Hyperteksti tieteellisessä julkaisutoiminnassa

Laura Vuorinen
Seminaariesitelmä
Hypermediajärjestelmät
26. huhtikuuta 2002

Helsingin Yliopisto
Tietojenkäsittelytieteen laitos

Hyperteksti tieteellisessä julkaisutoiminnassa

Laura Vuorinen
Seminaariesitelmä
Hypermediajärjestelmät
26. huhtikuuta 2002, 12 sivua

Helsingin Yliopisto
Tietojenkäsittelytieteen laitos

Esitelmässä pohditaan tieteellisen julkaisutoiminnan ongelmia. Tiedon ja julkaistujen artikkeleiden määrä on kasvanut räjähdysmäisesti. Yksittäisen ihmisen on vaikea pysyä julkaisuvauhdissa mukana, edes oman erikoisalan uusien artikkeleiden kattava seuraaminen ei onnistu. Artikkelien lukijoilla on erilaisia tarpeita tiedonhankinnassaan mm. erilaisten pohjatietojen takia. Esitelmässä esitellään kolme erilaista ratkaisuehdotusta näihin ongelmiin. Ratkaisuehdotukset ovat lähestymistavoiltaan teoreettisia, eikä niistä yhdestäkään ole tehty konkreettista käytössä olevaa sovellusta.

Amsterdamin yliopiston Communication in Physics Project tutkii artikkeleiden pilkkomista pieniin, itsenäisiin osiin, moduuleihin. Moduulit linkitetään toisiinsa sekä rakenteen että sisällön perusteella.

James Blustein on tutkinut menetelmiä olemassaolevien tieteellisten artikkeleiden muuttamiseksi hypertekstimuotoon. Artikkeli säilytetään yhtenäisenä, mutta artikkeliin lisätään sekä rakennetta että sisältöä kuvaavia linkkejä. Linkitys tehdään automaattisesti.

Robert Cameron on esittänyt ajatuksen bibliografisia tietoja ja lähdetietoja sisältävästä tietokannasta, joka sisältäisi tiedot kaikista maailman tieteellisistä artikkeleista. Tietokannassa oleva artikkeli olisi linkitetty sekä kaikkiin niihin artikkeleihin joihin se viittaa että kaikkiin niihin artikkeleihin jotka siihen itseensä viittaavat.

Aiheluokat (Computing Reviews 1998): H.5.4 Hypertext/Hypermedia
I.7.2 Document Preparation
I.7.4 Electronic Publishing

Avainsanat: hyperteksti, tieteellinen kirjoittaminen, elektroninen julkaisutoiminta, tieteellinen julkaisutoiminta

Sisällys

1 Tieteellinen kommunikointi 1

1.1 Historiasta nykypäivään 1

1.2 Erilaisia tarpeita 2

1.3 Tiedonhaku artikkeleista 2

2 Artikkeleiden kirjoittaminen modulaariseen muotoon 3

2.1 Moduulien perusajatus 3

2.2 Moduulijaon tekeminen sisällön mukaan 4

2.3 Linkkityypit 4

2.4 Modulaarisuuden tuomat edut 6

3 Blusteinin menetelmä artikkelien muuttamiseksi hypertekstimuotoon 6

3.1 Linkkityypit 7

3.2 Sisällön samankaltaisuutta osoittavien linkkien automaattinen luominen 7

3.3 Linkkien merkitseminen tekstiin 8

4 Cameronin maailmanlaajuinen lähdetiedot sisältävä viitetietokanta 9

4.1 Tietokannan perusidea 9

4.2 Tiedonhaku tietokannasta 10

4.3 Tietokannan tuomat edut 10

4.4 Tietokannan toteutus 11

Lähteet 12

1 Tieteellinen kommunikointi

Ensimmäisen luvun lähteinä on käytetty kaikkia lähdeluettelossa mainittuja artikkeleita. Historiallinen katsaus on peräisin Harmszen väitöskirjasta [Har00].

Tieteen kehittymisen edellytys on se, että tieteen harjoittajat kommunikoivat keskenään. Yksi ihminen ei ehdi tutkia kaikkea, kuorma on järkevä jakaa monen tutkijan kesken. Toinen kommunikoinnin merkitys on ajatusten kehittely ja kritisointi, ideoiden pohdinta yhdessä. Kommunikointi voi olla epämuodollista keskustelua tai muodollista, laajan piirin tavoittavaa tiedonvälitystä.

Esitelmässä pohditaan kommunikoinnin ongelmia, ja esitellään kolme erilaista ratkaisuehdotusta näihin ongelmiin. Ratkaisuehdotukset ovat lähestymistavoiltaan teoreettisia, eikä niistä yhdestäkään ole tehty konkreettista käytössä olevaa sovellusta.

1.1 Historiasta nykypäivään

Kommunikointi on aina ollut oleellinen osa tieteen tekoa. Varhaisista ajoista lähtien tieteenharjoittajat olivat kirjeenvaihdossa keskenään. Englannissa perustettiin 1662 tieteellinen yhteisö Royal Society. Seuran sihteeri valikoi oleelliset ja tärkeät kirjeet ja luki ne ääneen seuran kokouksissa. Kirjeenvaihdon käydessä laajemmaksi ja vaikeasti hallittavaksi sihteeri päätti julkaista kirjeet painettuna lehtenä. 1700-luvun loppupuolella tieteellisiä lehtiä oli jo niin paljon, että niiden seuraaminen kävi yksittäiselle ihmiselle vaikeaksi. Vastauksena tähän ongelmaan lehdet eriytyivät tieteenalakohtaisiksi 1800-luvulla. Kapeaan alaan erikoistuminen jatkui 1900-luvulla. Lehtien lukumäärä lisääntyi räjähdysmäisesti.

