Tiedonlouhinta rakenteisista dokumenteista (seminaarityö)

Miika Nurminen (minurmin@jyu.fi)

19.11.2003

Tiivistelmä

Seminaarityössä käsitellään tiedonlouhintaa ja sen tärkeimpiä osa-alueita, kuten tekstitiedon ja WWW:n louhintaa. Erityisesti keskitytään rakenteisten dokumenttien käsittelyyn ja klusterointimenetelmiin. Rakenteisuus mahdollistaa tietokoneen tulkittavissa olevan tiedon sisällyttämisen dokumenttiin. Linkeistä ja metatiedosta saatavaa lisätietoa voidaan käyttää dokumentin analysoinnissa. Lopuksi hahmotellaan dokumenttityypin mukaan mukautettavissa oleva XML-dokumenttien klusterointisovellus.

1 Johdanto

Tietokannoissa ja Internetissä olevan tiedon määrä on kasvanut viime vuosina kiihtyvää vauhtia. Miljoonia tietueita sisältävien tietovarastojen hahmotus ei ole manuaalisin keinoin käytännöllistä tai edes mahdollista. Eräs mahdollinen ratkaisu ”infoähkyyn” on tiedonlouhinta.

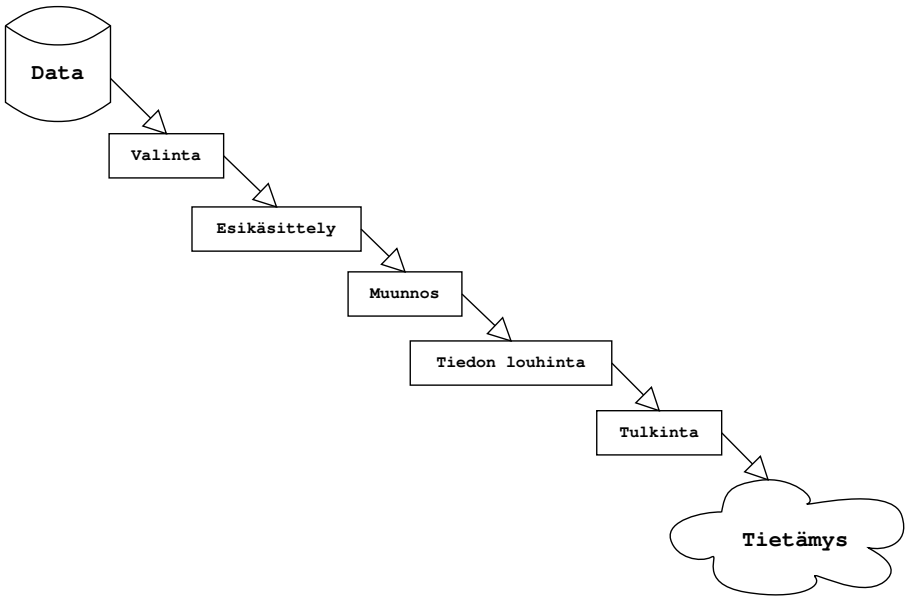
Tiedonlouhinta (tiedonrikastus) on suurten tietojoukkojen (tietokantojen, dokumenttien, mittaustietojen...) analyysia ja mallien muodostamista. Seminaarityössä keskitytään erityisesti XML-dokumentteihin kohdistuvaan tiedonlouhintaan. Tiedonlouhintamenetelmistä käsitellään klusterointia, eli tietoalkioiden ryhmittelyä niiden keskinäisen samanlaisuuden perusteella.

Seminaarityö jakautuu seuraaviin osiin: luvussa 2 määritellään tiedonlouhinta ja käydään läpi sen lähialueita, menetelmiä ja sovelluksia. Luvussa 3 käsitellään tarkemmin rakenteisten dokumenttien käsittely. Luvussa 4 hahmotellaan pro gradu -tutkielman yhteydessä toteutettavan sovelluksen vaatimuksia ja tavoitteita.

2 Mitä tiedonlouhinta on?

Hand et al. [16, sivut 1-4] määrittelevät tiedonlouhinnan (*Data Mining*) olevan suurten, tiettyä tarkoitusta varten kerättyjen tietojoukkojen analyysia, jonka tarkoituksena on löytää odottamattomia suhteita ja tiivistää dataa uusilla tavoilla, jotka ovat sekä ymmärrettäviä että käyttökelpoisia. Tiedonlouhinta on tieteidenvälistä toimintaa, jonka piiriin kuuluu joukko erilaisia menetelmiä ja algoritmeja. Tiedonlouhinnan lähitieteitä ovat tilastotiede, tietokannat, koneoppiminen, hahmontunnistus, tekoäly ja visualisointi.

