*Paavo Arvola**

Tekstiympäristön vaikutus XML -tiedonhaun täsmäytyksessä ja arvioinnissa

Olemassa olevan tiedon saatavuus on kaiken teknillisen, tieteellisen ja taloudellisen kehityksen perusedellytyksiä. Digitaalisen materiaalin tuotanto on viime aikoina kasvanut rajusti, ja valtaisa määrä materiaalia on yhtäkkiä tullut yleisesti saataville. Pelkkä saatavilla olo ei luonnollisestikaan riitä, ja ilman tehokkaita tiedonhakujärjestelmiä, tai kansanomaiselta nimeltään hakukoneita, digitaalisen materiaalin saatavuus ja siten niistä saatava hyöty jäisi olemattomaksi.

Sähköisillä hakumenetelmillä tehtävä tiedonhaku on arkipäiväistynyt internetin tiedonhakujärjestelmien myötä, ja yhä enenevässä määrin hakuja tehdään käyttäen pieniruutuisia mobiililaitteita. Tämä on johtanut siihen, että eräs keskeisimmistä tiedonhaun menetelmien kehittämisen tavoitteista on saavuttaa alati tarkempia hakutuloksia. Tämä tarkoittaa muun muassa sitä, että pitkistäkin dokumenteista oleellinen sisältö pyritään osoittamaan hakijalle tarkasti. Tiedonhakija pyritään siis vapauttamaan turhasta dokumenttien selaamisesta.

Tiedonhaun tutkimus on monitieteistä, perustuen ainakin tietojenkäsittelytieteeseen , matematiikkaan, informaatiotutkimukseen, kielitieteeseen ja tilastotieteeseen. Tiedonhaun tutkimuksen tarkoituksena on kehittää menetelmiä ja järjestelmiä, jotka löytävät hakijalle hyödyllistä ja oikeaa tietoa tietolähteistä, ja pienentämään käyttäjälle tiedon etsimisestä koituvan vaivan määrää. Väitöskirjani käsittelee XML-tiedonhakujärjestelmän kehittämistä ja arvioimista. Tällaisen tiedonhakujärjestelmän tavoitteena on juuri tarkasti osoittaa käyttäjälle hänen tarvitsemansa tieto. Apuna tämän tavoitteen saavuttamisessa on hyödyntää tietolähteiden rakennetta.

Tietojenkäsittelyn perinteessä tietolähteet luokitellaan akselille, jonka toisessa päässä ovat data-orientoituneet lähteet, kuten tietokannat, ja toisessa päässä teksti-orientoituneet lähteet, kuten tekstidokumentit. Data-orientoituneista lähteistä oikeaa vastausta haetaan eksaktilla kyselyllä, kun taas teksti-orientoituneessa tiedon haussa esitetään epätarkempi, yleensä hakusanoihin perustuva kysely, ja vastaukset, yleensä dokumentit, pyritään järjestämään niiden oletetun hyödyllisyyden mukaan.

Jaottelu teksti ja data-orientoituneisiin tietolähteisiin ei kuitenkaan ole tarkkarajainen, vaan lähteissä on aste-eroja niiden teksti ja dataorientoituneisuuden suhteen. Rakenteisen ja data-orientoituneen, tiedon seassa voi olla vapaata tekstiä sisältäviä osia, joiden avulla voidaan järjestää tietokannassa olevia rakenteellisia yksiköitä käyttäjän kyselyn perusteella.

Toisaalta tekstidokumentissa saattaa usein olla lisäksi data-orientoituneita piirteitä. Esimerkiksi kirja voi tekstin lisäksi sisältää selkeästi data-orientoituneita osia, kuten tekijän nimen, isbn -numeron, viiteluettelon, painopaikan, ja niin edelleen. Lisäksi kirjaan liittyy metadataa, kuten kirjan hinta ja kirjasta mahdollisesti esitettyjä arvioita. Käyttäjä voi tarkentaa kyselyään tällaisten data-orientoituneiden osien avulla.

