HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tieteellisen kirjoittamisen kurssi 2. Lähteiden hakeminen ja lukeminen

Antti Leino ‹antti.leino@cs.helsinki.fi›
19.1.2009

Tietojenkäsittelytieteen laitos

Perustuu Timo Aallon, Jaakko Kurhilan ja Seppo Sipun aiempiin luentomateriaaleihin

Tiedon hakemisen ja lukemisen ongelma

- Tutkimusprojekti perustuu aina aiempaan työhön
  - tunnettava aiheen kannalta relevantti lähdekirjallisuus
- Kaksi haastetta
  - aiemman tutkimuksen *löytäminen*
  - merkityksellisen tutkimuksen *tunnistaminen*
- Tietojenkäsittelytieteellisiä julkaisuja on paljon
  - lisää syntyy jatkuvasti
  - vain murto-osa merkityksellistä oman työn kannalta

Tiedonhankinta tutkimuksen eri vaiheissa

- Eri tyyppiset lähteet tutkimusprosessin alussa, keskellä ja lopussa
  - tietääkö mitä hakee vai kartoittaako aluetta
  - tutkimus iteratiivista, samoin tiedonhaku
- Verkon hyötykäyttö
  - tutkimusryhmien sivustot
  - hakupalvelut
  - blogosfääri, verkkoyhteisöt
- Tieteen henkilöityminen
  - haku aiheen / kirjoittajan mukaan

Ei kaikkea yksin

- Muutkin viittaavat aiempaan
  - kätevää erityisesti työn alkuvaiheessa
  - koosteartikkelit erityisen hyödyllisiä
- Lähdekritiikki tarpeen
  - tutkija kommentoi edeltäjiään yleensä vain sen verran kuin oman työn kannalta on tarpeen
  - väärinymmärtämisen mahdollisuus
  - oman työn kannalta tärkeät artikkelit – ne, joihin itse viittaa – on aina luettava myös itse

Tiedonhankintaprosessin vaiheet

- Ongelman tai aiheen määrittely
- Tiedonhaun suunnittelu
  - tiedontarpeen määrittely
  - tiedonlähteiden valinta
  - hakusanojen ja -lauseiden muotoilu
- Tiedonhaun toteutus
  - tiedonhaun tekniikat
- Hakutuloksen arviointi

Tiedonhankintaprosessin vaiheet

- Julkaisujen paikantaminen
  - WWW-lähteiden haku
  - Helka-kokoelmatietokannan käyttö
- Tiedon käyttö ja arviointi
  - lopullisten tulosten arviointi
  - lähdekritiikki
- Ongelman ratkeaminen

Tiedonhaun suunnittelu
Tarpeen määrittely

- Mistä aiheesta tietoa tarvitaan?
  - *paikannimidatan louhinta*
- Mitä aiheesta tiedetään ennestään?
  - *paikkatieto- ja tiedonlouhintakurssit*
- Mihin laajempaan aihepiiriin aihe kuuluu?
  - *tiedon louhinta*
- Mitä osa-alueita siinä on?
  - *yhteisesiintymä- ja assosiaatiosäännöt*
  - *yleiskuvan muodostaminen*
  - *visualisointi*

Tiedonhaun suunnittelu
Tarpeen määrittely

- Mistä näkökulmasta asiaa tarkastellaan?
  - *tietojenkäsittelytiede*
  - *nimistöntutkimus*
- Mikä kuuluu aiheeseen, mikä ei, mikä siinä on keskeistä?
  - *rajaus järvennimiin ja yhteisesiintymäsääntöihin*
- Mistä / keneltä löytyy aiheesta lisää tietoa?
  - *tiedon louhinta - ja paikkatiedon hallinta -aiheiset konferenssit ja lehdet*

Tiedonhaun suunnittelu
Bibliografiahakupalveluita

- ACM Guide to Computing Literature
  - > 1 000 000 tietojenkäsittelytieteen julkaisua
- DBLP (Digital Bibliography & Library Project)
  - > 2^{20} tietojenkäsittelytieteen julkaisua
- CiteSeer / CiteSeer^X
  - > 760 000 / > 1 100 000 tietojenkäsittelytieteen julkaisua
- Bibliografiset tiedot
- Usein myös linkki itse julkaisuun

Tiedonhaun suunnittelu
Kokotekstipalveluita

- ACM Digital Library
  - Kaikki ACM:n julkaisemat artikkelit
- IEEE Xplore
  - IEEE/IEE:n julkaisuja
- Lecture Notes in Computer Science
  - Springerin tietojenkäsittelyn julkaisusarja
- Saatavilla Nelli-portaalin kautta (≈ käytettävissä yliopiston verkosta)

Tiedonhaun suunnittelu
Yleistieteellisiä hakupalveluita

- Google Scholar
  - tieteellistä kirjallisuutta eri aloilta
- Scirus
  - kirjallisuutta ja tutkijoiden / laitosten www-sivuja
- Hakutuloksessa linkki artikkeliin, jos saatavilla verkosta
  - osa näistä kustantajien sivuilta
  - osa tutkijoiden omilta sivuilta – mahdollisesti vasta keskeneräinen käsikirjoitusversio!

