Tiedonlouhinta rakenteisista dokumenteista (seminaarityö)

Miika Nurminen
(minurmin@jyu.fi)
Jyväskylän yliopisto
Tietotekniikan laitos

Kalvot ja seminaarityö verkossa:
http://users.jyu.fi/~minurmin/gradusem/

Tiedonlouhinta (*Data Mining*)

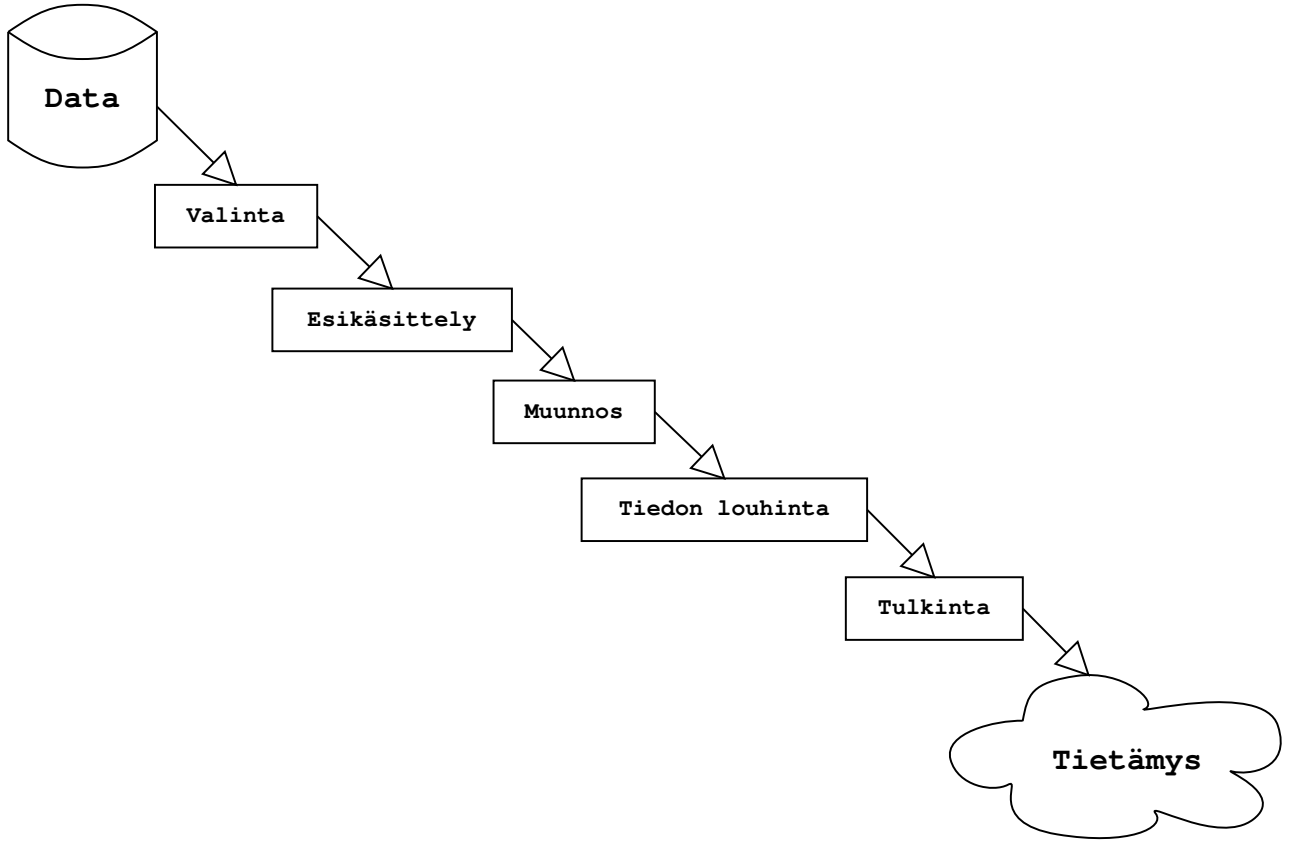
- Hand et al: Suurten, tiettyä tarkoitusta varten kerättyjen tietojoukkojen analyysiä, jonka tarkoituksena on löytää odottamattomia suhteita ja tiivistää dataa uusilla tavoilla, jotka ovat sekä ymmärrettäviä että käyttökelpoisia.

- Monitieteistä toimintaa: lähialueita ovat tilastotiede, tietokannat, koneoppiminen, hahmontunnistus, tekoäly ja visualisointi.

- Kehittyvä tieteenala: teoreettinen perusta on vielä muotoutumassa, mutta tiettyjen menetelmien ja algoritmien katsotaan kuuluvan tiedonlouhinnan osaksi.

