Informaatiotutkimuksen päivät 2010
21. - 22. lokakuuta, Tampere

ABSTRAKTI

Paavo Arvola, Marko Junkkari, Jaana Kekäläinen

XML tiedonhaku

Paavo Arvola, Informaatiotutkimuksen ja interaktiivisen median laitos, Tampereen yliopisto, paavo.arvola@uta.fi

Marko Junkkari, Tietojenkäsittelytieteen laitos, Tampereen yliopisto, junken@cs.uta.fi

Jaana Kekäläinen, Informaatiotutkimuksen ja interaktiivisen median laitos, Tampereen yliopisto, jaana.kekalainen@uta.fi

Tiedonhaun tarkoituksena on löytää hakijalle relevantti informaatio. Perinteisesti informaatiolähteet luokitellaan data-orientoituneeksi tai teksti-orientoituneeksi (Elmasri & Navathe 2004). Data-orientoituneessa lähestymistavassa etsitään oikeaa vastausta, kun taas teksti-orientoituneessa lähestymistavassa etsitään parasta vastausta. Perinteiset tietokantakyselyt ovat data-orientoituneita ja pohjautuvat tietokantaratkaisuihin, joissa hakijan oletetaan tuntevan haettavien kohteiden ominaisuudet ja niiden väliset suhderakenteet. Haun tarkoitus on tyypillisesti myös eroteltava kohde. Tarkastellaan esimerkiksi kyselyä: ”Anna kirjat, joita Vonnegut on kirjoittanut”, joka sisältää tarkasti haun kohteen (kirja) sekä sen ominaisuuden (Vonnegut). Tällainen kysely esitetään tietokantakyselykielelle (kuten SQL) tai se syötetään ennalta ohjelmoidun käyttöliittymän kautta järjestelmään, joka automaattisesti kääntää sen tietokannan kyselykielelle. Tulokset voidaan järjestää mm aakkosjärjestyksen perusteella. Teksti-orientoineessa lähestymistavassa käyttäjä antaa tietokantakyselyn sijasta vain avainsanoja mahdollisine laajennuksineen (Fraasit, +/- etuliitteet, Boolen operaattorit). Tuloksena saadaan dokumentteja, jotka sopivat parhaiten annettuun hakulausekkeeseen. Tyypillisesti dokumentit on järjestetty niiden relevanssin todennäköisyyden perusteella. Järjestämiseen voidaan käyttää mm. sanojen esiintymistiheyttä, käsitteellistä kyselyn laajentamista tai ontologioita. Hakusana ”Vonnegut” tuottaisi laajemman hakutuloksen, koska mukana olisivat kaikki kokoelman dokumentit, jotka sisältävät hakusanan sekä myös dokumentteja, joita mahdollinen laajennus on tuottanut (esim. dokumentteja tralfamadorilaisista).

Data-orientoitunut lähestymistapa perustuu ennalta määrättyyn tietokantakaavioon kun taas teksti-orientoitunut lähestymistapa kohdistuu vapaamuotoiseen tekstiin. Tämä jaottelu on kuitenkin vanhanaikainen, eikä se vastaa nykypäivän tiedonhaun tarpeisiin. Toisin sanoen dokumentti sisältää tyypillisesti sekä data- että teksti-orientoineita piirteitä. Otetaan esimerkiksi satunnainen kirja. Tämä sisältää teksti-orientoituneita osia (varsinainen teksti), mutta myös data-orientoitunutta informaatiota (tekijä, nimi, isbn etc.). Lisäksi kirjaan liittyy metadataa, joka on jokaiselle tietojärjestelmälle spesifiä. Esimerkiksi kirjastoissa yksittäinen kirja voi olla lainattavissa tai vain luettavissa. Lisäksi teksti-orientoituneella tiedolla on lähes poikkeuksetta rakenne: Kirja on jaettu

lukuihin, jotka jaettu alilukuihin jne. Rakenne on tyypillisesti hierarkkinen. Uusi lähestymistapa pyrkii mallintamaan data-orientoineet piirteet, teksti-orientoineet piirteet sekä dokumenttien rakenteen.

XML (eXtensible Mark-up Language) on standardi, jolla voidaan kuvata tekstin osien merkitystä sekä osien välisiä hierarkkisia suhteita (http://www.w3.org/XML/). XML:n eräs käyttötapa onkin mallintaa vapaata tekstiä sisältävien dokumenttien, kuten kirjojen, rakennetta, ja antaa roolit eri dokumentin osille, esimerkiksi otsikoille tai bibliografisille tiedoille. XML on metakieli, jonka avulla luodaan kohdedokumenttien rakennetta kuvaava merkkauskieli. Merkattuja dokumentteja kutsutaan rakenteisiksi dokumenteiksi. XML-dokumenteille on tyypillistä hierarkkinen eli puumainen rakenne: dokumentti jakautuu osaelementteihin; yläelementit sisältävät alaelementtejä. XML soveltuu niin tekstin kuin datankin mallintamiseen ja kuvaamiseen.