Tiedonlouhinta voidaan nähdä myös osana tietämyksen muodostamista tietokannoista (*KDD, Knowledge Discovery in Databases*). Tietämyksen muodostaminen on epätriviaali prosessi, jossa pyritään muodostamaan päteviä, uusia, potentiaalisesti käyttökelpoisia ja lopulta ymmärräviä malleja datasta. Prosessi on interaktiivinen ja iteratiivinen. Prosessin keskeisiä työvaiheita ovat tiedonlouhinnan ohella mm. tiedon esikäsittely ja tulkinta. Esikäsittelyssä eli datan puhdistamisessa datasta pyritään poistamaan ”kohinaa” ja täyttämään tyhjät tietoalkiot. Tulkinta tarkoittaa tietonlouhinnasta saatujen tulosten arviointia sekä mahdollista jatkokäsittelyä, kuten visualisointia. Prosessin pääkohdat on esitetty kuvassa 1. [12]



Kuva 1: Tiedonlouhinta osana tietämyksen muodostamisprosessia.

Edellisten (sinänsä lähellä toisiaan olevien) määritelmien ohella termejä *tiedonlouhinta* ja *tietämyksen muodostaminen* käytetään kirjallisuudessa synonyymeinä. Tietämyksen muodostamisprosessi painottaa tietokannoisssa olevan tiedon analysointia, mutta tiedonlouhintaa voidaan tarkastella myös yleisemmin ot-

tamatta kantaa syötetiedon formaattiin. Tiedonlouhintaa voidaan soveltaa kaikeen digitaalisessa muodossa olevaan dataan. [23]

2.1 Louhintatieteiden perhe

Tiedonlouhinnassa käytetty tietojoukko on perinteisesti matriisi, joka sisältää numeeristen tai lueteltujen muuttujien arvoja [16, sivut 4-9]. Yksittäinen relaatiotietokannan taulu tai tietovarasto (*Data Warehouse*) soveltuvat tietojoukoksi. Numeerisen tiedon louhinnan rinnalle on kehittynyt lukuisia uusia monimuotoisen syötetiedon louhintaan keskittyviä tutkimusalueita. Näistä keskeisimpiä ovat tekstitiedon louhinta (*Text Mining*), web-louhinta (*Web Mining*) ja relaatiotiedon louhinta (*Multirelational Data Mining*).

Tekstitiedon louhinta [9] on tiedonlouhintaa (rakenteettomasta tai puolirakenteisesta) tekstidatasta tai dokumenttijoukosta. Web-louhinta [10, sivut 195-206] on tiedonlouhintaa WWW-ympäristössä. Web-louhinta jakautuu edelleen WWW-sivujen sisällön, rakenteen ja käytön analysointiin. Relaatiotiedon louhinta [11] etsii malleja ympäristössä, joka käsittää useita erilaisilla suhteilla toisiinsa liittyviä tietojoukkoja. Tietojoukot voivat olla esim. relaatiotietokannan tauluja tai rakenteisia dokumentteja.

2.2 Menetelmiä ja sovelluksia

Tiedonlouhinnan menetelmillä pyritään muodostamaan malleja syötedatasta löydettyjen hahmojen ja säännönmukaisuuksien pohjalta. Mallit jakautuvat kuvaileviin ja ennustaviin. Tunnetuimmat menetelmät ovat seuraavat: [4]

- **Klusterointi** on kuvaileva menetelmä, jossa tietoalkiot ryhmitellään äärelliseen määrään klustereita (ryppäitä) niiden keskinäisen samanlaisuuden perusteella. ”Samanlaisuus” määritellään etäisyysfunktiolla.
- **Luokittelu** ja **regressio** ovat ennustavia menetelmiä, joista luokittelussa tietoalkiolle pyritään määrittämään jokin ennalta määrätyistä luokista. Regressiossa tietoalkiolle määritetään numeerinen arvo.
- **Assosiaatioiden** sekä **peräkkäisten toimintojen** etsintä. Assosiaatiot ovat kuvaileva malli toistuvasti yhdessä esiintyville tietueille, peräkkäiset toimintosäännöt ovat ennustavia assosiaatioita, joiden kohdalla on tiedossa tietueiden suhteellinen järjestys.