1900-luvun loppupuolella alkoi paperijulkaisujen ohella ja lisäksi artikkelien julkaiseminen internetissä. Osa tieteellisistä lehdistä julkaisee samat artikkelit sekä paperimuodossa että internetissä. On syntynyt elektronisia lehtiä, jotka julkaisevat aineistonsa pelkästään internetissä. Lisäksi merkittävä määrä artikkeleita on julkaistu pelkästään tutkijoiden omilla kotisivuilla. Internetin käyttöä artikkeleiden julkaisemisessa on lisännyt kirjastojen pienentyneet määrärahat ja tieteellisten lehtien määrän lisääntyminen. Kirjastoilla ei ole enää taloudellisia mahdollisuuksia hankkia kaikkia julkaistavia lehtiä kokoelmiinsa.

Tiedon määrän kasvaa jatkuvasti. Julkaisuja syntyy lisää ja artikkelit ovat pidempiä. Lisäksi yhden tutkimuksen tulokset voivat olla hajotettuna moneen artikkeliin. Suurimpana syynä artikkelien määrän lisääntymiseen on tieteen kehittyminen ja tieteentekijöiden suuri määrä. Osatekijä artikkelitulvaan on tieteellisen työn tulosten arviointi ja pisteyttäminen tekijän artikkeleiden saamien viittausten lukumäärän

perusteella. Pistelasku voi johtaa artikkeleiden määrän lisääntymiseen, sillä mitä enemmän tekijä julkaisee artikkeleita, sitä paremmat mahdollisuudet hänellä on saada viitteitä niihin. Tiedon tulva johtaa siihen että tieteenharjoittaja ei pysty perehtymään kaikkeen uuteen tietoon. Jopa tiukasti rajatun oman alan julkaisujen kattava seuraaminen voi olla mahdotonta.

1.2 Erilaisia tarpeita

Tieteellisen tiedon tuottajilla, etsijöillä ja lukijoilla on erilaisia tiedonhankinnan tarpeita. Artikkelin kirjoittaja haluaa saada palautetta tekemästään työstä. Lukija voi etsiä tiettyä yksittäistä tietoa, hakea uusia ajatuksia ja näkökulmia tai haluta pysyä ajantasalla aihepiirin kehityksessä. Sekä artikkeleiden, artikkelikokoelmien että viitetietokantojen pitäisi sopia näihin erityyppisiin käyttötilanteisiin.

Artikkelin tyypilliset lukutavat ovat koko artikkelin lukeminen alusta loppuun, ja mielenkiintoisten, itselle tärkeiden kohtien lukeminen. Artikkelin lukeminen alusta loppuun saakka on tyypillistä silloin, kun artikkeli käsittelee lukijalle ennestään tuntematonta aihetta. Artikkelin pitää johdattaa lukija aiheeseen, edetä johdonmukaisesti, käsitellä asia kattavasti ja tehdä selvä ero pääkohtien ja sivuseikkojen välille. Aiheeseen entuudestaan perehtyneet lukijat haluavat nähdä, mitä uutta artikkeli tuo aiheeseen. Tällaiset lukijat pomppivat kohdasta toiseen kiinnostuksensa mukaan. Artikkelin rakenteen pitäisi tukea kaikkia näitä erilaisia lukutapoja.

1.3 Tiedonhaku artikkeleista

Tiedonhaku on usein kaksiosainen tapahtuma: ensin pitää löytää relevantti artikkeli, ja sen jälkeen pitää artikkelista löytää halutun tiedon sisältävä kohta. Jotta artikkelin löytäminen olisi mahdollista, pitää artikkeliin liittää kuvailu artikkelin sisällöstä. Kuvailu voi olla tiivistelmä artikkelista tai esimerkiksi lista asiasanoja. Kuvailun pitää olla oikein tehty, kattava, tarkka ja selkeä. Artikkelin etsinnän onnistuminen tarkoittaa, että tulosjoukossa ei ole asiaankuulumattomia tai merkityksettömiä artikkeleita.

Löydetty artikkeli ei välttämättä käsittele pelkästään etsittävää tietoa. Jotta lukija ei joutuisi turhaan kahlaamaan koko artikkelia läpi, pitäisi artikkelista korostaa se kohta, jossa etsitty tieto on. Jos etsitty kohta on helposti pääteltävissä, voidaan korostus tehdä esimerkiksi muuttamalla tekstin fonttia kyseisessä kohdassa. Tällainen tilanne on esimerkiksi silloin kun tehdään asiasanahaku. Löydetty sana ja sen ympäristö korostetaan fonttimuutoksella. Toinen lähestymistapa ongelmaan on jakaa artikkeli pienempiin osiin.

Jokainen osa kuvaillaan erikseen, jolloin etsitty kohta artikkelista löytyy osan kuvailun perusteella. Tämä lähestymistapa perustuu artikkelin osien sisältöön, eikä yksittäisiin sanoihin.