Perinteiset tietokantakyselyt ovat data-orientoituneita ja pohjautuvat tietokantaratkaisuihin, joissa hakijan oletetaan tuntevan haettavien kohteiden ominaisuudet ja niiden väliset suhderakenteet. Kysely tällaiseen tietolähteeseen esitetään erityisellä tietokantakyselykielellä tai se syötetään käyttöliittymänrajapinnan kautta lomakkeella, jonka jälkeen järjestelmä automaattisesti kääntää sen tietokannan kyselykielelle. Jos vastauksia

---

* Paavo Arvolan väitöskirja The Role of Context in Matching and Evaluation of XML Information Retrieval (Tekstiympäristön vaikutus XML -tiedonhaun täsmäytyksessä ja arvioinnissa) tarkastettiin Tampereen yliopistossa 18.6.2011.

kyselyyn on useita, niiden järjestys voi perustua esimerkiksi aakkosjärjestykseen tai on käyttäjän näkökulmasta sattumanvarainen. Data-orientoitunut haku perustuu ennalta määrättyyn tietokantakaavioon kun taas tekstiorientoitunut sanahaku kohdistuu vapaamuotoiseen tekstiin.

Vaikka teksti-orientoitunut tieto onkin vapaamuotoista, se ei suinkaan ole rakenteetonta, vaan sillä on rakenne, joka on yhtäältä hierarkkinen, eli kirja on jaettu lukuihin, jotka puolestaan on jaettu alilukuihin ja nämä otsikoihin ja tekstikappaleisiin. Toisaalta rakenne on sekventiaalinen, eli tekstin osat seuraavat toinen toistaan lukemisjärjestyksessä.

XML soveltuu niin tekstin kuin datankin mallintamiseen ja kuvaamiseen. Se on erittäin yleisesti käytössä oleva sähköisen julkaisemisen standardi, jolla voidaan kuvata tekstin osien merkitystä sekä osien välisiä hierarkkisia ja sekventiaalisia suhteita. XML:n eräs käyttötapa onkin mallintaa vapaata tekstiä sisältävien dokumenttien, kuten kirjojen, rakennetta, ja antaa roolit eri dokumentin osille, esimerkiksi otsikoille tai bibliografisille tiedoille. Tällaisia dokumentteja kutsutaan rakenteisiksi dokumenteiksi. Rakenteisille dokumenteille on tyypillistä hierarkkinen eli puumainen rakenne: dokumentti jakautuu osaelementteihin, ja yläelementit sisältävät alaelementtejä. XML mahdollistaa näin dokumenttien sisäisen rakenteen hyödyntämisen. Toisin sanoen XML rakennetta voidaan hyödyntää kehitettäessä tarkkuusorientoituneita eli kohdennettuja tiedonhakujärjestelmiä ja -menetelmiä.

Digitaalisen tiedon yhteydessä, tiedon rakenne on olennaista tarkan prosessoinnin takia. Käyttäjä sen sijaan ei tunne rakenteita, joko kokonaan tai osittain, tai rakenteet voivat olla hyvin monimutkaisia. Tietomäärän kasvun myötä, myös tiedon rakenne on käynyt monimuotoisemmaksi. Käytäntö on osoittanut, että käyttäjät haluavat saada tarkkaa tietoa rakenteen suhteen epätarkoilla hakuilmauksilla, kuten esimerkiksi käyttämällä hakusanoja. Järjestelmien kehityksen kannalta tämä muodostaa haasteen.

XML-tiedonhaun tutkimuksen ja kehityksen tavoitteena on auttaa käyttäjää hakemaan tietoa data- ja tekstiorientoituneista lähteistä yksinkertaisilla ja sumeilla hauilla tuntematta tietokannan rakennetta kenties ollenkaan. XML-tiedonhaussa haku suoritetaan tietokantaan pelkistetyimmillään vain käyttäen sanahakua, mutta myös rakenteeseen kohdistuvat kyselyt ovat mahdollisia. Tässäkin tapauksessa XML-tiedonhaku tarjoaa saatujen hakutulosten oletettuun hyödyllisyyteen perustuvan lajittelun.

