hyväksymispäivä arvosana

arvostelija

Solmukokoelma dokumenttina Hypermediajärjestelmät - seminaari 26.9.2004

Joonas Muhonen

Helsinki 19.11.2004

Seminaaritutkielma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Matemaattis-luonnontieteellinen tiedekunta | Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Joonas Muhonen

Työn nimi — Arbetets titel — Title

Solmukokoelma dokumenttinaHypermediajärjestelmät - seminaari26.9.2004

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Seminaaritutkielma	19.11.2004	11 sivua

Tiivistelmä — Referat — Abstract

Laajat hajautetut hypermediaympäristöt sisältävät paljon web-sivuja jotka voivat olla joko yksittäisiä tai muodostaa koosteisia dokumentteja Nykyiset tiedonhakumenetelmät käsittelevät pääosin yksittäisiä web-sivuja, ja tämä rajoittaa niiden toimintaa.

Koosteisten dokumenttien käsittely parantaisi merkittävästi nykyisin käytössä olevien hakukoneiden toimintaa. Tällainen toiminnallisuus mahdollistaisi haut jotka kohdistuvat yhden sivun sijasta koko dokumenttiin.

Koosteisten dokumenttien löytäminen on ongelmallista. Suurimpana ongelmana on se, että koosteiset dokumentit on suunniteltu erottumaan lähinnä ulkoasunsa perusteella.

Jonkinlaisena ratkaisuna tähän voidaan pitää HTML-määrityksen käyttämistä koosteisten dokumenttien rakenteen esittämiseen, sekä hakukoneiden laajentaminen ymmärtämään koosteisia dokumentteja.

ACM Computing Classification System (CCS): H.5.4 Hypertext/Hypermedia

Avainsanat — Nyckelord — Keywords
koosteiset dokumentit

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto **1**

2 Koosteiset dokumentit ja haku **2**

2.1 Koosteiset dokumentit 2

2.2 Haku koosteisista dokumenteista 3

3 Menetelmät **3**

3.1 Aloituspisteiden löytäminen 4

3.2 Dokumentin sivujen löytäminen 5

3.2.1 Tuotantoprosessin mallintaminen 5

3.2.2 Linkkien rakenne 6

3.2.3 Linkkiverkon rakenne 7

3.2.4 Linkkien muoto 7

3.2.5 Linkkityypit 8

3.3 URL-osoitteiden muoto 8

4 Yhteenveto **9**

Lähteet **10**

1 Johdanto

Tämän tutkielman kannalta tärkeimmät käsitteet ovat dokumentit, *URL-osoitteet* (Uniform Resource Locator address) ja *web-sivut* (Web page) [EiM03]. Aina ei ole kuitenkaan selvää mikä on näiden termien välinen yhteys. Teknillisessä kirjallisuudessa yhdestä web-sivusta puhutaan usein dokumenttina, mutta termin yleisesti käytetty merkitys ei kuitenkaan ole tämä. Usein useat web-sivut muodostavatkin yhdessä yhden dokumentin.

Toisaalta URL-osoiteen ymmärretään usein viittaavan yhteen web-sivuun. URL-osoitteen ja web-sivun välisen yhteydet löytämisen voidaankin olettaa tuottavan ongelmia, sillä aina ei ole suinkaan selvää mikä osa URL-osoitteesta identifioi yksittäisen web-sivun.

Dokumentin erottamisen perustarve liittyy siihen, että hypermedian etsinnän tavoitteena on usein löytää yksi dokumentti [LCV02]. Useasta osasta, esimerkiksi web-sivusta, koostuvaa dokumenttia kutsutaan *koosteiseksi dokumentiksi* (compound document). Tällä hetkellä hakukoneiden käytön tehokkuutta rajoittaa se seikka, että pääosa niistä kohdistaa hakunsa yksittäisiin web-sivuihin. Vain muutama hakukone käyttää hyväksi koosteisten dokumenttien tietoja [CHC99]. Jotkin hakukoneet ottavat huomioon sivuun kohdistavat viittaukset muilta web-sivuilta [BrP98], mutta eivät välitä näiden sivujen sisällöstä.