Tietokannoissa ja Internetissä olevan tiedon määrä on viime vuosina kasvanut kiihtyvää vauhtia. Massiivisten tietovarastojen tehokas hyödyntäminen vaatii tiedonlouhintaa. Sovellusalueita ovat esimerkiksi asiakasprofiilien muodostaminen,

hakukoneiden toiminnan tehostaminen, roskapostin suodatus, vakuutuspetosten jäljitys, televerkon virheiden analysointi ja geenitutkimus. [23]

Esimerkkinä tekstitiedon louhinnasta käytännössä Bohnacker et al. [3] esittävät yrityksen ja asiakkaan välisen kommunikoinnin (B2C) tehostamisen. Esimerkkiyrityksessä käytössä olleella palautejärjestelmällä asiakkaat voivat antaa palautetta ja parannusehdotuksia yrityksen tuotteista. Palautteet jaettiin manuaalisesti yli 10000 eri kategoriaan, jotka käsittävät joukon eri tuotteita ja tunnistettuja vikatyyppejä. Palautetietojen arviointi ja etsiminen valtavasta kategoriajoukosta vei työaikaa. Lisäksi järjestelmä sopeutui huonosti odottamatomaan, mutta silti relevanttiin asiakaspalautteeseen.

Tekstitiedon louhinnalla palautejoukko jaettiin automaattisesti klustereihin. Tällöin työntekijä saa nopeasti yleiskäsityksen palautemassasta ja voi arvioida manuaalisen luokituksen onnistumista. Sovelluksen klusterointi perustuu yksinkertaiseen ideaan verrata palautteita niissä esiintyvien sanojen jakauman perusteella. Sovelluksen kehittäjien mukaan kiinnostavia (uusia ongelma-alueita ilmaisevia) klustereita ovat sellaiset, jotka on jaoteltu manuaalisesti eri kategorioihin, mutta silti muistuttavat sanatasolla toisiaan.

3 Tutkimusalue

Pro graduni tutkimusalue on tiedonlouhinta rakenteisista dokumenteista. Dokumenttiformaateista rajoitutaan XML-muotoiseen tietoon. Testiaineistona on kokoelma samaan dokumenttityyppiin kuuluvia, mahdollisesti keskenään linkitettyjä dokumentteja. Tiedonlouhintamenetelmistä keskitytään dokumenttien klusterointiin. Dokumenttien samanlaisuutta mitattaessa pyritään käyttämään hyödyksi niiden rakennetta ja tekstisisältöä sekä dokumenttijoukosta riippuen myös linkkejä ja metatietoa.

Tutkielman kannalta kiinnostavimpia tiedonlouhinnan osa-alueita ovat tekstin ja relaatiotiedon louhinta. Pääosa rakenteisten dokumenttien informaatiosta on ilmaistu luonnollisella kielellä, jonka analysointi vaatii tekstitiedon louhintaa. Dokumenttien rakenteen ja linkkien analysoinnissa tarkoitukseni on soveltaa relaatiotiedon louhintaa.

3.1 Tekstitiedon louhinnasta

Tekstitiedon louhinta sivuaa läheisesti ajallisesti varhaisempia luonnollisen kielen käsittelyyn (*Natural Language Processing, NLP*) liittyviä tutkimusalueita, joita ovat tiedonhaku (*Information Retrieval, IR*) ja tiedon eristäminen (*Information Extraction, IE*). Tiedonhaku on tietojenkäsittelytieteen osa-alue, joka tutkii tiedon (ei datan) hakua dokumenttikokoelmasta. Dokumenttien haun tarkoitus on

tyydyttää käyttäjän (yleensä luonnollisella kielellä) ilmaisema tiedon tarve [1, sivu 444]. Tiedon eristäminen on prosessi, jonka syötteenä on tiedonhaussa saatu dokumenttikokoelma ja tuloksena analysoitua ja tiivistettyä tietoa dokumenteista yhtenäisessä kehyksessä [7].