Teksti-orientoineessa lähestymistavassa käyttäjä antaa tietokantakyselyn sijasta tyypillisesti vain hakusanoja. Dokumenttihaussa vastauksena saadaan dokumentteja, jotka sopivat parhaiten, eli täsmäävät, annettuun kyselyyn. Toisin sanoen tiedonhakujärjestelmien tehtävänä on järjestää tekstidokumentit niiden oletetun hyödyllisyyden suhteen. Järjestelmä päättelee dokumenttien hyödyllisyyden käyttäjän antamasta kyselystä ja tämä päättely toteutetaan automaattisesti täsmäyttämällä käyttäjän antama kysely haun kohteena oleviin dokumentteihin. Esimerkiksi sanomalehden arkistossa täsmäytys perustuu hakusanojen tilastolliseen esiintyvyyteen dokumentissa. Järjestelmät tyypillisesti olettavat, että mitä enemmän ja tiheämmin kyselyssä esiintyvät sanat esiintyvät myös dokumentissa, sen parempi dokumentti. Toisaalta kyselyssä, jossa on useampi hakusana, yhden hakusanan erottelukyky on tyypillisesti parempi kuin toisen. Kyselyn automaattisessa prosessoinnissa näiden sanojen keskinäinen painotus riippuu niiden yleisyydestä koko haettavassa aineistossa. Harvinaiset sanat saavat suuremman painon kuin yleiset. Lisäksi täsmäytyksen apuna voidaan käyttää kyselyilmauksen automaattista laajentamista esimerkiksi erilaisten käsitehierarkioiden, tai dokumenttien välisten linkitysrakenteiden avulla.

Dokumenttia pienempien hakuyksiköiden, kuten XML -elementtien, täsmäytys on periaatteessa samankaltaista kuin dokumenttihaussa, mutta pienemmissä yksiköissä oleva tekstievidenssi, eli sanojen määrä, on huomattavasti vähäisempää, ja näin lyhyen elementin hyödyllisyyden automaattinen tunnistaminen on vaikeampaa kuin pitkän. Toisaalta voidaan olettaa, että lähekkäin olevat elementit kertovat usein osapuilleen samasta asiasta ja tähän oletukseen perustuen väitöskirjassa esitetään kontekstualisointi, menetelmä, jossa tätä XML-tiedonhaulle tyypillistä tekstievidenssin vähäisyyttä täydennetään hyödyntämällä XML-hierarkian ylempien tai rinnakkaisten osien sisältöä. Tämä menetelmä on väitöskirjassa osoitettu empiirisin mittauksin tehokkaaksi. Tehokkuudella tarkoitetaan tässä sitä, kuinka hyvin järjestelmät löytävät vastauksia niille esitettyihin kyselyihin.

Tiedonhakujärjestelmien tehokkuuden kehittäminen perustuu siis mittaamiseen, ja tämä mittaaminen tapahtuu erityisen, yksinkertaistetun tutkimusasetelman avulla. XML-tiedonhaun tutkimusasetelma on perinteisen dokumenttihaun tutkimusasetelman johdos. Dokumenttihaun tutkimusasetelmassa asetelmassa järjestelmän palauttama dokumentti katsotaan relevantiksi eli hyväksi vastaukseksi, mikäli se vastaa aiheensa puolesta käyttäjän antamaa kyselyä. Esimerkiksi, jos käyttäjä haluaa dokumentteja Talvisodassa olleesta Lemetin motista, vain tästä tapahtumasta kertovat dokumentit ovat relevantteja tämän hakuaiheen suhteen. Järjestelmiä palkitaan sen mukaan kuinka hyvin ne näitä relevantteja dokumentteja löytävät ja toisaalta rangaistaan epärelevanttien löytämisestä. Toisin sanoen tiedonhakujärjestelmälle syötetään kysely, ja järjestelmä on sitä parempi, mitä paremmin se kunkin hakuaiheen relevantit dokumentit tunnistaa.

Relevanssin määritelmä, joka perustuu dokumentin aiheenmukaisuuteen, on periaatteessa objektiivinen, mutta koska tutkitaan automaattista järjestelmää, dokumentin relevanssin arvioimista ei voida tehdä automaattisesti. Sen sijaan koeasetelmaa varten tutkijat kehittelevät joukon testikyselyitä ja arvioivat etukäteen määrätyn kokoelman dokumentteja joko relevanteiksi tai epärelevanteiksi kunkin testikyselyn suhteen. Eli esimerkiksi kertooko dokumentti Lemetin motista vai ei. Nämä relevanssiarviot tarvitsee tehdä vain kerran, jonka jälkeen näitä voidaan käyttää yhä uudelleen eri hakujärjestelmien vertailussa.