- ”Maailma hukkuu tietoon” $\Rightarrow$ Tulevaisuudessa yhä tärkeämpi tutkimuskohde.

Tietämyksen muodostaminen tietokannoista (*Knowledge Discovery in Databases*)



Tiedonlouhinta osana tietämyksen muodostamista

Louhintatieteiden perhe

Tiedonlouhinnan kohteena on perinteisesti ollut numeerisista ja luetelluista arvoista koostuva matriisi ($\approx$ relaatiotietokannan taulu).
$\Rightarrow$ Tarvitaan uusia tutkimusaloja monimuotoisen tiedon louhintaan.

- Tekstitiedon louhinta (*Text mining*, luonnollisen kielen käsittelyä)
- Web-louhinta (*Web mining*, WWW:n sisällön, rakenteen ja käytön analysointi.)
- Relaatiotiedon louhinta (*Multirelational Data Mining*, tutkii erilaisilla suhteilla toisiinsa liittyviä tietojoukkoja)
- Uusia ”tutkimusalueita” odotettavissa kaupallisiin ohjelmistoihin markkinointitarkoituksessa keksittyjen termien myötä...

Louhintamenetelmistä

- Klusterointi (samanlaisten alkioiden ryhmittely)
  - Sisältää mm. hierarkkiset, osittavat (K-Means...), ruudukkoon ja keskinäiseen esiintymiseen perustuvat menetelmät.
  - Samanlaisuus määritellään etäisyysmitan avulla.
- Luokittelu (tietoalkio pyritään sovittamaan johonkin ennalta määrätyistä luokista)
- Regressioanalyysi (tietoalkiolle määritetään numeerinen arvo)
- Assosiaatioiden ja peräkkäisten toimintojen etsintä (yhdessä esiintyviä tietueita)

Tekstitiedon louhinta

- Tiedonlouhinnan ja luonnollisen kielen käsittelyn (*Natural Language Processing*) soveltamista tekstiin.
- Teksti $\approx$ rakenteetonta tai puolirakenteista tekstiä sisältävä dokumenttikokoelma.
- Hyödyntää ja laajentaa tiedonhaun (*Information Retrieval*) ja tiedonkeruun (*Information Extraction*) tekniikoita.
- Sovellusalueita: tiedonhaun tehostus klusteroinnilla, lyhennelmien muodostaminen, metatiedon generointi (digitaaliset kirjastot, semanttinen web)

Tekstitiedon louhinta: Sovellusesimerkki

Survey Of Clustering Data Mining Techniques
(2002) (Make Corrections)
Pavel Berkhin
Accrue Software

CiteSeer Home/Search Context Related

View or download:
accrue.com/product.../cluster_review.ps
Cached: PS.gz PS PDF DjVu Image Update Help

From: accrue.com/produ.. researchpapers (more)
(Enter author homepages)

(Enter summary)

Rate this article: 1 2 3 4 5 (best)
Comment on this article

Abstract: Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters neccessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historial perspective rooted in mathematics, statistics and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters in unsupervised learning and the resulting system represents a data concept.... (Update)

Active bibliography (related documents): More All
3.1: Learning Simple Relations: Theory and Applications - Berkhin, Becher (2002) (Correct)
1.1: How Many Clusters? Which Clustering Method? Answers Via.. - Fraley, Raftery (1998) (Correct)
0.9: Clustering Algorithms for Spatial Databases: A Survey - Kolatch (Correct)

Similar documents based on text: More All
0.1: Local Decomposition Algorithms - Alonso, Mora, Raimondo (1990) (Correct)
0.1: An Efficient Approach to Clustering in Large Multimedia.. - Hinneburg, Keim (1998) (Correct)
0.1: Hierarchical Taxonomies using Divisive Partitioning - Boley (1998) (Correct)

BibTeX entry: (Update)

```
@techreport{ berkhin02survey,
  author = "Pavel Berkhin",
  title = "Survey Of Clustering Data Mining Techniques",
  institution = "Accrue Software",
  address = "San Jose, CA",
  year = "2002",
  url = "citeseer.nj.nec.com/berkhin02survey.html",
  url = "http://citeseer.nj.nec.com/berkhin02survey.html" }
```

Citations (may not include all citations):
1833 Genetic Algorithms in Search (context) - Goldberg - 1989
1735 Maximum likelihood from incomplete data via the EM algorithm (context) - Dempster, Laird et al. - 1977
1719 Pattern Classification and Scene Analysis (context) - Duda, Hart - 1973

CiteSeer-kirjasto (http://citeseer.nj.nec.com/cs).