Tiedonhaku ja tiedon eristäminen suhtautuvat dokumentteihin toisistaan poikkeavilla tavoilla. Tiedonhaun syötteenä on käyttäjän kysely ja tuloksena *relevantteja* dokumentteja. Tiedon eristämisessä syötteenä on dokumenttijoukko ja tuloksena määrämuotoinen kooste dokumenteissa olevasta tiedosta [18]. Tekstitiedon louhinta muistuttaa molempia, mutta on niitä kehittyneempi tekniikka. Tiedonhaussa käyttäjän täytyy tietää vähintään avainsana etsimästään tiedosta ja tiedon eristämisessä käyttäjän on määriteltävä tarkasti, millä säännöillä tietoa erotellaan dokumenteista [17]. Tekstitiedon louhinnan tavoitteena on löytää uutta ja mahdollisesti odottamatonta tietoa automaattisesti. Louhintatekniikoita voidaan hyödyntää sekä tiedonhaussa että -keruussa. Esimerkiksi hakutuloksia voidaan ryhmitellä klusteroinnin avulla.

Kodratoff [18] kutsuu tekstidokumenttien hakuun ja käsittelyyn liittyvää prosessia nimellä tietämyksen muodostaminen tekstiaineistosta (*Knowledge Discovery in Texts*). Prosessin keskeisin ominaisuus on sen tietämystä (*Knowledge*) luova vaikutus. Termejä *tekstitiedon louhinta*, *dokumenttien louhinta* ja *tiedonhaku teksteistä* on käytetty myös synonyymeina [22] [8]. Kaupallisiin sovelluksiin markkinointitarkoituksessa kehitetyt uudet termit ilman todellista uutta sisältöä vaikeuttavat käsitteiden täsmällistä määrittelyä. Yleensä tekstitiedon louhinta ymmärretään kuitenkin tiedonlouhinnan ja luonnollisen kielen käsittelyn tekniikoiden soveltamiseksi tekstimuotoiseen tietoon [9].

3.2 Rakenteisten dokumenttien käsittelyn erikoispiirteet

Tekstisisällön analysoinnin kannalta rakenteiset dokumentit eivät eroa muista dokumenteista, koska molempia analysoidaan käyttäen (rakenteettoman) tekstin louhintaa. Muilta osin rakenteisten dokumenttien käsittely on kuitenkin helpompaa kuin rakenteettoman (esim. tekstitiedostot) tai puolirakenteisen (esim. ulkoasuun keskittyvä HTML). Rakenteisuus mahdollistaa tietokoneen luettavissa ja tulkittavissa olevan tiedon sisällyttämisen dokumenttiin.

Huolellisesti valituilla elementeillä dokumentti voi kuvata sisältönsä merkityksiä rakenteellaan (”Itsekuvailevat dokumentit”). Tällöin dokumenttien rakennetta voi käyttää hyväksi esim. verrattaessa eri sivujen rakenteellista samanlaisuutta [15] tai poistamalla dokumentista osat, jotka eivät ole merkityksellisiä tutkittavan ongelman kannalta. Rakenne mahdollistaa myös täsmälliset viittaukset dokumentin eri osiin esim. XPath-kielellä [6].

Linkit ovat toinen keskeinen tekijä WWW-ympäristössä ja XML-kielisissä dokumenteissa yleensäkin. Linkkejä analysoimalla voidaan arvioida tietyn doku-

mentin luotettavuutta tai linkitettyjen dokumenttien samanlaisuutta [20]. Linkkien analysointi voidaan tulkita relaatiotiedon louhinnaksi: yhtä DTD:tä (*Document Type Definition*) noudattavat dokumentit vastaavat relaatiotietokannan taulua, elementit vastaavat kenttiä ja linkit suhteita muihin tietueisiin [13].

Metatiedon avulla dokumenteista voidaan ilmaista lisätietoa yhtenäisellä kielellä; joidenkin sovellusalueiden sisällä myös yhtenäisellä käsitteistöllä eli ontologialla. W3C:n määrittelemä RDF (*Resource Description Framework*) [19] ja ISO:n standardoima[1] aihekartat (*Topic Maps*) [21] ovat kieliä metatiedon kuvaamiseen. Metatietoa voidaan käyttää hyväksi tiedonhaussa tai tehtäessä päätelmiä dokumenttien ominaisuuksista edellyttäen, että sovellusalue on hyvin tunnettu.

Kuva 2: CiteSeer-kirjaston käyttöliittymä.