Tämä seminaaritutkielma jakautuu viiteen kappaleeseen. Tämä kappale toimii johdantona tutkielmaan. Toisessa kappaleessa esitellään tutkielman keskeiset käsitteet, ongelmat ja syyt aiheen tutkimukseen. Kolmannessa kappaleessa esitetään menetelmiä ja ratkaisuja koosteisten dokumenttien erottamiseen. Neljännessä kappaleessa esitellään muutamia järjestelmiä jotka liittyvät tutkittavaan aiheeseen. Viimeisessä kappaleessa tehdään yhteenveto aiheesta.

2 Koosteiset dokumentit ja haku

Laajat hajautetut hypermediaympäristöt sisältävät paljon web-sivuja jotka voivat olla joko yksittäisiä tai muodostaa koosteisia dokumentteja [EiM03]. Nykyiset tiedonhakumenetelmät käsittelevät pääosin yksittäisiä web-sivuja, ja tämä rajoittaa niiden toimintaa.

2.1 Koosteiset dokumentit

Hypertekstin historian alkuvaiheissa dokumentit olivat yleensä yhdestä URL-osoitteesta löytyviä valmiita kokonaisuuksia [EiM03]. Hypertekstin navigaation kehittyessä huomattiin, että dokumentin jakaminen useaan eri web-sivuun parantaa selailumahdollisuuksia, vähentää turhaa verkkokuormaa ja helpottaa usean kirjoittajan samanaikaista työskentelyä.

Sama voidaan todeta myös muun verkkomedian osalta. Varsinkin aiemmin web-sivuilla oli tapana jakaa isot kuvat osiin, jotta näitä osia voitaisiin ladata nopeammin valmiiksi. Nykyään tämä tapa on vähentynyt, mutta esimerkiksi videoleikkeitä paloitellaan yhä. Tavoitteena voi olla esimerkiksi yhden ison tiedoston jakaminen jollenkin tietylle tallennusmedialle mahtuviksi osiksi.

Hypertekstisiä dokumentteja voidaan siis julkaista kahdessa muodossa: yhtenä yhdestä web-sivusta koostuvana dokumenttina tai monesta sivusta koostuvana koosteisena dokumenttina [LCV02]. Usein on tarkoituksen mukaista, että dokumentista on kaksi versiota, esimerkiksi yksi selailua ja yksi tulostamista varten. Voidaan kuitenkin todeta, että verkosta löytyvät dokumentit ovat tulevaisuudessa pääosin koosteisia dokumentteja, joissa kaikki hypertekstin tarjoamat mahdollisuudet on käytetty hyväksi. Tulostamista varten on kuitenkin olemassa muita siihen tarkoitukseen paremmin sopivia julkaisumuotoja.

2.2 Haku koosteisista dokumenteista

Verkossa tapahtuvien hakujen tavoitteena on yleensä löytää jokin tietty dokumentti [LCV02]. Tutkimus altavista-hakukoneen tietokantaan [SHM98] osoittaa, että verkkohaut koostuvat usein vain muutamasta termistä. Tällaisten yksinkertaisten hakujen syynä voidaan pitää sitä, että useamman termin käyttö johtaisi hakutulosten karsiutumiseen, sillä useat useat termit eivät välttämättä löydy etsityn dokumentin samalta web-sivulta. Käyttäjät joutuvatkin karsimaan hakusanoja, ja ennalta arvaamaan, mitkä hakusanat ovat varmasti samalla web-sivulla heidän hakemissaan dokumenteissa.

Tutkimuksessa huomattiin myös, että suuri osa näistä lyhyistä muutaman termin kyselyt oli itse asiassa lauseita. Valmiit lauseet lienevät kaikista helpompia arvauksia tekstistä, joka varmasti on samalla web-sivulla. Tästä voidaan muotoilla johtopäätös, että kyselyn tekijät varmistelevat myös tällä kyselyn sanojen asettumista samalle verkkosivulle.