2 Artikkeleiden kirjoittaminen modulaariseen muotoon

Amsterdamin yliopistossa toimii tutkimusryhmä, Communication in Physics Project (http://www.science.uva.nl/projects/commphys/), joka on paneutunut tieteellisen tiedon välitykseen fysiikassa [HaK98], [Har00], [KiH00]. Tavoitteena on löytää ratkaisuja tietotulvan hallintaan. Ryhmän tutkimus on teoreettista, eikä ryhmä ole ensisijaisesti kiinnostunut tulosten teknisestä toteuttamisesta. Artikkeleiden kirjoittaminen ryhmän ehdottamaan modulaariseen muotoon vaatisi kirjoittajan avuksi suunnitellun työkalun, jollaista ei ainakaan vielä ole olemassa.

2.1 Moduulien perusajatus

Tieteelliset artikkelit on perinteisesti kirjoitettu siten, että kukin artikkeli muodostaa itsenäisen kokonaisuuden. Toisaalta monet artikkelit muodostavat jatkoa edellisille artikkeleille, niissä kuvataan tietyn tutkimuksen uusimmat löydökset. Lukijan mielenkiinto ei yleensä kohdistu koko artikkeliin, vaan esimerkiksi vain artikkelissa esitettyyn uuteen ajatukseen. Artikkelin jako pienempiin itsenäisiin osiin, moduuleihin, helpottaa halutun tiedon löytämistä artikkelista. Moduuli on yhden asiakokonaisuuden sisältävä yksikkö, joka voidaan paikantaa, noutaa ja tarkastella ilman artikkelin muita osia. Jokaisella moduulilla on sitä kuvaavaa metadataa: bibliografisia tietoja ja asiasanoja. Läheisesti toisiinsa liittyvät moduulit voivat yhdessä muodostaa koosteisen moduulin. Myös koosteisella moduulilla pitää olla oma metadatansa.

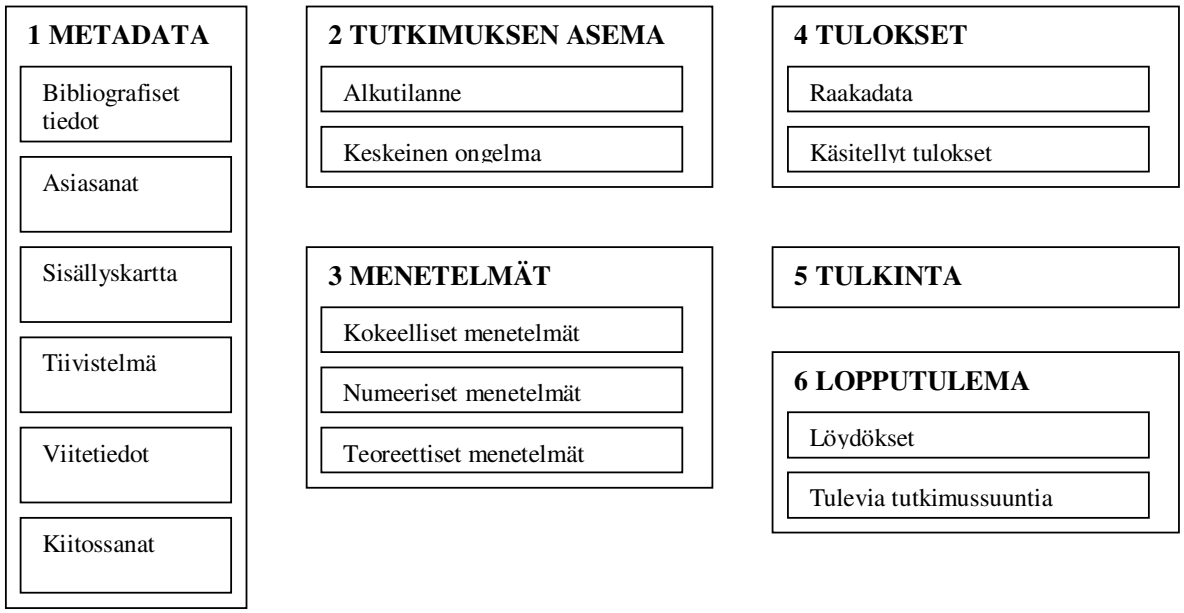
Moduulien väliset suhteet eivät rajoitu siihen, että ne voivat kuulua samaan koosteiseen moduuliin. Moduuleilla voi olla sekä rakenteellisia suhteita että sisällöllisiä suhteita saman artikkelin toisiin moduuleihin. Suhteita voi olla myös muiden artikkeleiden moduuleihin. Suhteita kuvataan linkeillä. Yksi linkki voi ilmaista monia ominaisuuksia. Linkki voi kuulua alkuperäiseen artikkeliin, mutta linkki voi olla myös myöhemmin lisätty viittaus toisesta artikkelista. Tämän takia linkkeihinkin pitää lisätä metadatana bibliografisia tietoja.

Moduulit voidaan jaotella niiden sisältämän tiedon vaikutusalueen mukaan. Makroskooppinen tieto on tieteenalan yleistä tietoa. Tällainen tieto on kerrottu alan perusteoksissa. Mesoskooppinen tieto on yhteistä kaikille tai monille tutkimusprojektin julkaisuille. Tällaista tietoa on esimerkiksi yleiskuvaus ratkaistavasta

ongelmasta tai selitys käytetystä mittauslaitteistosta, kun samaa laitteistoa käytetään pitkän aikaa. Mikroskooppinen tieto on vain kyseiseen artikkeliin kuuluvaa tietoa, uusi saavutettu tulos tai ajatus, varsinainen syy artikkelin julkaisemiseen. Yhdessä moduulissa ei saa olla kuin yhden tyyppistä tietoa.