[1] ISO/IEC 13250

Jos dokumenteilla ei ole metatietoja, niitä voidaan myös luoda louhinnan tuloksena. CiteSeer[2] on WWW:ssä toimiva digitaalinen kirjasto ja indeksointijärjestelmä, joka etsii automaattisesti tieteellisiä artikkeleita, indeksoi ne ja tutkii niiden keskinäiset viittaukset. CiteSeeria voidaan pitää tekstinlouhintajärjestelmänä, koska se pystyy analysoimaan eri formaateissa olevia tekstejä ja tunnistamaan eri tavoin merkityt viittaukset samaan dokumenttiin. CiteSeer ei erityisesti käsittele rakenteisia dokumentteja, mutta järjestelmän linkkien käsittelylogiikka, dokumenttien samankaltaisuuden vertailu ja metatietojen luonti soveltuisivat hyvin myös niihin. Kuvassa 2 on esimerkkki järjestelmän käyttöliittymästä. [14]

3.3 Klusterointimenetelmistä

Klusterointi on tietoalkioiden ryhmittelyä lähellä toisiaan olevista alkioista koostuviin klustereihin (ryppäisiin). Tiedonlouhinnassa klustereiden ajatellaan kuvaavan tietojoukossa piilossa olevia hahmoja, yleisemmin kyse on tiheysjakauman estimoinnista. Datan esittäminen pienemmällä määrällä klustereita helpottaa tiedon tulkintaa, mutta samalla mallista häviää tietoa. Jokainen tietoalkio kuuluu yleensä korkeintaan yhteen klusteriin, mahdolliset poikkeamat (*outliers*) poistetaan aineistosta.

Kaikkiin klusterointimenetelmiin liittyy tavalla tai toisella tietoalkioiden vertaamiseen käytetty mitta. Yleisiä mittoja ovat esimerkiksi L_p-etäisyydet

$$d(x,y) = \| x - y \|_p = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p},$$

missä x ja y ovat verrattavien alkioiden piirrevektoreita. Erikoistapauksena L_2-etäisyys on tuttu euklidinen etäisyys. Muista mitoista mainittakoon esim. piirrevektorien välinen kulma samanlaisuuden määrityksessä. Kaikkea dataa ei voida muuntaa numeeriseen muotoon (laadulliset muuttujat), jolloin mitan käsite täytyy määritellä jollakin toisella tavalla. Myös rakenteisten dokumenttien välisen ”etäisyyden” mittaaminen osoittautunee hankalaksi tehtäväksi. Tärkeimmät klusterointimenetelmät voidaan ryhmitellä seuraavasti:

- **Hierarkkiset** menetelmät jakautuvat kokoaviin ja jakaviin menetelmiin. Datasta muodostetaan puurakenne, jonka solmut edustavat klustereita tietyllä tarkkuustasolla. Kokoavissa menetelmissä puun muodostaminen aloitetaan yksittäisistä havaintoalkioista, jakavissa menetelmissä koko havaintoaineisto tulkitaan alussa yhdeksi klusteriksi, jota jaetaan. Hierarkkisten menetelmien etuna on mahdollisuus tarkastella havaintoaineistoa monella tarkkuustasolla.