Tiedonhaun suunnittelu
Hakusanat ja -lausekkeet

- Käsitteiden analysointi
  - liian yleinen/suppea?
- Käsite hakusanaksi
  - sijamuoto, kirjoitusasu, fraasit, katkaiseminen
- Hakusanojen yhdistäminen
  - rakenteeton haku (vrt. Google)
  - yhdistäminen loogisilla operaattoreilla (and / or / not)
- Kenttiin kohdistaminen
  - tekijä, nimeke, asiasanat, tiivistelmä, kaikki

Tiedonhaun toteutus
Hakutekniikoita

- Pikahaku
  - nopea haku parilla keskeisellä sanalla
- Helmenkasvatus
  - analysoi hyvä löytynyt artikkeli ja paranna hakua siitä löytyvillä käsitteillä
- Viiteanalyysi
  - mihin artikkelissa viitataan ja missä siihen viitataan
- Laajenna hakua synonyymeillä sekä ylä- ja alakäsitteillä
- Tarkenna hakua
- Pidä kirjaa onnistuneista hauista

Hakutulosten arviointi

- Yleensä 10–20 ensimmäisen tuloksen arviointi riittää
- lukemisjärjestys: otsikko – avainsanat – tiivistelmä – johdanto – (kaaviot) – johtopäätökset
  - vasta lopuksi koko artikkeli, jos tarpeen
- Lähdekritiikki!
  - tieteellisessä julkaisussa ilmestyminen ei (aina) laadun tae
  - WWW-tuloksissa kriittisyys erityisen tärkeätä

Hakutulosten arviointi
Verkkolähteiden arviointikriteereitä

- Tekijyys
  - tekijän asema ja asiantuntemus
  - lähteiden käyttö
  - tekstin ulkoasu
  - aineiston sijainti
- Puolueettomuus
- Ajantasaisuus
- Kattavuus
- Tekninen toimivuus

Julkaisujen paikantaminen

- Sähköisesti
  - jo hakutuloksessa yleensä linkki artikkeliin
- Paperilla
  - kirjastot
  - kirjakaupat, usein verkko-
- Maksutta saatavilla
  - verkosta
  - paperilla kirjastosta
  - merkittävä osa näistä maksullista aineistoa, josta joku (kuten yliopisto) jo maksanut
- Maksullista
  - maksullisista verkkopalveluista
  - ostettuna kirjakaupasta

Julkaisujen paikantaminen
Painetut aineistot

- Mieluiten kirjastosta
  - edullista
  - nopeaa
- Kumpulan tiedekirjasto
- Helka-tietokanta
  - Helsingin yliopiston kirjastot
  - muutama muu pääkaupunkiseudulla toimiva tieteellinen kirjasto

Julkaisujen paikantaminen
Luokittelu

- Julkaisut *hyllyissä* UDC-luokituksen mukaan
  - http://www.helsinki.fi/kumpula/tiedekirjasto/kokoelmat/hyllyluokitus.htm
  - http://www.udcc.org/outline/outline.htm
- Tkt:n aineistot *kuvailtu* ACM:n CCS-luokituksen mukaisesti
  - http://www.helsinki.fi/kumpula/tiedekirjasto/kokoelmat/_acm.htm
  - http://www.acm.org/class/1998/
  - kuvaus CCS → UDC yleensä suoraviivainen
    - ACM A. → UDC 004.01

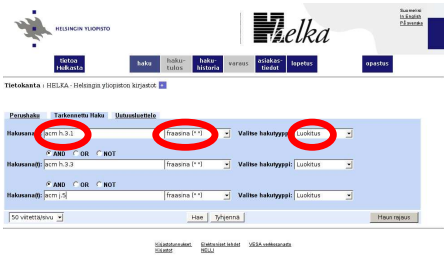
Julkaisujen paikantaminen
UDC-luokitus

- 0 Generalities
  - 00 Prolegomena. Fundamentals of knowledge and culture. Computer science.
    - 004 Computer science and technology. Computing
- 1 Philosophy. Psychology
- 2 Religion. Theology
- 3 Social sciences
- 4 Vacant
- 5 Natural sciences
  - 51 Mathematics
    - 519.4 Information theory. Automata theory
    - 519.6 Computational methods
  - 52 Astronomy. Astrophysics. Space research. Geodesy
    - 528.9 GIS
- 6 Technology
- 7 The Arts
- 8 Language. Linguistics. Literature
- 9 Geography. Biography. History