2.2 Moduulijaon tekeminen sisällön mukaan

Perinteiset artikkelit noudattavat vakiintunutta jaottelua johdantoon, menetelmiin, tuloksiin, pohdintoihin ja loppupäätelmiin. Moduulijako on tehty tämän jaottelun pohjalta. Moduulijako on esitetty kuvassa 1.



Kuva 1. Tieteellisen artikkelin moduulit.

Numeroidut moduulit 1-6 ovat koosteisia moduuleja. Niiden osina olevat moduulit voivat olla joko koosteisia moduuleita tai perusmoduuleita. Poikkeuksen tähän muodostaa tulkintamoduuli, joka yleensä ei ole koosteinen, koska tulosten tulkinta on usein esseetyylistä pohdiskelua, joka olisi vaikea jakaa pienempiin osiin. Tulkintakin voidaan tarvittaessa jakaa laadulliseen ja määrälliseen tulkintaan.

2.3 Linkkityypit

Moduulien välisiä suhteita voidaan kuvata kuudella rakenteellisella ja kolmella sisältöön liittyvällä ominaisuudella. Ominaisuudet eivät ole toisiaan poissulkevia. Ominaisuudet voivat olla symmetrisiä, jolloin molemmat moduulit ovat samassa asemassa toisiinsa nähden, tai epäsymmetrisiä.

Rakenteelliset suhteet:

Hierarkkisuus (hierarchical relations) kuvaa perusmoduulin ja koosteisen moduulin välistä suhdetta. Ominaisuus on epäsymmetrinen (koostuu, on osana).

Läheisyys (proximity-based relations) kuvaa kahden moduulin etäisyyttä toisistaan. Samaan koosteiseen moduuliin kuuluvat moduulit ovat läheisiä. Läheisyyttä voi kuvata myös artikkelitasolla, moduulilla on läheisempi suhde samaan artikkeliin kuuluvaan moduuliin kuin toisen artikkelin moduuliin. Ominaisuus on symmetrinen.

Vaikutusalue (range-based relations) osoittaa siirtymisiä mikroskooppisten, mesoskooppisten ja makroskooppisten moduulien välillä. Ominaisuus on epäsymmetrinen.

Polut (path relations) vievät lukijan moduulista toiseen ennalta määritellyn suunnitelman mukaisesti. Poluilla annetaan lukijalle mahdollisuus tutustua artikkeliin ikään kuin se olisi lineaarista tekstiä. Ominaisuus on epäsymmetrinen (seuraava, edellinen).

Esitysmuotosuhde (representational relations) ilmaisee moduulien käsittelevän samaa asiaa eri muodossa, kuten esimerkiksi numerodata ja graafinen kaavio. Ominaisuus on epäsymmetrinen.

Hallinnollinen suhde (administrative relations) yhdistää metadatamoduuliin kaikki muut moduulit. Ominaisuus on epäsymmetrinen.

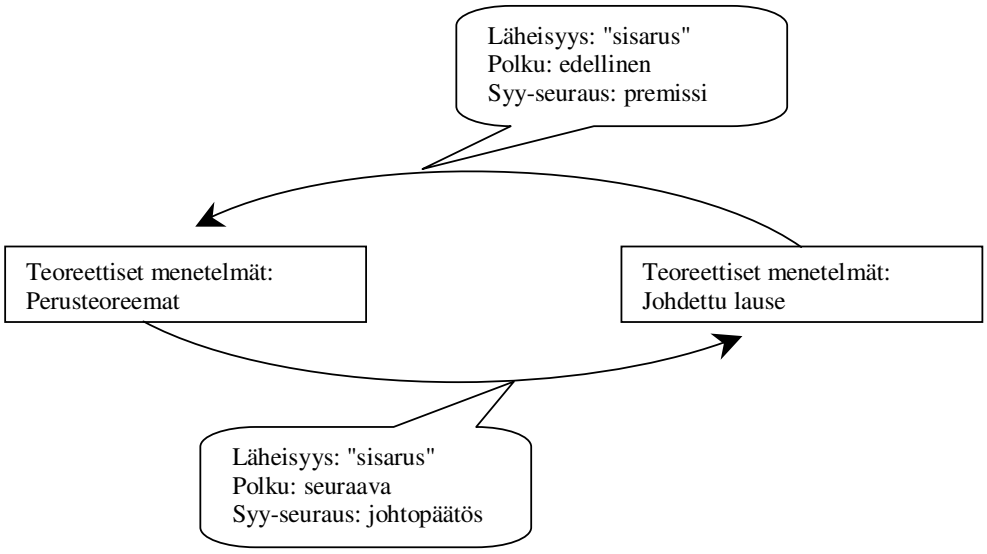
Sisällölliset suhteet:

Syy-seuraus-suhde (causal relations) yhdistää lähtöoletukset ja johtopäätöksen.

Vertailu osoittaa (comparison relations) tukeeko toinen moduuli lähdemoduulissa olevaa väitettä vai ei.

Luonteenomaiset suhteet (symptomatic relations) osoittavat jotain yhteyttä tai riippuvuutta moduulien välillä.

Kuvassa 2 on esimerkki kahden moduulin välisistä suhteista. "Teoreettiset menetelmät" on koosteinen moduuli. Perusmoduulit "Perusteoreemat" ja "Johdettu lause" ovat sisarussuhteessa, syy-seuraus-suhteessa ja artikkelin läpi vievän polun peräkkäiset moduulit.



Kuva 2. Kahden moduulin väliset suhteet.