Sähköisten kokoelmien kasvu, hakujen arkipäiväistymisen ja pieniruutuisten mobiililaitteiden myötä tiedonhaun menetelmien kehittämisen tavoitteena on saavuttaa alati tarkempia hakutuloksia. Tämä tarkoittaa sekä haun kohdistamista tulosdokumenttien relevantteihin osiin että vähäisempää tarvetta tulosdokumenttien selailuun, milloin vastaus käyttäjän hakuun on jossain dokumentin keskellä. Lisäksi tulosdokumentti voidaan koostaa useiden eri dokumenttien osista. Dokumenttien merkkaaminen XML:n tai sitä muistuttavien merkkauskielien, kuten HTML tai XHTML, avulla niin internetissä kuin muussa sähköisessä julkaisemisessa on johtanut siihen, että tiedonhakujärjestelmiä ja -menetelmiä on kehitetty käyttämään dokumenttirakenteita. XML -muotoiseen tietoon kohdistuvaa hakua kutsutaan XML-tiedonhauksi (XML information retrieval) (Lalmas 2009). XML-tiedonhaku perustuu dokumentin rakenteen ja osien merkkauksen hyödyntämiseen tiedonhaussa, ja sen yhtenä tavoitteena on osoittaa hakijalle dokumentin relevantit osat.

XML-tiedonhaun tutkimuksen ja kehityksen tavoitteena on auttaa käyttäjää hakemaan tietoa data- ja tekstikokoelmista yksinkertaisilla ja sumeilla hauilla tuntematta tietokannan rakennetta kenties ollenkaan (Trotman & Lalmas 2006). XML-tiedonhaussa haku suoritetaan tietokantaan pelkistetyimmillään vain käyttäen avainsanahakua, mutta myös rakenteeseen kohdistuvat kyselyt ovat mahdollisia (Trotman & Sigurbjörnsson 2004). Lisäksi XML-tiedonhaku tarjoaa saatujen hakutulosten oletettuun relevanssiin perustuvan lajittelun. Yhteisten hakukielien ja -järjestelmien kehittäminen data- ja teksti-orientoituneelle tiedolle on yhä aktiivinen tutkimuskohde.

XML-tiedonhaun tutkimuksen haasteet ovat paitsi relevanssin arvioimisessa kokodokumenttihakua vähäisemmästä tekstimassasta, myös tulosten mielekkäässä ja käyttäjäystävällisessä esittämisessä. TRIX (Tampere Retrieval and Indexing for XML) ryhmä on vuodesta 2004 kehittänyt menetelmiä tehokkaaseen XML –tiedonhakuun ja tiedonhaun evaluointiin. Eräs tällainen menetelmä on kontekstualisointi (contextualization) (Arvola & al. 2005, Kekäläinen & al. 2009), jossa XML-tiedonhaulle tyypillistä tekstievidenssin vähäisyyttä on kompensoitu hyödyntämällä XML hierarkian ylempiä osia dokumentin alempien osien painotuksessa.

XML–tiedonhaun tuloksellisuuden mittaaminen muodostaa oman erityisongelmansa. XML-tiedonhaun luonteen vuoksi perinteiset tiedonhaun laboratoriomenetelmät ja -mittarit eivät

sellaisenaan sovellu XML-tiedonhaun tehokkuuden mittaamiseen. TRIX–ryhmä onkin kehittänyt XML-tiedonhaun evaluointiin menetelmiä, jotka perustuvat käyttäjien ja käyttöliittymien simulointiin (Arvola & Kekäläinen, 2010). Tähän simulointipohjaiseen malliin perustuvia mittareita (Arvola & al. 2010) käytetään kansainvälisen INEX hankkeen virallisissa menetelmäarvioissa (http://www.inex.otago.ac.nz/). TRIX -yhmän toimintaa on rahoitettu Suomen Akatemian projekteissa n:o 115480 ja 130482 .

Arvola, P., Junkkari, M., Kekäläinen, J. (2005). Generalized contextualization method for XML information retrieval, In Proc. of CIKM 2005, 20-27.

Arvola, P., Kekäläinen, J. (2010). Simulating User Interaction in Result Document Browsing, In proc. of SimInt at SIGIR 2010.

Arvola, P., Kekäläinen, J., Junkkari, M. (2010) Expected reading effort in focused retrieval evaluation, To appear in Information Retrieval. 25 sivua.

Elmasri, R., Navathe, S. (2004). Fundamentals of Database Systems, 4th ed., Addison-Wesley.

Kekäläinen, J., Arvola, P., Junkkari, M. (2009). Contextualization. Encyclopedia of Database Systems 2009. 474-478.

Lalmas, M. (2009). XML Retrieval (Synthesis Lectures on Information Concepts, Retrieval, and Services), Gary Marchionini Editor, Morgan and Claypool Publishers, 112 sivua.

Trotman, A. and Lalmas, M. (2006). Strict and vague interpretation of XML-retrieval queries. In Proc. of SIGIR 2006. 709-710.

Trotman, A., Sigurbjörnsson, B. (2004). Narrowed Extended XPath I (NEXI). In Proc. of INEX 2004, 16-40.