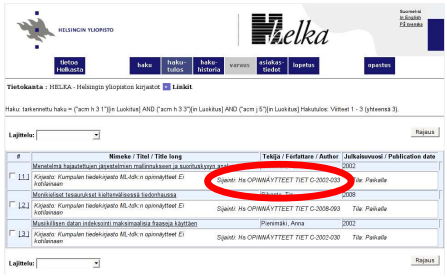
Julkaisujen paikantaminen
ACM:n CCS-luokitus

- A. General Literature
- B. Hardware
- C. Computer Systems Organization
- D. Software
- E. Data
- F. Theory of Computation
  - F.0 General
  - F.1 Computation by Abstract Devices
  - F.2 Analysis of Algorithms and Problem Complexity (B.6, B.7, F.1.3)
  - F.3 Logics and Meanings of Programs
    - F.3.0 General
    - F.3.1 Specifying and Verifying and Reasoning about Programs (D.2.1, D.2.4, D.3.1, E.1)
    - F.3.2 Semantics of Programming Languages (D.3.1)
    - F.3.3 Studies of Program Constructs (D.3.2, D.3.3)
    - F.3.m Miscellaneous
  - F.4 Mathematical Logic and Formal Languages
  - F.m Miscellaneous
- G. Mathematics of Computing
- H. Information Systems
- I. Computing Methodologies
- J. Computer Applications
- K. Computing Milieux

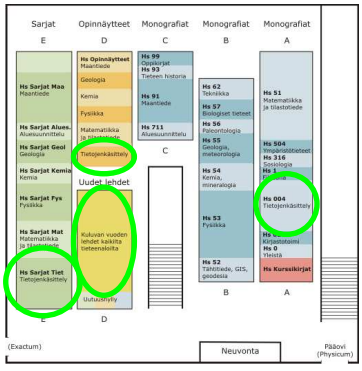
Julkaisujen paikantaminen
Haku Helka-tietokannasta ACM-luokituksella



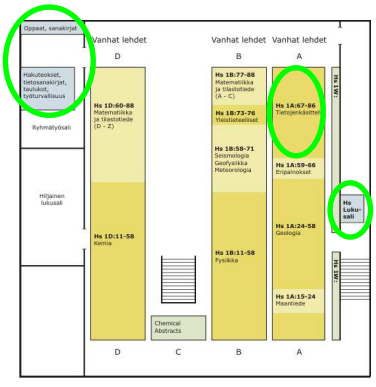
Julkaisujen paikantaminen
Haku Helka-tietokannasta ACM-luokituksella



Julkaisujen paikantaminen
Kumpulan tiedekirjasto, 2. kerros



Julkaisujen paikantaminen
Kumpulan tiedekirjasto, 1. kerros



Tiedon käyttö ja arviointi
Erilaisia lähteitä

- Formaalit tiedonlähteet
  - julkista, vahvistettua ja yleensä kirjallista tietoa
  - oman tekstin lähdeluetteloon yleensä vain näitä
  - jako ensisijaisiin ja toissijaisiin
- Informaalit tiedonlähteet
  - yksityisluontoista, epävirallista tietoa
  - usein hyvää taustamateriaalia

Tiedon käyttö ja arviointi
Ensisijaiset lähteet

- Uutta tietoa ensi kertaa tai täydellisesti ilmaistuna
- Artikkelit tieteellisissä lehdissä ja konferenssijulkaisuissa
- Standardit, patentit, tekniset raportit
- Lait, viranomaismääräykset, asiakirjat
- Tieteellisten julkaisujen oikeellisuus ja arvo pyritty takaamaan esim. vertaisarvioinnilla

Tiedon käyttö ja arviointi
Esimerkkejä ensisijaisista lähteistä

Lei10 Freiherr von Leibniz, G. W., Brevis designatio meditationum de Originibus Gentium, ductis potissimum ex indicio linguarum. *Miscellanea Berolinensia ad incrementum scientiarum, ex scriptis Societati Regiae Scientiarum exhibitis edita* I, sivut 1–16. Johan. Christ. Papenius, Berliini 1710. http://www.bbaw.de/bibliothek/digital/ [7.9.2008].

