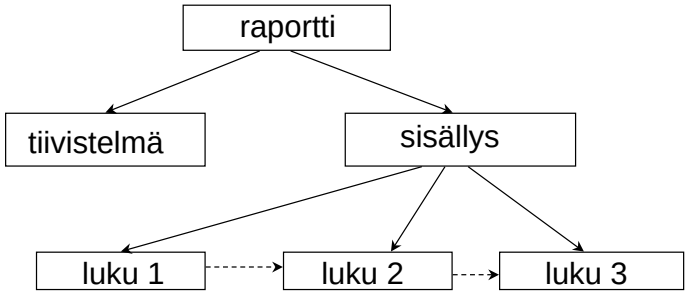
6. Hyperteksti ja tiedonhaku

6.1 Yleistä

Hypertekstimäinen informaation jäsentely:
- solmut
- linkit (yhteydet)

Yhteys tiedonhakuun:
solmu = dokumentti tai sen osa



Hypertekstissä solmu on vahva, riittävä peruskäsite.

Tiedonhaussa pelkkä solmunäkemys on riittämätön.

- haun kohde voi olla solmu tai solmujoukko!

Hypertekstin perusidea:
- ei-lineaarinen lukeminen, selailu
  (browsing, navigating)
- vuorovaikutteisesti:
  käyttäjä löytää (tunnistaa) tiedot
- apuvälineitä: karttoja. luetteloita, historialistoja,
  kirjanmerkkejä, maamerkkejä

Mihin selailu sopii?
- pieni ‘tietokanta’
- materiaali jollain tavalla tuttua
  sisältö, rakenne, muodot, …
- hakujen yhteydessä myös oppii (rakenteesta, sisällöstä)

Tehokkaampiakin hakuvälineitä
tarvitaan!
- WWW (koko!)
-automatisoinnin pyrkimys yleensä,
  selailun rajoitukset

Linkkien tärkeys:
- mahdollistavat (perus)liikkumisen
- ilmaisevat osien välisiä yhteyksiä (erilaisia!)

Linkkien ominaisuuksia (tyyppejä):
- erilaisia ankkureita: teksti, kuva (esitysteknisiä eroja)
-teknisiä, periaatteellisia eroja:
  ulkoinen, (sivun)sisäinen;
  rakenteellinen, viittaava

- esimerkkejä rakenteellisista:
  kokonaisuus → osa
  osa →kokonaisuus
  alaviite

- semanttisia:
  määritelmä
  esimerkki
  erikoistapaus
  kommentti
  vastaus
  (periaatteessa mikä tahansa
  ‘speech act’ tms.)

Tyyppien erottelu?
- käyttäjälle, esim. visuaalisesti
- tiedonhaussa, sopivalla määrittelyllä (metatiedolla)

- Esim. HTML – linkkityyppejä:

start sivujoukon alkudokumenttiin
next sivujoukon seuraavaan
prev sivujoukon edelliseen

contents sisällysluetteloon
index indeksiin (?)
chapter lukuun
section kappaleeseen
subsection alikappaleeseen
appendix liitteeseen

alternate sivun toiseen versioon, esim. toisella kielellä

Esim.

```
<HEAD>
        ...other head information...
        <TITLE>Chapter 5</TITLE>
        <LINK rel="Index" href="../index.html">
        <LINK rel="prev" href="chapter4.html">
        <LINK rel="next" href="chapter6.html">
</HEAD>
```

(käytön laajuus nykyisin ??)

Hypertekstiin kohdistuvan tiedonhaun
erityispiirteitä / ongelmia:

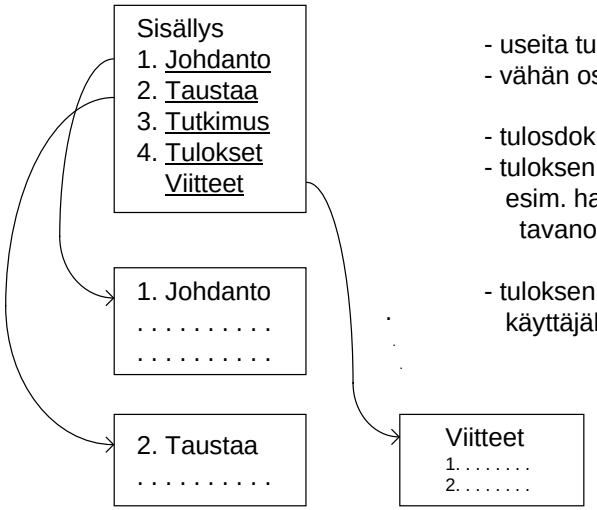
1) haun kohde: solmu vai solmujoukko?