Kaikenkaikkiaan voidaan tästä voidaan vetää se johtopäätös, että jo nyt olisi tarvetta tehokkaalle hakukoneelle joka osaisi ottaa huomioon samaan koosteiseen dokumenttiin kuuluvat sivut palauttaessaan hakutuloksia. Nykyisten hakukoneiden käytössä on nahtävissä selvästi käytösmalleja joilla yritetään korjata hakukoneiden puutteita.

3 Menetelmät

Koosteisia dokumentteja tai *koosteita* (aggregate) on tutkittu lähes siitä lähtien kun hypertekstiä on ollut olemassa [BoS91]. Läpi hypertekstin historian ihmisillä on ollut vaikeuksia luoda käsitystä siitä, mitkä web-sivut itse asiassa muodostavat yhdessä suuremman kokonaisuuden. Hyperteksti on kuitenkin kehittynyt, ja web-sivujen navigaation kehittymisen myötä käyttäjät osaavat ymmärtää ja navigoida paremmin

koosteisia dokumentteja muodostavilla web-sivustoilla. Tämä ei kuitenkaan helpota dokumenttien etsintää, vaan navigaation lisäksi tarvitaan menetelmiä koosteisten dokumenttien automaattiseen löytämiseen.

Koosteisten dokumenttien löytäminen alkaa dokumentin *aloituspisteen* (entry point, leader) etsinnällä. Aloituspisteen löydyttyä aletaan dokumentin linkkejä käydä läpi, ja pyritään tätä kautta löytämään ne sivut jotka kuuluvat dokumenttiin.

3.1 Aloituspisteiden löytäminen

Aloituspisteellä tarkoitetaan sitä URL-osoitetta, ja sen kautta web-sivua, joka toimii jonkinlaisena aloituspisteenä koosteiselle dokumentille [EiM03]. Tämä on usein sisällysluettelo tai otsikkosivu. Mikäli sellaista ei ole, se on sivu, joka lukijoiden on tarkoitus nähdä ensimmäisenä, tai URL-osoite joka on tarkoituksella määritetty ulkoista linkitystä varten, sillä kun koosteinen dokumentti tuodaan verkkoon, sen aloituspisteeseen luodaan usein hyperlinkki.

Näiden aloituspisteiden tunnistaminen on ensisijaisen tärkeää, sillä niistä on parasta aloittaa koosteisen dokumentin läpi kahlaaminen. Lisäksi aloituspistettä voidaan käyttää hyväksi koosteiseen dokumentin kohdistuvan etsinnän tuloksia esiteltäessä.

Aloituspisteen on hyvä olla mahdollisimman keskeinen dokumentin kannalta. Siitä tulisi olla mahdollisimman lyhyet yhteydet muualle dokumenttiin – mahdollisimman tasapuolisesti.

Useissa hakemistoissa on oletusarvoinen aloitussivu, ja se toimiikin usein hyvänä ehdokkaana koosteisen dokumentin aloituspisteeksi. Tällaiset sivut on usein helppo tunnistaa tiedostonimen ulkoasusta.

Mikäli dokumentista löytyy hakemisto tai sisällysluettelo, se on usein erittäin hyvä aloituspiste. Hakemisto tai sisällysluettelo voidaan löytää siihen viittaavien linkkien ulkoasun perusteella. Usein siihen viittaavissa linkeissä lukee juuri “hakemisto” tai

"sisällysluettelo".

Kun koosteiseen dokumenttiin viitataan, kohteena on usein juuri sellainen web-sivu dokumentista, josta on mahdollisimman hyvät yhteydet muualle dokumenttiin. Sivu johon on useita linkkejä muualta on myös hyvä ehdokas aloituspisteeksi.