2.4 Modulaarisuuden tuomat edut

Moduulin tiedon vaikutusaluetta voidaan käyttää hyväksi tietotulvan rajoittamisessa. Kun aiheeseen perehtynyt lukija haluaa pysyä ajantasalla alan kehityksessä, riittää että hän lukee artikkelista vain mikroskooppiset moduulit. Hän voi ohittaa artikkelista toiseen samanlaisina toistuvat mesoskooppiset moduulit. Mesoskooppisten moduulien käyttö vähentää myös artikkelin kirjoittajien työtä.

Modulaarisuus tehostaa tiedonhakua. Koska jokaisella moduulilla on omat asiasanansa, on yksittäisen tiedon löytäminen artikkelista helppoa. Haun tuloksena annetaan vain se moduuli, jossa hakusanat esiintyvät. Hakuja voi rajoittaa myös moduulin nimen perusteella. Esimerkiksi tietystä laitteistosta kiinnostunut lukija voi rajata haun koskemaan vain eri artikkeleiden "Kokeelliset menetelmät" -moduuleja.

3 Blusteinin menetelmä artikkelien muuttamiseksi hypertekstimuotoon

James Blustein on väitöskirjassaan [Blu99] tutkinut menetelmiä olemassaolevien tieteellisten artikkeleiden muuttamiseksi hypertekstimuotoon. Artikkeli säilytetään yhtenäisenä, mutta artikkeliin lisätään sekä rakennetta että sisältöä kuvaavia linkkejä.

Blustein pitää automaattisen linkityksen käyttöä välttämättömänä. Automaattinen linkitys perustuu annettuihin sääntöihin, ja linkitys noudattaa sääntöjä täsmällisesti. Linkityksestä tulee yhtenäisempi kuin käsin tekemällä. Jos säännöt ovat onnistuneesti laaditut, linkkejä ei puutu oleellisista kohdista, eikä linkkejä myöskään ole liikaa tai turhissa kohdissa. Oikea linkkien määrä on artikkelin käytön kannalta tärkeä tekijä. Jos linkkejä on liikaa, ei linkki tarkoita lukijalle enää mitään erityistä, se ei ilmaise tärkeää kohtaa tai kytköstä. Jos linkkejä on liian vähän, ei teksti juurikaan poikkea tavallisesta tekstistä. Automaattinen linkitys on myös nopea tehdä.

3.1 Linkkityypit

Artikkeleihin lisättiin kolmen tyyppisiä linkkejä: rakenteellisia linkkejä, määritelmälinkkejä ja sisällön samankaltaisuutta osoittavia linkkejä. Artikkeleihin ei lisätty retorisia linkkejä, kuten linkkejä vastaväitteisiin, väitettä tukeviin todisteisiin tai esimerkkeihin. Blustein käytti materiaalinaan tietojenkäsittelyyn liittyviä artikkeleita, eivätkä ne sisältäneet keskustelutyyppistä väittelyä.

Rakenteelliset linkit (structural links) kuvaavat tekstin rakennetta. Tällaisia ovat esimerkiksi sisällysluettelot, otsikot, alaviitteet ja lähdeviitteet. Linkit ovat jo tekstissä valmiina olemassa, ne vain pitää muuttaa hypertekstilinkeiksi. Hypertekstilinkit luodaan tekstin rakennemerkintöjen perusteella.

Määritelmälinkit (definition links) liittävät tekstin seassa olevan termin tai käsitteen termin selitykseen. Blusteinin tarkoituksena oli luoda määritelmälinkitkin automaattisesti. Hän lähti oletuksesta että määriteltävät termit esitetään tieteellisessä tekstissä kursivoituna. Tämä oletus ei pitänyt paikkaansa, kirjoittajat eivät käytä kursivointia säännönmukaisesti ja yhtenäisesti. Toinen oletus oli että määritelmät esiintyvät tekstissä tietynlaisissa lauserakenteissa, kuten "x tarkoittaa että" tai "määritellään x". Tämäkään oletus ei pitänyt paikkaansa. Kolmas mahdollisuus löytää määriteltävät termit perustui siihen että tällaiset termit esiintyvät tekstissä huomattavasti useammin kuin muissa artikkeleissa. Ajatusta ei lähdetty toteuttamaan, koska sanojen esiintymistiheyksien vertailu olisi ollut liian raskas operaatio. Niinpä Blustein päätyi luomaan määritelmälinkit käsin.

Määritelmälinkkiä ei lisätä, jos termi ja määritelmä ovat samassa tai peräkkäisissä kappaleissa. Lisäksi yhdessä kappaleessa saa olla tietystä termistä enintään yksi määritelmälinkki.

Sisällön samankaltaisuutta osoittavat linkit (semantic similarity links) jaetaan kahteen aliluokkaan. *Yhteenvetolinkit* (summary links) liittävät yhteenvedossa esiintyvän asian siihen kohtaan tekstiä, missä asiaa käsitellään. Esimerkiksi tiivistelmän tai loppuyhteenvedon lauseesta on linkki siihen kohtaan varsinaista tekstiä jossa asia käsitellään. *Aiheen perusteella yhdistävät linkit* (related links) liittävät yhteen ne tekstin osat, jotka käsittelevät samaa aihetta. Näiden linkkien on tarkoitus auttaa niitä lukijoita, jotka eivät lue artikkelia lineaarisesti alusta loppuun, vaan hyppivät artikkelissa kiinnostuksensa mukaan.