- käyttäjälle kumpi tahansa voi olla 'oikea';
vaikea kuvata etukäteen
- riippuu hypertekstin sisällöstä (jäsentelystä),
linkityksestä sekä tiedontarpeen luonteesta

- samanlaisuuden määritys:
· solmujoukossa enemmän osumia, skaalaus?
· samanlaisuuden laskenta vain solmusta /
ympäristöä huomioiden / solmujoukosta?
- läheiset solmut: mitkä, kuinka monta?

2) tuloksen esittäminen
- pelkkä tulossolmu (sivu) / ympäristöä

Dokumentti hypertekstinä, versio 1:

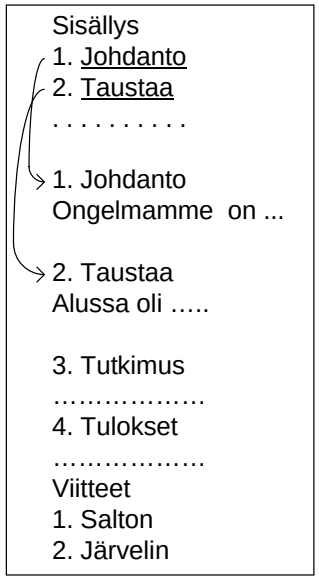


- useita tulosdokumentteja
- vähän osumia / solmu (?)

- tulosdokumenttien järjestys?
- tuloksen esittäminen?
esim. hakukoneiden
tavanomaiset linkkilistat

- tuloksen kokoaminen jää
käyttäjälle

Dokumentti hypertekstinä, versio 2:



- yksi tulosdokumentti
- useita osumia,
  niiden löytäminen selaamalla‘
  erottuminen visuaalisesti?
- haku dokumentin sisällä
  jää käyttäjälle:
  FIND xyz …,
  FIND AGAIN ...

3) tiedonhaun (algoritmisen) ja selaamisen yhdistäminen

Käytännön tilanteita:

1° suorita kysely, selaa joidenkin tulossivujen ympäristöä
2° navigoi jollekin sivulle (ei-algoritmisesti),
hae algoritmisesti
- samanlaisia sivuja
- ympäristöstä

(lisäksi sivunsisäisen haun ongelma, jos sivu on laaja)

Hypertekstin suunnittelun ongelma
- sekä kokonaisuuden (solmujoukon) että
  solmun tasolla

- luettavaksi (lähinnä lineaarisesti, tekstinä)?
- selattavaksi?
- etsittäväksi?

Esim. aikataulusivut lähiliikenteessä

Jäsentely 1: Helsinki, Espoo, Vantaa
Jäsentely 2: bussi, metro, juna
Jäsentely 3: alueittain (mistä mihin)

-tiedontarve esim. Leppävaara → Myyrmäki

- rakennelinkit
- viittauslinkit
- multimediapiirteet

6.2 Hypertekstin tiedonhakumalleja

1° Solmun ympäristön huomiointi haussa

Luonnollinen oletus on, että linkillä yhdistetyillä sivuilla on muutakin kuin teknistä yhteyttä:
- ne ovat ehkä tiedonhaun mielessä lähekkäisiä, tai
- ne muodostavat yhdessä hakuun paremmin sopivan kokonaisuuden

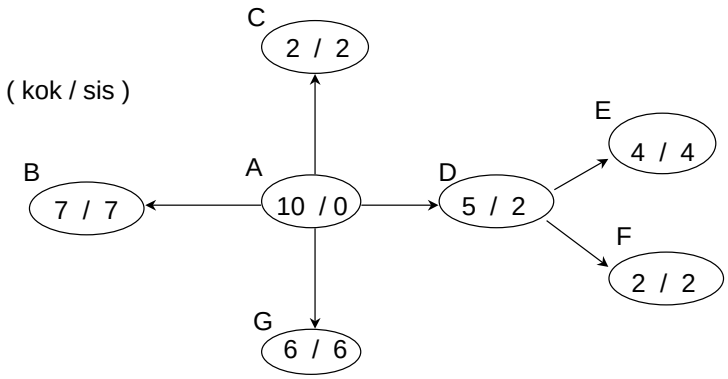
Ts. pelkästään termien esiintymät solmussa eivät ratkaise solmun samanlaisuutta kyselyn nähden; otetaan huomioon myös läheisiä solmuja.