3.2 Dokumentin sivujen löytäminen

Kun dokumentille on löydetty aloituspiste, täytyy määrittää ne web-sivut jotka kuuluvat kyseiseen koosteiseen dokumentiin [EiM03].

Yksi lähtökohta dokumentin rajaamiseen lähtee siitä periaatteesta, että jos joukko web-sivuja voidaan ajatella suunnattuna verkkona, pitää koosteisen dokumentin muodostaa puu tähän verkkoon. Toisin sanoen joltain verkossa olevalta sivulta pitää olla yhteys jokaiseen dokumentin sivuun.

Tämä määritelmä ei missään tapauksessa ole tarpeeksi rajaava tarpeeseemme. Kun tarkastellaan lähemmin olemassa olevia koosteisia dokumentteja, huomataan että nämä muodostavat lähes täydellisen verkon, eli lähes jokaiselta dokumentin sivulta lähtee linkki lähes jokaiselle muulle dokumentin sivulle.

Toinen seikka jolla dokumenttia voidaan rajata, on URL-osoitteiden hierarkkinen rakenne, joka määritellään URI-määrityksessä [BFM98]. Osoittautuu, että yksittäiset koosteiseen dokumentin web-sivut kuuluvat lähes aina samaan URL-osoitteen määrittelemään osoitteeseen aina polkuosan viimeiseen *jakoviivamerkkiin* ("/", slash) saakka.

3.2.1 Tuotantoprosessin mallintaminen

Mikäli dokumentti on tuotettu jollain työkalulla, syntyy tuloksista yleensä tietyn kaltaisia [EiM03]. Työkalujen tuottamien sivujen malleja tutkimalla on helppo lu-

otettavasti tarkistaa mitkä web-sivut kuuluvat samaan koosteiseen dokumenttiin.

Tämä tuottaa kuitenkin merkittäviä ongelmia, sillä työkalujen on valtava määrä, ja niiden uusien versioiden seuraaminen olisi vielä suurempi tehtävä. Lisäksi tämä ratkaisutapa ei käy tilanteisiin joissa dokumenttin tuottamiseen ei käytetä mitään ennestään tunnettua työkalua.

3.2.2 Linkkien rakenne

Koosteisissa dokumenteissa hyperlinkit voidaan jakaa kahteen ryhmään [EiM03]. Dokumentin sisäisiin linkkeihin, ja siitä pois johtaviin linkkeihin. Käytännössä linkin ryhmä voidaan yleensä määritellä linkin rakenteen perusteella.

Linkit voidaan jakaa rakenteensa perusteella viiteen eri linkkityyppiin:

1. ulkoiset linkit: linkki *web-sivustolta* (web site) toiselle,
2. ristikkäiset linkit: linkki hakemistohierarkkiasta toiseen,
3. alaspäin osoittavat linkit: linkki hakemistohierarkkiassa alaspäin,
4. ylöspäin osoittavat linkit: linkki hakemistohierarkkiassa ylöspäin ja
5. sisäiset linkit: linkki hakemistohierarkkiassa samalle tasolle.

Jokainen linkkityyppi saattaa sisältää tietoa koosteisten dokumenttien löytämisestä. Sisäiset linkit ovat pääasiassa linkkejä samaan koosteiseen dokumenttiin. Ulkoiset ja ristikkäiset linkit osoittavat usein jonkin toisen koosteisen dokumentin aloituspisteeseen. Ylöspäin ja alaspäin osoittavat linkit taas osoittavat jonkinverran keskimääräistä useammin samaan koosteiseen dokumenttiin, ja jos näin on, niin silloin ne yleensä osoittavat tietyn ylä- tai alaosion aloituspisteeseen.