3.2 Sisällön samankaltaisuutta osoittavien linkkien automaattinen luominen

Yhteenvetolinkkien luonnissa oletetaan, että yhteenvedon jokainen lause kohdistuu eri osiin tekstissä, joko eri lukuihin tai eri kappaleisiin. Joka kappaleesta tutkitaan vain kaksi ensimmäistä lausetta, ja niiden

perusteella päätetään samanlaisuuden aste yhteenvedon lauseen kanssa. Kahden ensimmäisen lauseen tutkiminen perustuu havaintoon, että kirjoittajat heti kappaleen alussa kertovat kappaleen oleellisen sisällön, loppukappale vain syventää asiaa. Jos lauseen ja kappaleen välinen samankaltaisuus on riittävän suuri, luodaan välille linkki. Jos samasta yhteenvedon lauseesta lähtee useita linkkejä, valitaan niistä se jonka samankaltaisuus on suurin. Jos kaksi linkkiä on yhtä arvokkaita, tutkitaan linkkien osoittamien kappaleiden aliosia eli alilukuja ja seuraavia kappaleita. Linkeistä valitaan se, jonka kohteena olevan kappaleen ympäristössä samankaltaisuus on suurempi.

Aiheen perusteella yhdistäviä linkkejä luotaessa käytetään hyväksi James Allanin luomaa linkkien automaattista generointia. Allanin menetelmää on tässä seminaarissa käsitelty jo aiemmin Ilpo Lyytisen linkkityyppiesitelmässä. Käsiteltävä teksti pilkotaan virkkeisiin ja sanaryhmiin. Jokaisesta sanaryhmästä lasketaan samankaltaisuus tekstin virkkeisiin. Jos samankaltaisuus ylittää määritellyn rajan, tulee siihen kohtaan linkkiehdokas. Linkki hylätään saman tien, jos linkin toisena osapuolena on yhden sanan mittainen sanaryhmä tai virke. Samoin hylätään ne linkit, jotka yhdistävät saman luvun sisällä olevia sanaryhmiä ja virkkeitä.

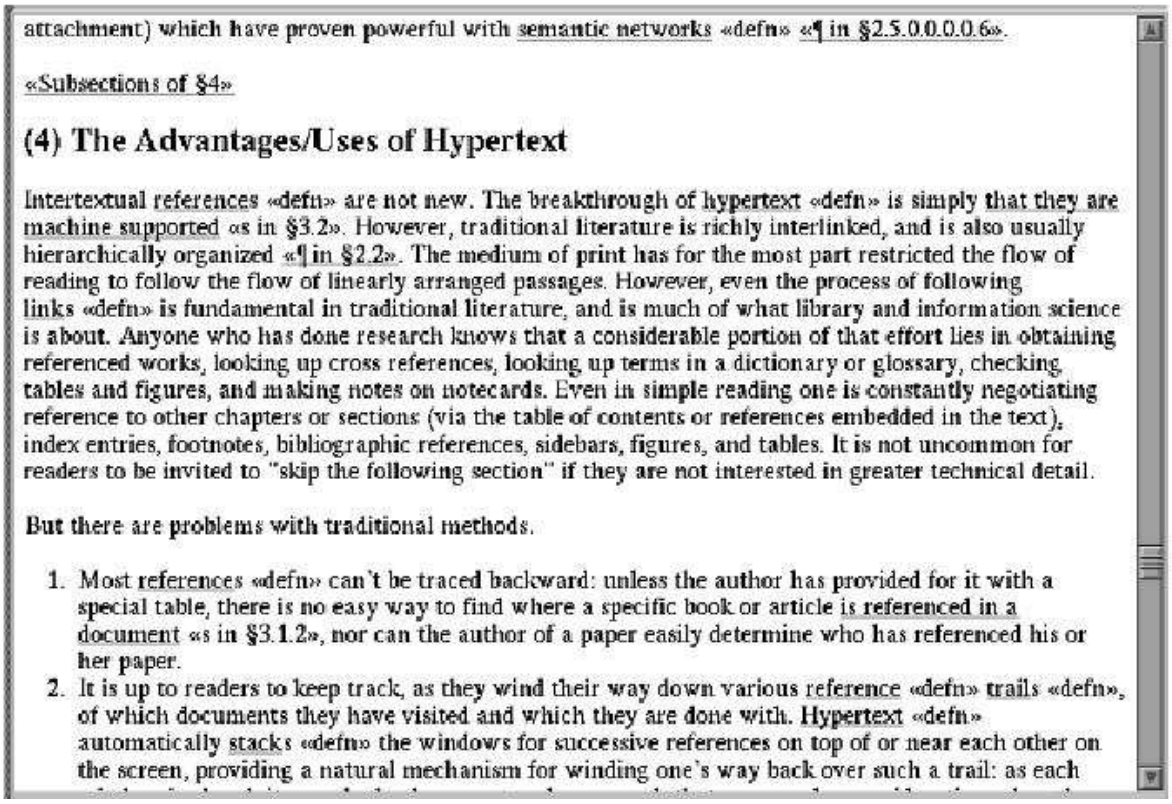
Sanaryhmien tarkastelun jälkeen tutkitaan kokonaisia virkkeitä. Ensin tutkitaan virkkeen samankaltaisuutta tekstin lukujen kanssa. Jos samankaltaisuus luvun kanssa ylittää määritellyn rajan, tutkitaan samankaltaisuutta kyseisen luvun alilukujen kanssa. Jos jokin aliluvuista antaa suuremman samankaltaisuuden, korvataan lukuun viittaava linkki alilukuun viittaavalla linkillä. Tämän jälkeen tutkitaan samankaltaisuutta aliluvun kappaleisiin, jos jokin kappaleista antaa suuremman samankaltaisuuden kuin alikuku, korvataan alilukuun viittaava linkki kappaleeseen viittaavalla linkillä. Menetelmää sovelletaan vielä kyseisen kappaleen virkkeisiin. Linkki yritetään tällä menetelmällä saada kohdistettua mahdollisimman tarkasti samaa aihetta käsittelevään tekstikohtaan.