solmun **sisäinen paino** riippuu termiesiintymistä
- lukumäärä, tai
- $w_{ij} = tf_{ij} \times \log(N/dfj)$

solmun **ulkoinen paino** = jälkeläissolmujen (L) vaikutus, esim. niiden sisäisten painojen summa

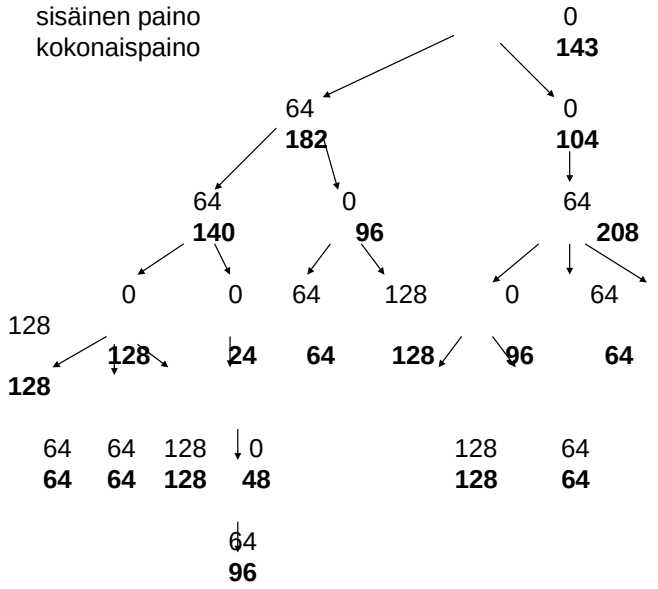
solmun kokonaispaino = $W_{sis} + \alpha \sum_{i \varepsilon L} W_{i,kok}$

Esim. $\alpha = 0.5$



Järjestys: B, G, D, E, C, F; A (sis.)
A, B, G, E, C, D, F (kok.)

Havainnollisempi kuva syntyy laajemmasta hierarkiasta:



Erilaisia laskutapoja:

- jälkeläisten painojen keskiarvo
- linkkityypeille erilaiset kertoimet

- sykliset rakenteet: kuinka lasketaan?
- edeltäjäsolmujen vaikutus?
- etäisyys, kuuluminen kokoelmaan? (jos määritelty)

Kokoelman (solmujoukon) käsite?

- hierarkia
- lähekkäiset (naapurit)
- sisällöllisesti samanlaiset
  (näiden tekijöiden yhdistelmä?)

- harvoin eksplisiittinen
- sivuston suunnittelijalla yleensä selvillä
  (ainakin osittain; WWW??)

2° Linkkitopologiaan perustuva solmujen valinta

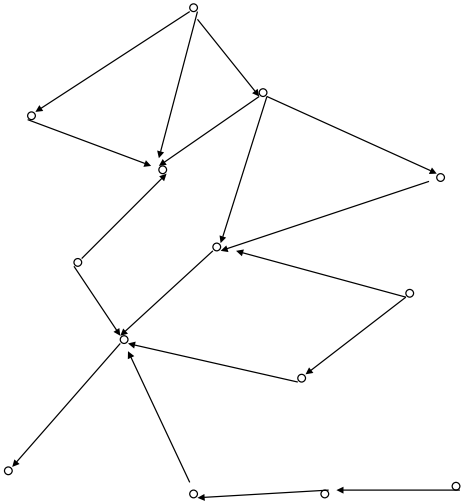
Kleinberg, J., Authoritative sources in a hyperlinked environment. Cornell University, 1998.

- tavoitteena löytää hyviä tulossivuja
  väljästi määriteltyyn kyselyyn (“broad search topic”)
  Esim. “Java-ohjelmointikieli”, “hakukoneet”, “Harvard”

  - tulosjoukko on aivan liian suuri, relevantteja paljon, mutta mitkä parhaita?