[2] http://citeseer.nj.nec.com/cs

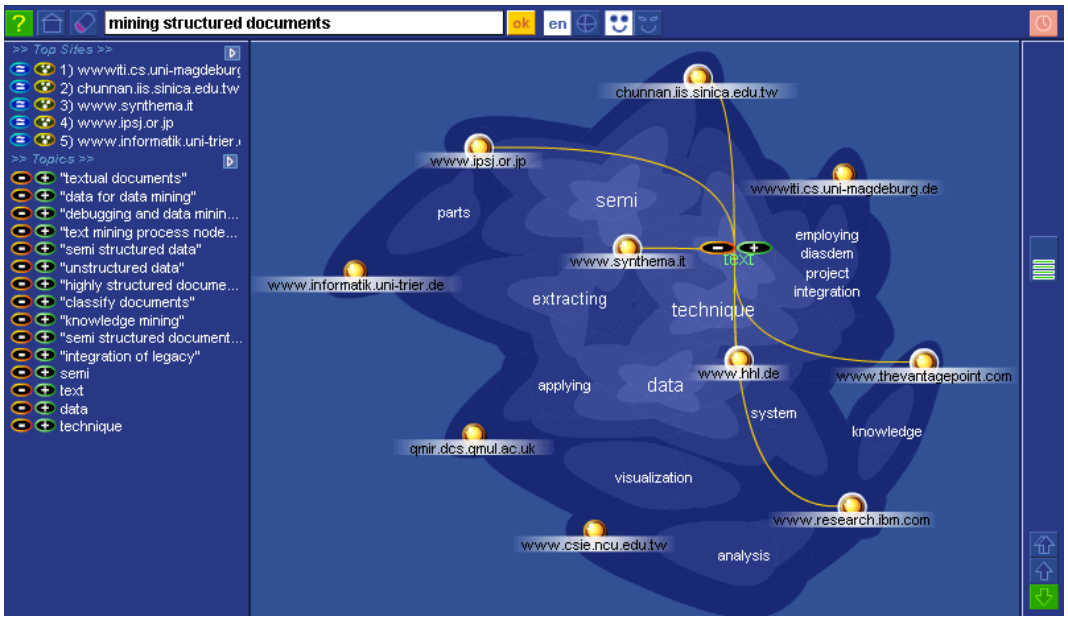
- **Osittavat** menetelmät perustuvat havaintoalkioiden iteratiiviseen ryhmittelyyn. Hierarkkisista menetelmistä poiketen klusterien paikat muuttuvat algoritmin edetessä, mutta klusterien määrä on ylensä sidottu. Tunnetussa K-means -algoritmissa määritetään etukäteen klusterien määrä, joista kukin esitetään pisteidensä painotettuna keskiarvona. Toinen lähestymistapa havaintopisteiden ositukseen ovat tiheyteen perustuvat menetelmät, joilla lähellä toisiaan olevat pisteet luokitellaan samaan klusteriin.
- **Ruudukkoon** (*grid*) perustuvissa menetelmissä on päinvastainen lähestymistapa hierarkkisiin ja osittaviin menetelmiin verrattuna. Kun edellisissä menetelmissä keskityttin klustereiden muodostamiseen havaintopisteiden perusteella, ruudukkoon perustuvissa menetelmissä lähdetään havaintoavaruudesta ja sen osittamisesta. Menetelmät eivät tällöin ole riippuvaisia havaintoalkioiden valinnan järjestyksestä.
- **Keskinäiseen esiintymiseen** (*co-occurrence*) perustuvat menetelmät soveltuvat laadullisten muuttujien käsittelyyn. Havaintoalkio on äärellinen bittivektori, jonka arvot ilmaisevat jonkin laadullisen ominaisuuden mukana oloa tai puuttumista (esim. asiakkaan ostoskori). Havaintoalkioiden läheisyyttä voidaan mitata esim. yhteisten lähimpien naapureiden (*Shared Nearest Neighbors, SNN*) määrällä.

Klusterointiin on sovellettu myös laskennallisesti älykkäitä tekniikoita, kuten geneettisiä algoritmeja ja itseorganisoitavia karttoja (SOM). Laajojen tietokantojen, paikkatiedon ja korkeadimensioisen datan käsittelyyn on kehitetty erityismenetelmiä. Kaikenkaikkiaan erilaisten klusterointimenetelmien ja algoritmien kirjo on hyvin laaja. [2]

4 Kehitettävä sovellus

Pro graduni empiirisenä osuutena on tarkoitus kehittää XML-dokumenttien klusterointisovellus. Sovellus ryhmittelee tietyn DTD:n mukaisia XML-dokumentteja sisällöltään läheisiin klustereihin. Dokumenttikokoelma klustereineen on mahdollista visualisoida käyttäjälle esim. itseorganisoituvan kartan avulla (ks. esim. WebSOM[3] tai KartOO[4]). Järjestelmällä pitäisi pystyä myös tekemään yksinkertaisia hakuja vertaamalla dokumentteja käyttäjän antamaan dokumenttifragmenttiin [5]. Dokumentit voisivat sijaita XML-tietokannassa tai WWW:ssä. Klusterointialgoritmia tai dokumenttien etäisyysmittaa ei ole vielä päätetty menetelmien kirjavuuden takia.

[3] http://websom.hut.fi/websom/
[4] http://www.kartoo.com/



Kuva 3: KartOO-hakukoneen käyttöliittymä.