Rakenteisista dokumenteista

Rakenteisuus (rajoittuen XML-muotoon) helpottaa tiedon louhintaa:

- Dokumentti voi kuvata sisältönsä merkityksiä rakenteellaan (Käyttäjän haun kannalta oleelliset elementit voidaan erotella).
- Pelkän tekstidatan lisäksi voidaan tutkia myös sivujen rakenteellisia yhtäläisyyksiä.
- Linkkejä analysoimalla voidaan arvioida dokumentin luotettavuutta tai linkitettyjen dokumenttien samanlaisuutta.
- Metatiedon avulla dokumenteista voidaan ilmaista lisätietoa yhtenäisellä kielellä (joilllakin sovellusalueilla myös yhtenäisellä käsitteistöllä $\Rightarrow$ ontologiat).

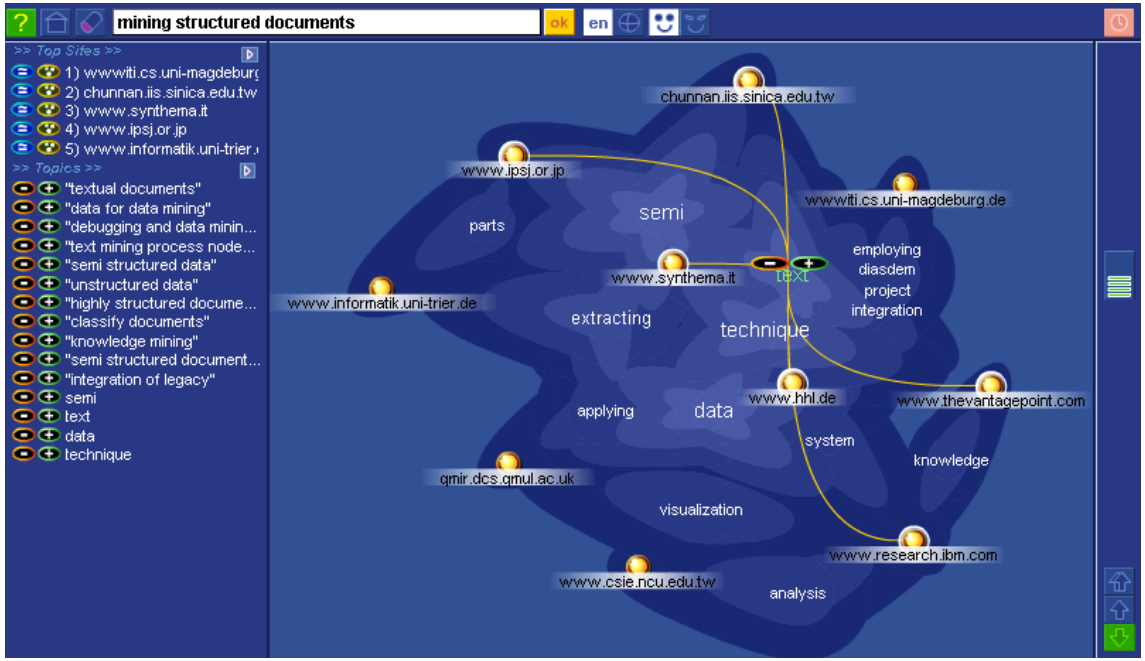
Empiirinen osuus

XML-dokumenttien klusterointisovellus (suunnitteilla).

- Ryhmittelee tietyn DTD:n mukaisen XML-dokumenttikokoelman sisällöltään läheisiin klustereihin ja visualisoi ne esim. itseorganisoituvalla karttalla (*SOM*).
- Dokumenttifragmentteihin perustuva hakutoiminto.
- Konfigurointi eri dokumenttityypeille (esim. määritellään etukäteen prosessointitapoja, joita asetetaan dokumenttityypin soveltuville elementeille).
- Klusterointialgoritmi ja dokumenttien etäisyysmitta ovat vielä työn alla (mahdollisia menetelmiä hyvin runsaasti...).

Käyttöliittymä

Tavoitteena visuaalinen näkymä dokumenttien esiintymistiheyksiin ja avainsanoihin KartOO-hakukoneen tyyliin.



KartOO-hakukone (`http://www.kartoo.com/`).

Elementtien prosessointitapoja

Elementtityyppejä sovellettuna XHTML-tyyliseen dokumenttityyppiin.

- Otsikko (`h1`,`h2`)
- Sisältö (`p`,`span`,`b`,`i`,`img`-elementin `alt`-attribuutti)
- Linkki (`a`, `href`-attribuutit)
- Rakenteen kuvaus (`body`,`div`,`table`,`ul`,`li`)
- Ei-sisältö (`script`, kommentit)
- Metatieto (RDF-kuvaukset)