3.2.3 Linkkiverkon rakenne

Kun koosteisen dokumentin web-sivujen muodostamaa verkkoa analysoidaan, huomataan että se yleesä muistuttaa vähintään yhtä seuraavista verkkorakenteista [EiM03]:

1. suora polku: koosteisen dokumentin läpi kulkee yksi järjestetty polku,
2. täydellinen verkko: jokailta dokumentin web-sivulta on linkki jokaiselle toiselle,
3. ratas: yhdeltä sivulta on linkki jokaiselle muulle sivulle ja
4. monitasoinen dokumentti: dokumentti muodostaa selkeän hierarkkisen rakenteen.

Edellä mainitut rakenteen eivät ole toisiaan poissulkevia. Myöskään tällaisen rakenteen olemassaolo ei osoita varmuudella että rakenne muodostaisi yksittäisen koosteisen dokumentin.

3.2.4 Linkkien muoto

Yksi tapa luokitella web-sivuja on käyttää hyväksi linkkien muotoa [EiM03]. Koska koosteiset dokumentit on yleensä kirjoitettu suhtellisen lyhyessä ajassa, ja dokumentin yksittäisten sivujen luontiin on käytetty yleensä samaa tekniikkaa, osoittautuu että ulos dokumentista viittaavat linkit ovat muodoltaan samankaltaisi. Käytännössä tämä johtuu usein siitä, että linkit on luonut sama julkaisujärjestelmä tai -työkalu samaa valmista pohjaa käyttäen.

Toinen linkkien muotoa käyttävä heuristiikka on yleisten linkkitekstien käyttö websivujen luokitteluun [EiM03]. Tällaisia yleisiä linkkitekstejä ovat “seuraava” ja “edellinen” suoran polun muodostavissa dokumenteissa, “hakemisto” ja “sisällysluettelo” ratasmuotoisisssa dokumenteissa ja numeroidut linkit täydellisen verkon muodostavissa dokumenteissa. Näitä yleisiä tekstejä voidaan myös tunnistaa samassa samal-

la URL-hierarkkialla olevia web-sivuista joilla on samalla tavalla nimettyjä sisäisiä linkkejä. Tällä menetelmällä päästään eroon kieli- ja julkaisutyökaluriippuvuudesta.

3.2.5 Linkkityypit

HTML-määritys (HTML-specification) [RHJ99] sisältää menetelmän kertoa koosteisten dokumenttien sivujen keskinäisten suhteiden kuvaamiseen "link-type" attribuutilla. Tällä attribuutilla kirjoittaja voi kuvata osoitettavan sivun suhteen selattavaan sivuun, ja sitä kautta hyvin yksinkertaisella tavalla paljastaa koko dokumentin rakenteen.

Ikävä kyllä tätä rakennetta ei yleisesti juuri käytetä. Lisäksi hyvin harvat julkaisutyökalut luovat tuotettavien web-sivujen linkkeihin tämän attribuutin.

Syynä rakenteen vähäiseen käyttöön lienee se, että sen käytöllä ei toistaiseksi ole juurikaan merkitystä. Erittäin harva HTTP-asiakas esittää mitenkään näitten attribuuttien arvoja. Kun linkkityyppien tuki paranee HTTP-asiakkaissa on odotettavissa, että myös niiden tuottamiseen käyttään enemmän vaivaa.

3.3 URL-osoitteiden muoto

Edelä mainitut linkkien hierarkkiseen suhteeseen perustuvat tekiikat eivät toimi mikäli web-sivut on järjestetty jollain muulla tavalla. Theodor Nelson on todennut [Nel97], että hierarkkinen dokumenttien organisointi on epäluonnollista.

Yksi yleinen tapa organisoida web-sivuja on käyttää jotain tiettyä URL-osoitteen *perusosaa* (base URI), ja yksilöidä dokumentti jollain attribuutilla *kyselyosassa* (query string). Tällöin dokumenttien hierarkiaa ei voida mitenkään päätellä pelkästää osoitteen perusteella.

Harvinaisempaa sen sijaan on se, että koosteinen dokumentti sijoitettaisiin selkeästi

eri hakemistoihin, jonkin muun hierarkian mukaan [EiM03].

