Informaatiotulva ja sen hallinta

Tallennetun informaation räjähdysmäinen kasvu on saanut aikaiseksi sen, että tiedon tallennuksesta ja –hausta ja sen tutkimisesta on tullut eräs tärkeimmistä informaatiotutkimuksen osa-alueista. Lisäksi tiedon tallennuksesta ja –hausta on tullut eräs mielenkiintoinen rajankäyntikohde tietojenkäsittelytieteen ja informaatiotutkimuksen välimaastoon. Tämäkin raja alkaa olla kuin veteen piirretty viiva, molemmat alat joutuvat ylittämään sen jatkuvasti. Vielä kymmenisen vuotta sitten tietojenkäsittelyoppi pelkisti tiedon haun pitkälti perusdatan käsittelyyn ja indeksointiin. Viime vuosina tietojenkäsittelyn alalla on jouduttu ottamaan entistä enemmän huomioon merkityksen muodostumisen prosessi kokonaisuutena ja inhimillisenä toimintana. Pelkkien merkkijonojen tunnistaminen ei riitä välineeksi ympäristössä, jossa sekalaisia dokumentteja syntyy vähintäänkin miljoonan tallennetun yksikön päivävauhdilla.

Erääksi tärkeäksi tutkimuskohteeksi onkin noussut erilaisten dokumentaaristen rakenteiden automaattinen tunnistaminen ja toisaalta, varsinkin informaatiotutkimuksessa, sellaisten kuvailuvälineiden laatiminen, jotka tehokkaammin kuvailisivat merkityksiä, joita dokumenttien avulla halutaan välittää. Tietojenkäsittelytieteessä ja teknisissä tieteissä tämä on tarkoittanut viime aikoina dokumenttien dataan sisäänrakennettujen merkitysrakenteiden löytämiseen tarkoitettujen algoritmien kehittämistä. Käytännössä tämä tarkoittaa esimerkiksi digitoitujen äänitteiden analysointia puheentunnistuksen keinoin ja digitoitujen kuvien tunnistusta hahmoanalyysin avulla.

Kielitieteessä on jo pitkään tiedetty, että jo foneettisella tasolla voidaan havaita määrättyjä merkityksiä. Toisaalta kielitieteessä ja viestinnän tutkimuksessa on todettu jo pitkään, että kaikkia merkityksiä ei voida palauttaa foneettiselle tasolle. Automaattisen tiedonhaun ja indeksoinnin kannalta onkin hyvin mielenkiintoista se, löydetäänkö näiltä matalimmilta esittämisen tasoilta riittävästi merkitystä välittäviä ja erottavia elementtejä, jotta informaation tallennus ja –haku voitaisiin toteuttaa tyhjentävästi automaattisilla algoritmeilla. Tällä alueella tapahtuukin merkittävää tutkimustoimintaa monella tieteenalalla.

Mielenkiintoista on myös se, että informaatioalan perinteiset välineet - dokumentteihin liitetty indeksointi, luokitus sekä tiivistelmät - ovat säilyttäneet ja todennäköisesti myös säilyttävät tehonsa dokumenttien tallennuksen ja -haun välineinä jatkossakin. Rakenteisten dokumenttien kuvailukielten ja standardien kehittyessä edelliset välineet tulevat hyödynnettyä tehokkaammin myös avoimessa tiedon julkaisemis- ja tallennusympäristössä. Allekirjoittanut onkin jo jonkin aikaa ollut sitä mieltä, että varsinkin luokitus tulee kokemaan renessanssin vaikka se välillä on jäänyt indeksoinnin jalkoihin. Tämä sen vuoksi, että luokitus on ja tulee olemaan tehokkain tapa jäsentää laajoja tietomassoja.

