetykset vuodessa
- Tavoite: tiedon louhinta luottokortin luvattoman käytön automaattiseksi tunnistamiseksi
- Data: korttifirman ostostapahtumatiedot

- Tehtävä: pohdi miten erilaisia tiedonlouhintamenetelmiä voisi soveltaa tähän ongelmaan.

Ratkaisumahdollisuuksia

- Luokittele kortit varastettuihin ja omiin
- Havaitse muutos käyttöprofiilissa
- Etsi poikkeukselliset tapahtumat
- Klusteroi ostajat (tai myyjät) ostostapojen mukaan
- Etsi varastetuilla korteilla toistuvia hahmoja
- ...

Toistuvat hahmot

Toistuvia hahmoja

- Mitä tuotteita kaupasta ostetaan usein yhdessä?
  - maito, leipä
  - maito, perunat
  - maito, leipä, perunat
  - valmispizza, kola, sipsit
  - iltapäivälehti, tupakka
  - vaipat, olut

Toistuvat hahmot

- Tavoitteena on kuvailla mahdollisesti mielenkiintoisia yksinkertaisia ilmiöitä
- Perusmenetelmä tuottaa kaikki hahmot, joiden frekvenssi > kynnysarvo
- Tuoteyhdistelmiä on valtavasti, läpikäynti käsin olisi mahdotonta
- Joukossa voi olla yllättäviäkin joukkoja
- Menetelmä on sovellusriippumaton

Ongelman laskennallinen vaativuus

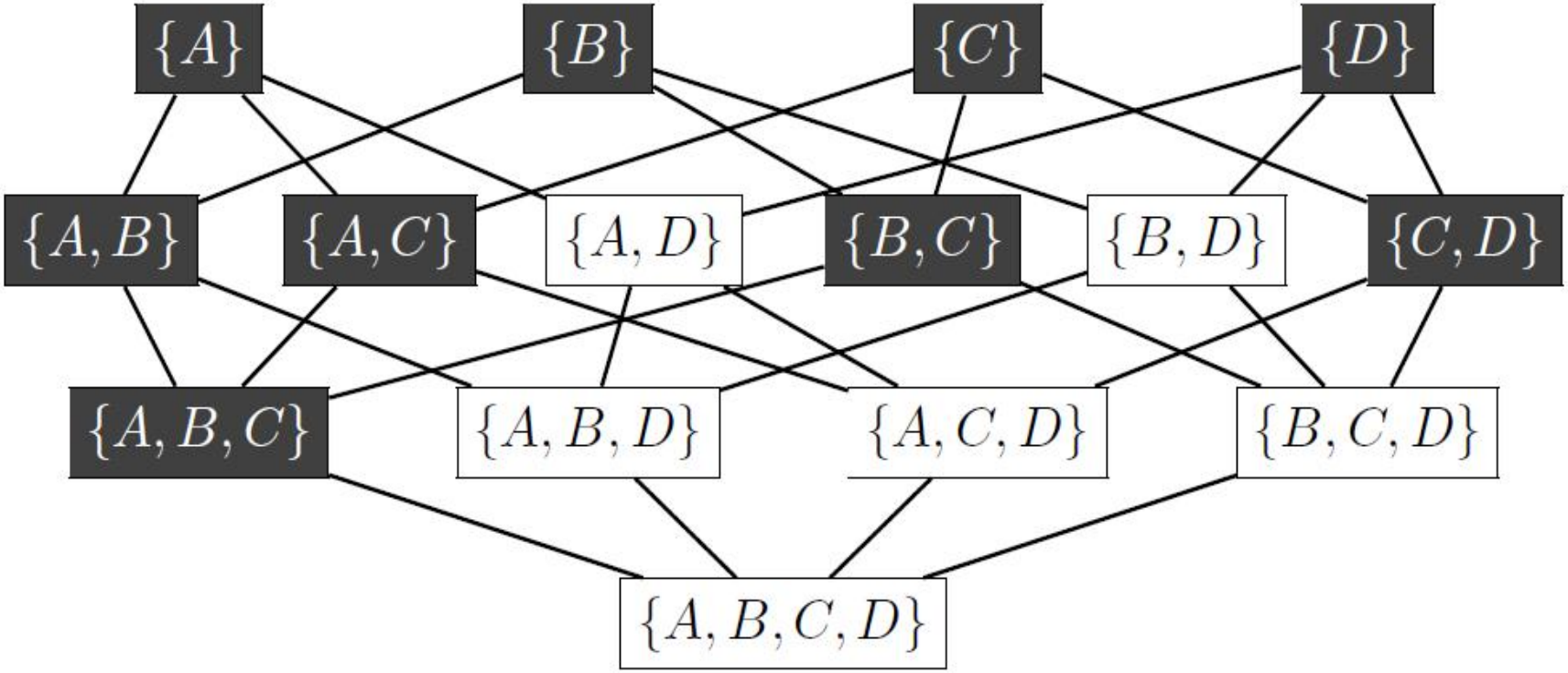
- Mahdollisia hahmoja voi olla valtavasti
  - Jos pienessä kioskissa on myynnissä 100 tuotetta, niin erilaisia tuoteyhdistelmiä on