Sovelluksen tulee olla konfiguroitavissa eri dokumenttityyppien mukaisesti. Konfiguroinniin voisi toteuttaa määrittelemällä joukon eri elementeille tyypillisiä prosessointitapoja, jotka asetetaan dokumenttityyppikohtaisesti. Taulukossa 1 on lueteltu mahdollisia elementtityyppejä ja niihin kuuluvia elementtejä XHTML-tyylisessä dokumenttityypissä.

Elementtityyppi	HTML-elementtejä
Otsikko	`h1`,`h2`
Sisältö	`p`,`span`,`b`,`i`,`img`-elementin `alt`-attribuutti
Linkki	`a`, `href`-attribuutit
Rakenteen kuvaus	`body`,`div`,`table`,`ul`,`li`
Ei-sisältö	`script`, kommentit
Metatieto	RDF-kuvaukset

Taulukko 1: XML-elementtityyppejä

Elementtien ryhmittely ohjaa klusterointialgoritmin toimintaa. Taulukon esimerkkityypeistä *otsikko-* ja *sisältö* -tyyliset elementit käsitellään suoraan niiden tekstisisällön perusteella, linkeissä hyödynnetään dokumenttien välisiä viittauksia ja rakenteellisille linkeille voidaan määritellä oma käsittelynsä. *ei-sisältö* -tyyliset elementit voidaan jättää kokonaan huomiota.

5 Yhteenveto

Tiedonlouhinta on suurten tietojoukkojen analysointia ja mallien muodostamista. Tiedonlouhinnan voi yleisimmillään tulkita minkä tahansa digitaalisessa muodossa olevan tiedon käsittelyksi tai osaksi tietämyksen muodostamisprosessia, jossa pyritään muodostaan raakadatasta uusia, käyttökelpoisia ja ymmärrettäviä malleja (tietämystä). Numerotiedon louhinnan rinnalle on kehittynyt myös muuntyyppisen tiedon louhintaan keskittyviä tutkimusalueita, kuten tekstitiedon louhinta, web-louhinta ja relaatiotiedon louhinta.

Tekstisisällön analysoinnin osalta rakenteisia dokumentteja käsitellään tekstitiedon louhinnnan keinoin. Rakenteisuus mahdollistaa lisäksi tietokoneen tulkittavissa olevan tiedon sisällyttämisen dokumenttiin. Tällöin louhinta-algoritmi voi tutkia dokumenttien rakenteellisia eroja ja analysoida linkki- tai metatietoa.

Suunnitteilla oleva XML-dokumenttien klusterointisovellus ryhmittelee tietyn DTD:n mukaisia dokumentteja sisällöltään läheisiin klustereihin. Klusterointialgoritmia tai dokumenttien etäisyysmittaa ei ole vielä menetelmien kirjavuuden takia päätetty, mutta sovelluksen tulee olla konfiguroitavissa eri dokumenttityypeille ja tulosten tulee olla visualisoitavissa käyttäjälle.

Lähteet

[1] R. Baeza-Yates ja B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[2] P. Berkhin. Survey of clustering data mining techniques. Tekninen raportti, Accrue Software, San Josa, CA, 2002. URL: http://citeseer.nj.nec.com/berkhin02survey.html.

[3] U. Bohnacker, L. Dehning, J. Franke ja I. Renz. Textual analysis of customer statements for quality control and help desk support. Kirjassa K. Jajuga, A. Sokolowski ja H.-H. Bock, toim., *Classification, Clustering, and Data Analysis. Recent Advances and Applications. Proceedings of the 8th Conference of the International Federation of Classification Societies (IFCS-2002), Krakow, Poland, July 16 - 19, 2002*, osa 21, ss. 437–445. Springer, 2002.

[4] C. Bounsaythip ja E. Rinta-Runsala. Overview of data mining for customer behavior modeling. Tekninen raportti TTE1-2201-18, VTT Information Technology, 2001. URL: http://citeseer.nj.nec.com/bounsaythip01overview.html.

[5] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass ja A. Soffer. Searching xml documents via xml fragments. Kirjassa *Procee-*

dings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ss. 151–158. ACM Press, 2003. URL: http://http://portal.acm.org/citation.cfm?id=860464&jmp=abstract&dl=portal&dl=ACM.

[6] J. Clark ja S. DeRose. Xml path language (xpath) version 1.0, w3c recommendation. Tekninen raportti, W3C, 1999. URL: http://www.w3.org/TR/xpath.