MTV94 Mannila, H., Toivonen, H. ja Verkamo, A. I., Efficient algorithms for discovering association rules. Teoksessa *Knowledge Discovery in Databases, Papers from the 1994 AAAI Workshop (KDD 94)*, Fayyad, U. M. ja Uthurusamy, R., toimittajat, sivut 181–192. AAAI Press, Menlo Park, CA 1994.

Por07 Portele, C., toimittaja, OpenGIS Geography Markup Language (GML) Encoding Standard. Open Geospatial Consortium Inc., 2007.

Säh04 Sähköisen viestinnän tietosuojalaki. Eduskunta 16.6.2004/516.

Tiedon käyttö ja arviointi
Toissijaiset lähteet

- Tiivistävät, jäsentävät ja auttavat löytämään ensisijaisissa lähteissä esitettyä tietoa
- Oppikirjat, käsikirjat, kokoomateokset, yleiskatsaukset, sanastot jne.
- Osa kirjoista / yleiskatsauksista ensisijaislähteiden veroisia
  - viittausten määrä voi olla keino selvittää teosten käyttöarvoa
- Lähdeluettelot hyvänää apuna ensisijaisten lähteiden paikallistamisessa
- Hyvän koosteartikkelin (»review», »survey») löytäminen voi säästää *paljon* aikaa

Tiedon käyttö ja arviointi
Esimerkkejä toissijaisista lähteistä

ATK08 ATK-sanakirja. 14. painos. Talentum, 2008.

HMS01 Hand, D. J., Mannila, H. ja Smyth, P., Principles of Data Mining. MIT Press, 2001.

MSP08 Miltchev, S., Smith, J. M., Prevelakis, V., Keromytis, A. ja Ioannidis, S., Decentralized Access Control in Distributed File Systems. *ACM Computing Surveys*, 40,3 (2008), artikkeli 10.

Wik08 Wikipedia, http://en.wikipedia.org/wiki/John_William_Mauchly [7.9.2008].

Tiedon käyttö ja arviointi
Formaalien lähteiden laatu

- Pelkkä julkaisun olemassaolo ei ole laadun tae
- Hyvämaineisessa lehdessä / konferenssissa julkaistut artikkelit vertaisarvioitu
- Muualta löytyvä artikkeli ei välttämättä
  - esiversio artikkelista
  - julkaistavaksi tarjottu artikkeli
  - seminaariesitelmä
  - plagiaatti
- Kyyninen nyrkkisääntö: muutamaa vuotta vanhempi artikkeli, jota ei ole julkaistu vakavasti otettavalla foorumilla $\Rightarrow$ jotain vikaa

Tiedon käyttö ja arviointi
Informaalit tiedonlähteet

- Yksityisluontoista / »epävirallista» tietoa
- Oma tietämys
  - »Jostain luennolta minä tämän kai opin, mutten muista mistä ja milloin.»
  - »Jokainen, joka on vähänkään ohjelmoinut Pythonilla, tietää tämän aivan varmasti.»
- »Nahkakantiset manuaalit»: asiantuntijat, kollegat, työn ohjaaja, . . .
- Virallisesti vielä julkaisemattomat tutkimustulokset
- Verkkokeskustelut, blogit ja muut yhteisölliset lähteet

Tiedon käyttö ja arviointi
Informaali tieto

- Saattaa olla vahvistamatonta ja epätäydellistä tai jopa väärää: ei kannata luottaa liikaa
- Hyvää vihjeiden ja hakuideoiden saamisessa sekä tiedon jäsentelyssä
- Viitattava harkiten – vain, jos tietoa ei ole saatavana formaalina

Tiedon käyttö ja arviointi
Muistilista artikkelin lukemiseen

- Mikä on päätulos?
- Kuinka tarkkoja väitteet ovat?
- Miten tuloksia voidaan käyttää?
- Miten tuloksia perustellaan?
- Miten perustelut on koottu?
- Miten mittaukset on suoritettu?
- Kuinka huolellisesti algoritmit ja kokeet kuvataan?
- Onko artikkeli luottamusta herättävä?
- Onko vertailua oikeaan taustakirjallisuuteen?
- Miten tulokset voitaisiin tuottaa uudestaan?

Ongelman ratkaiseminen

- Kun lähteet on löydetty, on aika ajatella itse
- Tutkielmassa omaa ajatustyötä
  - ei pelkkää lähteistä löytyvien tietojen toistoa
  - jonkinlaista omaa jäsentämistä / arviointia
- Iteratiivinen prosessi
  - alkuperäiseen kysymykseen vastaaminen herättää uusia kysymyksiä
  - tarve etsiä uusia lähteitä

Työn iloa