Yhdestä virkkeestä saa lähteä vain yksi aihelinkki, joten jos virkkeessä on jo sanaryhmällä linkki, ei uutta koko virkkeeseen kohdistuvaa linkkiä enää lisätä, vaikka virkkeen samankaltaisuus annetun rajan ylittäisikin. Jos samasta kappaleesta on useampi linkki yhteen tekstikokonaisuuteen (esim. virkkeeseen tai kappaleeseen), jätetään jäljelle vain se linkki jonka samankaltaisuus on suurin. Muut linkit poistetaan.

3.3 Linkkien merkitseminen tekstiin

Linkit ja linkkien tyypit merkitään tekstin sekaan. Linkkien käyttötapa näkyy kuvassa 3. Linkki merkitään tavalliseen tapaan alleviivaamalla linkkisana. Tyyppi merkitään linkin perään. Tyyppien merkinnässä käytetään erilaisia kaksoiskulmasulkujen sisällä olevia määritteitä. Tyyppimääritteitä ovat esimerkiksi:

<<defn>>	määritelmälinkki
<<summ>>	yhteenvetolinkki
<<s in §>>	aihelinkki lauseeseen luvussa
<<¶ in §>>	aihelinkki kappaleeseen luvussa
<<§>>	aihelinkki lukuun


attachment) which have proven powerful with semantic networks «defn» «¶ in §2.5.0.0.0.0.6».

«Subsections of §4»

(4) The Advantages/Uses of Hypertext

Intertextual references «defn» are not new. The breakthrough of hypertext «defn» is simply that they are machine supported «s in §3.2». However, traditional literature is richly interlinked, and is also usually hierarchically organized «¶ in §2.2». The medium of print has for the most part restricted the flow of reading to follow the flow of linearly arranged passages. However, even the process of following links «defn» is fundamental in traditional literature, and is much of what library and information science is about. Anyone who has done research knows that a considerable portion of that effort lies in obtaining referenced works, looking up cross references, looking up terms in a dictionary or glossary, checking tables and figures, and making notes on notecards. Even in simple reading one is constantly negotiating reference to other chapters or sections (via the table of contents or references embedded in the text), index entries, footnotes, bibliographic references, sidebars, figures, and tables. It is not uncommon for readers to be invited to "skip the following section" if they are not interested in greater technical detail.

But there are problems with traditional methods.

1. Most references «defn» can't be traced backward: unless the author has provided for it with a special table, there is no easy way to find where a specific book or article is referenced in a document «s in §3.1.2», nor can the author of a paper easily determine who has referenced his or her paper.
2. It is up to readers to keep track, as they wind their way down various reference «defn» trails «defn», of which documents they have visited and which they are done with. Hypertext «defn» automatically stacks «defn» the windows for successive references on top of or near each other on the screen, providing a natural mechanism for winding one's way back over such a trail: as each

Kuva 3. Linkkien ja linkkityyppien käyttö tekstin seassa.

Koska linkkityyppien näyttäminen tekstin seassa häiritsee tekstin lukemista, voisi käyttöliittymää hieman muokata. Itse linkki olisi edelleen tekstin seassa, mutta linkkityypin kuvaus ja muut mahdolliset lisätiedot olisivat marginaalissa linkin kohdalla.

4 Cameronin maailmanlaajuinen lähdetiedot sisältävä viitetietokanta

4.1 Tietokannan perusidea

Robert Cameron on esittänyt ajatuksen bibliografisia tietoja ja lähdetietoja sisältävästä tietokannasta (universal citation database) [Cam97]. Tietokannassa olisi tiedot kaikista maailman tieteellisistä artikkeleista riippumatta siitä, miten, koska ja missä muodossa artikkeli on julkaistu. Tietokantaan kelpuutettaisiin yhtä

lailla perinteiset tieteellisissä aikakauslehdissä ja kausijulkaisuissa (journals) ilmestyneet artikkelit ja konferenssijulkaisuiden artikkelit, kuin tekniset raportit ja artikkeleiden esiversiotkin (preprints). Jokainen artikkeli olisi linkitetty kaikkiin niihin artikkeleihin, joita se käyttää lähteinään. Lisäksi artikkeli olisi linkitetty kaikkiin niihin artikkeleihin, jotka käyttävät sitä lähteenään. Tietokantaa päivitettäisiin joka päivä, joten se olisi koko ajan ajantasalla.

4.2 Tiedonhaku tietokannasta

Maailmanlaajuinen lähdetiedot sisältävä viitetietokanta on tehokas tiedonhaun väline. Uudet artikkelit viittaavat tärkeisiin aiemmin julkaistuihin artikkeleihin, joten uusinta tietoa aiheesta voidaan hakea etsimällä ne artikkelit, jotka viittaavat aiempaan tärkeään artikkeliin. Koska tietokantaa päivitetään jatkuvasti, ja koska tietokantaan otetaan mukaan myös artikkeleiden esiversiot, ei uuden tiedon leviämisessä synny mitään viiveitä.