    $2^{100} \approx 10^{30} = 1000000000000000000000000000000$
- Jos yhdistelmiä tutkitaan miljoona sekunnissa, niin aikaa kaikkien tutkimiseen kuluu $4 \cdot 10^{16}$ vuotta
  - maailmankaikkeus on n. $14 \cdot 10^{9}$ vuotta
- Kuitenkin vain murto-osa yhdistelmistä toistuu edes kaksi kertaa
- Yksi tiedon louhinnan ydinalueita on ollut tehokkaiden menetelmien kehittäminen toistuvien hahmojen etsintään

Toistuvien hahmojen tehokas haku

- Opetetaan tiedon louhinnan kurssilla
  - seuraavan kerran keväällä 2010
- Apriori-algoritmin periaatteet toistuvien joukkojen löytämiseksi:
  - tarkastelee hakuavaruutta joukkojen muodostamana suunnattuna verkkona
  - leveyssuuntainen haku pienistä joukoista alkaen
  - isompi joukko karsitaan hausta, jos jokin sen osajoukko ei ollut toistuva (ks. luento 3, etsintä)
  - algoritmi päättyy kun kaikki jäljellä olevat joukot on karsittu

Toistuvien hahmojen hakuavaruus



Toistuvia hahmoja

- Mitkä geneettiset markkerit esiintyvät usein yhdessä?
- Millaiset hälytysyhdistelmät toistuvat televerkossa?
- Millaiset merkkijonot tai fraasit ovat yleisiä eri teksteissä?
- Millaiset kemialliset rakenteet ovat yleisiä karsinogeenisissä yhdisteissä?
- ...

Toistuvien hahmojen yleinen muotoilu

- Syöte
  - r: tietokanta
  - P: suuri joukko hahmoja tai hahmojen "kieli"
  - f: funktio, joka laskee hahmon $p \in P$ yleisyyden f(p, r) tietokannassa r
  - k: yleisyyden kynnysarvo
- Tulos
  - kaikki joukon P hahmot, joiden yleisyys ylittää kynnysarvon k tietokannassa r, eli joukko $\{p \in P \mid f(p,r) \geq k\}$
- Tehokkaita ratkaisuja tunnetaan, etenkin kun P:n hahmoille on sopiva yleistyshierarkia

Toistuvat hahmot tiedon louhinnan perustyökaluna

- Yksinkertaiset hahmot ovat helposti ymmärrettäviä
- Hyvin harvinaiset hahmot ovat harvoin mielenkiintoisia
- Etsintä on tehokasta
- Tarkasteltavat hahmotyypit voidaan räätälöidä sovelluksen mukaan
- Toistuvia hahmoja voidaan käyttää monimutkaisempien mallien rakennuspalikoina
  - assosiaatiosäännät (association rule):
    “jos vaippoja niin olutta”, säännön vahvuus (ehdollinen tn):
    P(olut | vaipat) = f(olut ja vaipat) / f(vaipat)
  - palikoina myös ennustuksessa, klusteroinnissa, yms.

Lopuksi vielä tiedon louhinnasta

Tiedon louhinta tieteenalana

- Tuottaa ja tunnistaa erilaisia datan ohjelmalliseen analysointiin ja kuvailemiseen liittyviä ongelmatyyppejä tai lähestymistapoja
- Analysoi ja kategorisoi niitä
- Kehittää niihin tehokkaita ratkaisuja

Tiedon louhinnan 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?

Miksi tiedon louhintaa?

- Soveltajat: tarpeita ja mahdollisuuksia on, samoin dataa
  - Kaupalliset ja tieteelliset mahdollisuudet
  - Datan kerääminen ja tallettaminen on helppoa
  - Ensisijaiset tarpeet on tyydytetty (tapahtumankäsittely, talletus)
  - Tietovarastot (data warehouse)
  - Usein louhitaan muita tarkoituksia varten kerättyä tietoa
- Tietojenkäsittelijät: uusia mielenkiintoisia ongelmia
  - Uusia ongelmatyyppejä...
  - ...joitten teoria on vasta muodostumassa
  - Teoria usein suoraan sovellettavissa käytäntöön
  - Poikkitieteellisen yhteistyön mahdollisuus

- Laitoksella Suomen Akatemian Algodan-huippuyksikkö

Harjoitustehtäviä

* Jos 30 % laitoksen opiskelijoista suorittaa ASAn, 15 % tiedon louhinnan ja 10 % molemmat, mikä on assosiaatiosäännön ”jos ASA niin tiedon louhinta” voimakkuus?
** Miten suhtautuisit siihen, että laitos louhisi kurssi-ilmoittautumis-tietoja? Millä edellytyksillä se olisi sinulle ok? (Onko sinulla muuten minkään kaupan kanta-asiakaskorttia…?)
*** Valitse jokin sinulle tuttu aineisto. Miten siihen voisi soveltaa toistuvien hahmojen etsintää, klusterointia ja luokittelua?
**** Kalvolla ”Toistuvien hahmojen hakuavaruus” tummat solmut ovat eräässä aineistossa toistuvia joukkoja. Apriori-algoritmi etsii toistuvat joukot leveyssuuntaisella haulla ja karsii joukon, jos jokin sen osajoukko ei ole toistuva. Mitä valkoisista (ei-toistuvista) joukoista Apriori ei kuitenkaan pysty karsimaan?
***** Jos kioskissa on myynnissä 100 tuotetta, erilaisia mahdollisia toistuvia joukkoja on 2^{100} kpl. Paljonko on olemassa sellaisia hahmoja, joissa voidaan erikseen mainita myös se, että jokin tuote ei esiinny ostoksessa? (Miten tämä vaikuttaa toistuvien hahmojen määrään?)