Lisäksi luokitus voidaan nähdä tapana tulkita olemassa olevaa informaatiota ja sen sisältöjä. Eli kuten Kwasnik asian ilmaisee, luokitus on tapa nähdä asioita, hahmottaa niiden välisiä suhteita ja rakenteita. Luokitus on aina myös kunkin (osa)kulttuurin tapa kertoa siitä, mitä se pitää näkemisen arvoisena. Luokituksen merkitystä informaation järjestäjänä kuvaa sekin, että ohjelmointitekniikassa luokat ja niihin perustuva olio-ohjelmointi otettiin käyttöön, kun perinteiset ohjelmointitavat eivät enää kyenneet tehokkaasti hallitsemaan merkityksiä ja niiden välittämistä.

Informaatiotutkimuksen kannalta haastavan ongelmakentän muodostaa myös tietokannoista ja niihin tallennetuista dokumenteista löytyvän, niin kutsutun kätketyn informaation etsiminen. Viime vuosina onkin esille noussut tämän tapaisen tietämyksen löytämiseen, tallennettuun dataan kohdistuva "kaivostoiminta" ja siihen liittyvien välineiden kehittäminen. Tähän kuuluvat tekniikat ovat hedelmällisiä etenkin laajoissa, kokonaiset dokumentit tallentavissa tietokannoissa. Näiden tekniikoiden avulla on mahdollista luoda myös uutta tietämystä ja tietoa tietokantoihin talletetun datan analyysin avulla. Tässä kehitystyössä on olennaista tieteitten välinen toiminta, perusdatan käsittelytekniikoista aina laajojen ja sisällöllisten rakenteiden tulkintaan ja tämän tulkinnan mallintamiseen.

Onkin hyvin mielenkiintoista nähdä, mitä voidaan löytää edellä mainittujen tekniikoiden avulla esimerkiksi vanhasta suomalaisesta painetusta aineistosta, kun sen laajamittainen digitointi saadaan käynnistettyä. Dokumenttien rakenteiden kuvauskielten ja niihin liitettyjen kohdentavien tiedonhakualgoritmien avulla pystytään todennäköisesti ratkaisemaan suuriin datamassoihin liittyviä tiedon tallennuksen ja –haun käytännön erityisongelmia. Varsinkin humanistiselle ja yhteiskuntatieteelliselle tutkimukselle tulee todennäköisesti avautumaan uusia näköaloja ja perus-

aineistoja laajojen tietomassojen rakenteiden ja dokumenttien välisten suhteiden analyysin avulla.

Erään käytännön ongelman uusien tekniikoiden käyttöönotossa muodostaa varsinkin suomalaisissa kirjastotietokannoissa se, ettei niissä ole kovinkaan paljoa tätä esille kaivettavaa dataa. Pelkkä perusluettelointidata muutamine sisällönkuvailuelementteineen kun ei anna tarpeeksi tartuntapintaa uusille tekniikoille ja välineille. Onkin harmi, ettei monessakaan suomalaisessa perustietokannassa ole tallennettu esimerkiksi tiivistelmiä puhumattakaan, että sisällönkuvailu – on se sitten toteutettu luokituksen tai asiasanoituksen avulla – olisi tehty monipuolisesti ja fasetoidusti. Tässä onkin alamme käytännön toiminnassa paljon parantamisen ja kehittämisen varaa.

Tämän lehden artikkelit käsittelevät tiedonhaun ongelmia. Tässä alallamme on tehty viime vuosina merkittävää työtä ja tästä saatu palaute on ollut aina kansainvälistä tasoa myöten kiittävää. Kuten artikkeleista huomaa, asettaa uusi, laajoihin datamassoihin perustuva digitaalinen ympäristö entistä suurempia vaatimuksia tiedontallennukselle ja -haulle. Erityisen merkittäväksi laajoissa tietokantaympäristöissä muodostuu relevanttien dokumenttien löytäminen ja tähän liittyvien menetelmien kehittäminen ja testaaminen. Liian laajat ja runsaasti hälyä sisältävät viitejoukot ovat jo pitkään olleet avointen ja laajojen tiedon tallennus- ja hakuympäristöjen ongelma. Tämän vuoksi tällä alalla tehtävä kehitystyö on ensiarvoisen tärkeää myös käytännön tietopalvelutoiminnalle.

Kuopiossa 20.11.2000.

Jarmo Saarti