Tietokantaan voi tehdä hakuja myös kirjoittajan nimen tai artikkelin otsikon sanojen perusteella. Hakua voi tehostaa suodattamalla tulosta. Artikkeleista voidaan esimerkiksi kelpuuttaa vain ne, jotka viittaavat kahteen tiettyyn artikkeliin. Suodatuksen voi tehdä myös vaatimalla, että tiivistelmässä esiintyy tietty sana. Aihepiirin tärkeimpien perusteosten löytämistä helpottaa viittausten määrän huomioiminen, tärkeisiin artikkeleihin viitataan paljon.

Lähdetietojen käyttö tiedonhaussa vaatii artikkelin kirjoittajalta huolellisuutta. Käytetyt lähteet pitää osata valita koko käsitellyn aihepiirin kattavasti, muuten kaikki kiinnostuneet lukijat eivät löydä artikkelia. Toisaalta tässä piilee liikaviittaamisen vaara, kirjoittajalle saattaa tulla houkutus ylenpalttiseen lähteiden käyttöön, jotta hän varmistaisi oman artikkelinsa näkyvyyden.

4.3 Tietokannan tuomat edut

Maailmanlaajuinen lähdetiedot sisältävä viitetietokanta tuo helpotusta uuden tiedon tulvaan. Tietokannan käyttäjä voi määritellä häntä kiinnostavat artikkelit ja antaa haluamansa suodatusehdot. Aina kun ehtoja vastaava artikkeli lisätään tietokantaan, saa käyttäjä ilmoituksen artikkelista. Koska tietokantaan kelpuutetaan kaikki artikkelit, eikä vain tietyissä lehdissä julkaistut artikkelit, saavat artikkelit yhtäläisen näkyvyyden julkaisutavasta riippumatta. Tämä saattaisi vähentää tarvetta hajottaa yksi artikkeli moneksi.

Artikkelin kirjoittaja saa tietokannan avulla nopeasti palautetta omasta työstään. Hän saa tiedon heti, kun joku käyttää hänen artikkeliaan lähteenään. Uusia näkökulmia tiettyyn aihepiiriin etsivä lukija hyötyy tietokannasta. Viittausketjuja seuraamalla hän pääsee etenemään aihepiiristä eri suuntiin. Uuteen aihepiiriin tutustuva lukija on hankalammassa asemassa. Jotta tietokantaa voisi käyttää tehokkaasti hyväksi, pitäisi jo etukäteen olla tiedossa aihepiiriin kuuluva artikkeli.

4.4 Tietokannan toteutus

Tietokanta voitaisiin toteuttaa internetin kautta toimivana hajautettuna tietokantana. Yliopistot ja tutkimuslaitokset ylläpitäisivät kukin omaa osuuttaan tietokannasta. Artikkelin kirjoittajan pitäisi lähettää tietokantaan artikkelinsa bibliografiset tiedot, lähdetiedot ja tiivistelmä. Jotta tietokannan kustannukset pysyisivät kurissa, ei artikkeleita indeksoitaisi lainkaan käsin. Artikkeleista jäisi siis puuttumaan asiasanat ja luokitus. Asiasanojen puuttuessa pitäisi artikkelin otsikon sisältää oleellisia avainsanoja.

Ehdotetunlaista tietokantaa ei ole tähän mennessä toteutettu [Law01]. Vaikeutena on riittävän yksimielisyyden ja tahdon saavuttaminen yliopistojen piirissä. Hanketta ei ehkä pidetä realistisena tai tarpeellisena. Kilpailevista hankkeista pitäisi luopua. Yhtenä suurena esteenä tietokannan toteutumiselle on artikkelien kirjoittajilta vaadittu aktiivinen panos viitetietojen ja lähdetietojen antamisessa.

Lähteet

Blu99 Blustein, J., Hypertext versions of journal articles: Computer-aided linking and realistic human-based evaluation. PhD thesis, University of Western Ontario, London, Ontario, March 1999.
ftp://ftp.csd.uwo.ca/pub/thesis/Blustein.PhD.Thesis.ps.gz

Lyhyt katsaus väitöskirjan keskeiseen sisältöön on sivulla J. Blustein's PhD Thesis in Brief
http://www.csd.uwo.ca/%7ejamie/Official/Proposal/thesis-brief.html

Cam97 Cameron, R. D., A universal citation database as catalyst for reform in scholarly communication. First Monday, 2, 4, 1997.
http://www.firstmonday.dk/issues/issue2_4/cameron/index.html

HaK98 Harmsze, F., Kircz, J., Form and content in the electronic age. IEEE-ADL'98 Advances in Digital Libraries Conference Session: Electronic Publishing: Defining the Technical and Scientific Information Package of the Future. Santa Barbara CA, USA. 22-25 April 1998.
www.science.uva.nl/projects/commphys/papers/adl98.pdf

Har00 Harmsze, F., A modular structure for scientific articles in an electronic environment. PhD thesis, Universiteit van Amsterdam, February 2000.
http://www.science.uva.nl/projects/commphys/papers/thesisfh/Front.html

KiH00 Kircz, J., Harmsze, F., Modular scenarios in the electronic age. Conferentie Informatiewetenschap 2000, De Doelen Rotterdam, April 5, 2000.
http://www.science.uva.nl/projects/commphys/papers/mod2k/mod2k.html

Law01 Lawrence, S., Access to scientific literature. The Nature Yearbook of Science and Technology, edited by Declan. Butler, Macmillan, London, England, pp. 86-88, 2001.
www.neci.nec.com/~lawrence/papers/access-nature01/access-nature01.pdf