  - keskeistä tietoa sisältävät sivut eivät ole välttämättä niitä, joissa termifrekvenssi on suurin (esim. Harvard)

  - esiintymiä voi olla hyvin vähän tai jopa ei yhtään (esim. tietyn hakukoneen “kotisivut”)



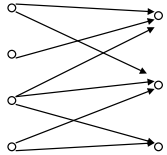
- määritellään **arvosivu** (authority page):

sivu, johon viitataan paljon
(tarkemmin: viittaukset tulevat sivuilta,
joilta viitataan monille arvosivuille)

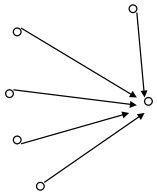
- paljon (arvosivuille kohdistuvia) linkkejä sisältävää

sivua sanotaan **napasivuksi** (hub page)

Arvosivut ja napasivut voidaan määrittää
pelkästään linkkiyhteyksiä tutkimalla.



napoja



arvosivuja

(ei arvosivu)

Oletuksia:

- Linkkiyhteys osoittaa piilevää sivun tekijän
harkintaan perustuvaa yhteyttä eli viitatun
sivun arvoa. (Kaikki linkit eivät ole tällaisia!)

- Tarkastellaan laajaa (globaalia) sivujoukkoa,
ei esimerkiksi paikallista sivustoa.

- Indeksiä (tai sivujen sisältöä, termejä) käytetään
vain aluksi, tavallaan epäsuorasti.

- Hypertekstiä tarkastellaan suunnattuna verkkona
G = (V, E);
särmä (p,q) $\in$ E on linkki sivulta p sivulle q

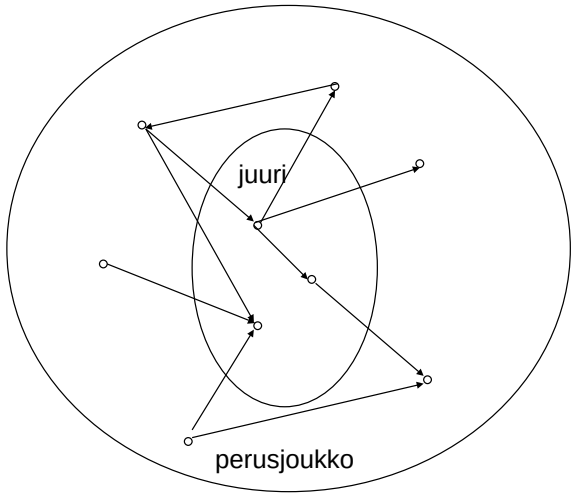
out-degree: solmusta lähtevien linkkien lkm
in-degree: tulevien linkkien lkm
aliverkko: solmut ja niiden väliset linkit

Arvo- ja napasivujen määritys:

1) suoritetaan kysely tavanomaisella hakukoneella ja valitaan sen tuloksesta t ensimmäistä sivua juurijoukoksi

2) laajennetaan juurijoukko perusjoukoksi liittämällä siihen kaikki ei-paikalliset solmut, joihin viitataan juurijoukon solmusta, ja osa juurijoukon solmuihin viittaavista ei-paikallisista solmuista

Paikalliset solmut (sama domain-nimi) tulkitaan
teknisesti yhteenliitetyiksi; linkki ei ilmaise toisen
solmun arvosta mitään.
(Muitakin heuristiikkoja voidaan käyttää: samalta
domain-alueelta
hyväksytään vain rajattu määrä linkkejä.)



3) arvo- ja napasivut määritetään iteroimalla operaatioita

I: sivun p arvopaino $x^{<p>} = \sum_{q:(q,p)\in E} y^{<q>}$

O: sivun p napapaino $y^{<p>} = \sum_{q:(p,q)\in E} x^{<q>}$

(aloitus painoilla 1, skaalauksia, tuloksena tasapainotila, jossa suuren x-painon saaneet sivut valitaan arvosivuiksi ja suuren y-painon saaneet valitaan napasivuiksi)

Huom.
-Solmujen termisisältöä käytetään vain alkukyselyssä.

- Lopulliseen tulokseen voi linkkiyhteyksien kautta tulla solmuja, jotka eivät kuulu alkukyselyn tulokseen ja joissa ei ehkä ole kyselytermin esiintymiä lainkaan.
- Mukaan voi tulla sivuja, joilla on esim. kuvia, mutta hyvin vähän tekstisisältöä.
- Voi löytyä erillisiä aliverkkoja, joissa on omat arvo- ja napasivuparinsa (mutta vähän yhteyksiä aliverkosta toiseen). Tällöin menetelmä lähestyy ryvästämistä ...
- Kokeillen on todettu, että menetelmän (Clever-hakukone) tulosjoukko on jopa parempi kuin Yahoon hakemistojen kautta löydetty.
- Menetelmälle on esitetty useita muitakin sovelluskohteita kuin tiedonhaku.