Hakuyksiköiden erilaisen luonteen vuoksi kohdennetun ja XML-tiedonhaun tutkimiseen tarvitaan lisäksi erikoismenetelmiä. Perinteisen dokumenttitiedon haun tutkimusasetelman perusoletukset ovat kyllä sinänsä katsottu perusteiltaan riittäviksi, mutta relevanssin aiheenmukaisuuden määritelmä ei yksin riitä. Kun tutkitaan kohdennettua hakua, tulee lisäksi pohtia mikä on lisäksi mahdollisimman tarkka, mutta silti riittävä vastaus käyttäjän antamaan kyselyyn. Esimerkiksi talvisotaa käsittelevässä dokumentissa tai kirjassa vain yksi luku tai jopa yksi tekstikappale käsittelee Lemetin mottia. Tällöin koko dokumentti on relevantti, mutta aihetta käsittelevä tekstikappale on paljon tarkempi vastaus kyselyyn. Kohdennettu tiedonhakujärjestelmä pyrkii siis sellaiseen otokseen relevantista dokumentista, jossa on mahdollisimman paljon relevanttia ainesta mutta mahdollisimman vähän epärelevanttia.

Dokumentin aiheenmukaisuuden arvioissa tutkijoiden tulee lisäksi osoittaa, mitkä osiot dokumentin sisällä ovat relevantteja. XML -tiedonhaun tutkimuksessa voidaan esimerkiksi arvioida erikseen kunkin elementin aiheenmukaista kattavuutta ja tarkkuutta. Nykyään molempien aspektien arviointi joka dokumentin osalle on katsottu kumminkin liian työlääksi ja arviointia on yksinkertaistettu siten, että relevantista dokumentista merkitään relevantti teksti. Muu teksti on epärelevanttia. XML -tiedon haun lisäksi tämä mahdollistaa muunkin kohdennetun tiedonhaun, kuten tekstikatkelmahaun tutkimisen. Järjestelmän palauttaman tekstikatkelman relevanssiarvo on relevantin tekstin määrän suhde koko tekstikatkelman tekstin määrään. Esimerkiksi, jos järjestelmä palauttaa tekstikatkelman, josta tasan puolet on arvioitu relevantiksi, järjestelmä saa tästä 0,5 pistettä.

Kohdennettujen tiedonhakujärjestelmien mittaaminen on ollut varsin järjestelmäorientoitunutta, ja mittaamisessa on sivuutettu muun muassa järjestys, jossa teksti luetaan dokumentin sisällä. Mittareissa on sisäänrakennettuna oletus, että käyttäjä lukee kaiken järjestelmän palauttaman materiaalin ja vain sen. Järjestelmän saamat pisteet perustuvat palautetussa tekstissä ja dokumentissa olevaan relevantin tekstin määrään ja sen suh-teeseen koko tekstin määrään. Tämä on johtanut siihen, että varsin pitkiä tekstikatkelmia palaut-taneet järjestelmät ovat saaneet korkeampia pisteitä kuin lyhyitä tarkkoja palauttaneet järjestelmät, kyseenalaistaen näin koko kohdennetun tiedon-haun hyödyllisyyden tekstiorientoituneissa tieto-lähteissä. Tällainen mittaaminen ei kuitenkaan ota huomioon monia käyttäjään liittyviä muuttujia. Todellisuudessa käyttäjä saattaa lukea dokumenttia vain sen verran kuin katsoo tarpeelliseksi ja siirtyä lukemaan seuraavaa dokumenttia, tai lukea myös järjestelmän osoittaman materiaalin ympäristöä, eli kontekstia. Väitöskirjassa kehitetyissä kohdennetun tiedonhaun arviointimenetelmissä on erilaisten käyttäjäsimulaatioiden avulla pyritty ottamaan huomioon edellä mainitut, dokumenttikontekstiin liittyvät seikat, ja näiden seikkojen huomioonottamisella on ollut merkittävä vaikutus mittaustuloksiin ja sitä kautta eri järjestelmien keskinäiseen paremmuusjärjestykseen.