[7] J. Cowie ja W. Lehnert. Information extraction. *Commun. ACM*, 39(1):80–91, 1996. URL: http://portal.acm.org/citation.cfm?id=234209&coll=portal&dl=ACM&CFID=14016653&CFTOKEN=92368801.

[8] M. Dixon. An overview of document mining technology. URL: http://citeseer.nj.nec.com/dixon97overview.html.

[9] J. Dörre, P. Gerstl ja R. Seiffert. Text mining: finding nuggets in mountains of textual data. Kirjassa *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ss. 398–401. ACM Press, 1999. URL: http://portal.acm.org/citation.cfm?id=312129.312299&dl=portal&dl=ACM&type=series&idx=SERIES939&part=Proceedings&WantType=Proceedings&title=Conference%20on%20Knowledge%20Discovery%20in%20Data.

[10] M. H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2002.

[11] S. Džeroski. Multi-relational data mining: An introduction. *ACM SIGKDD Newsletter*, 5(1):1–16, 2003. URL: http://www.acm.org/sigs/sigkdd/explorations/issue5-1/Dzeroski.pdf.

[12] U. Fayyad, G. Piatetsky-Shapiro ja P. Smyth. From data mining to knowledge discovery in databases. *Ai Magazine*, 17:37–54, 1996. URL: http://citeseer.nj.nec.com/fayyad96from.html.

[13] L. Getoor. Link mining: A new data mining challenge. *ACM SIGKDD Newsletter*, 5(1):84–89, 2003. URL: http://www.acm.org/sigs/sigkdd/explorations/issue5-1/Getoor.pdf.

[14] C. L. Giles, K. Bollacker ja S. Lawrence. CiteSeer: An automatic citation indexing system. Kirjassa I. Witten, R. Akscyn ja F. M. Shipman III, toim., *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, ss. 89–98, Pittsburgh, PA, June 23–26 1998. ACM Press. URL: http://citeseer.nj.nec.com/giles98citeseer.html.

[15] J. Han ja K. C.-C. Chang. Data mining for web intelligence. *IEEE Computer*, 35(11):64–70, 2002. URL: http://www-faculty.cs.uiuc.edu/~kcchang/Papers/dmweb-ieeecomputer02.pdf.

[16] D. Hand, H. Mannila ja P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[17] M. A. Hearst. Untangling text data mining. Kirjassa *37th Annual Meeting of the Association for Computational Linguistics*, ss. 3–10. Association for Computational Linguistics, 1999. URL: http://acl.ldc.upenn.edu/P/P99/.

[18] Y. Kodratoff. Knowledge discovery in texts: A definition and applications. Kirjassa Z. W. Ras ja A. Skowron, toim., *Foundations of Intelligent Systems, 11th International Symposium, ISMIS '99, Warsaw, Poland, June 8-11, 1999, Proceedings*, sarjan *Lecture Notes in Computer Science* osa 1609, ss. 16–29. Springer, 1999. URL: http://citeseer.nj.nec.com/kodratoff99knowledge.html.

[19] O. Lassila ja R. R. Swick. Resource description framework (rdf) model and syntax specification, w3c recommendation. Tekninen raportti, W3C, 1999. URL: http://www.w3.org/TR/REC-rdf-syntax/.

[20] D. S. Modha ja W. S. Spangler. Clustering hypertext with applications to web searching. Kirjassa *Proceedings of the eleventh ACM on Hypertext and hypermedia*, ss. 143–152. ACM Press, 2000. URL: http://portal.acm.org/citation.cfm?id=336351&coll=ACM&dl=ACM&CFID=14098497&CFTOKEN=25660503.

[21] S. Pepper ja G. Moore. Xml topic maps (xtm) 1.0, topicmaps.org specification. Tekninen raportti, TopicMaps.Org, 2001. URL: http://www.topicmaps.org/xtm/1.0/.

[22] A. Tan. Text mining: The state of the art and the challenges. Kirjassa *Proceedings of the Pacic Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases (KDAD'99)*, ss. 65–70, 1999. URL: http://citeseer.nj.nec.com/tan99text.html.

[23] S. Äyrämö ja T. Kärkkäinen. Data mining — principles and basic applications. Tekninen raportti, Jyväskylän yliopisto, Agora Center, 2003.