4 Yhteenveto

Koosteisten dokumenttien käsittely parantaisi merkittävästi nykyisin käytössä olevien hakukoneiden toimintaa [LCV02]. Tällainen toiminnallisuus mahdollistaisi haut jotka kohdistuvat yhden sivun sijasta koko dokumenttiin.

Koosteisten dokumenttien löytäminen on ongelmallista. Suurimpana ongelmana on se, että koosteiset dokumentit on suunniteltu erottumaan lähinnä ulkoasunsa perusteella [MiT99]. Dokumentin kirjoittajat olettavat usein että heidän dokumenttinsa ymmärretään kokonaisuudeksi ulkonäön perusteella.

Toinen ongelma on se, että kirjoittajat olettavat että heidän dokenttinsa löydetään selailemalla siten, että lukijalle on jäänyt käsitys koko sivuston rakenteesta [MiT99]. Näin selailijalla olisi koko ajan selkeä käsitys siitä mitä kaikkea sivustoon kuulu, ja missä yhteydessä yksittäisen web-sivu tieto esitetään. Todellisuus on kuitenkin toisenlainen. Useimmat web-sivut löydetään hakemalla, eivätkä hakukoneet toistaiseksi ymmärrä sivujen muodostamia rakenteita.

Jonkinlaisena ratkaisuna tähän voidaan pitää HTML-määrittelyn käyttämistä koosteisten dokumenttien rakenteen esittämiseen, sekä hakukoneiden laajentaminen ymmärtämään koosteisia dokumentteja. Tällöin koosteisiin dokumentteihin kohdistuvat haut löytäisivät kohteensa ilman että jokaisella web-sivulla on toistettu kaikki dokumentin avainsanat. Ongelmaksi saattaa kuitenkin osoittautua tällaisten laajojen hakujen aikavaatimukset.

Mielenkintoisena seikkana missään tutkimuksessa ei oltu tutkittu suhteellisten linkkien käytön merkitystä siihen, osoittaako linkki toiseen osaan dokumenttia. Ainakin allekirjoittaneella on sellainen käsitys, että dokumentteja tuotettaessa käytetään suhteel-

lisia linkkejä usein dokumentin sisäisissä linkeissä, kun taas dokumentista ulos viittaavat linkit sisältävät yleensä täydellisen URL-osoitteen.

Lähteet

BFM98 T. Berners-Lee, R. T. Fielding ja L. Masinter. Uniform Resource Identifiers (URI): Generic syntax. RFC 2396, 1998. URL `http://www.ietf.org/rfc/rfc2396.txt`.

BrP98 S. Brin ja L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

BoS91 R. A. Botafogo ja B. Shneiderman. Identifying aggregates in hypertext structures. *UK Conference on Hypertext*, sivut 63–74, 1991.

CHC99 M. Chen, M. A. Hearst, J. H. ja J. Lin. Cha-cha: A system for organizing intranet search results. *USENIX Symposium on Internet Technologies and Systems*, 1999.

EiM03 N. Eiron ja K. S. McCurley. Untangling compound documents on the web. *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, sivut 85–94. ACM Press, 2003.

LCV02 W.-S. Li, K. S. Candan, Q. Vu ja D. Agrawal. Query relaxation by structure and semantics for retrieval of logical web documents. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):768–791, 2002.

MiT99 Y. Mizuuchi ja K. Tajima. Finding context paths for web pages. *Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots*, sivut 13–22. ACM Press, 1999.

Nel97 T. H. Nelson. Embedded markup considered harmful. *World Wide Web J.*, 2(4):129–134, 1997.

RHJ99 D. Raggett, A. L. Hors ja I. Jacobs. The html 4.01 specification, 1999. URL `http://www.w3.org/TR/html4/`.

SHM98 C. Silverstein, M. Henzinger, H. Marais ja M. Moricz. Analysis of a very large altavista query log. Tekninen raportti 1998-014, Digital SRC, 1998.