WWW sivujen tietosisällön louhiminen

Matti Vuorinen, Kati Blomqvist, Veli-Pekka Ahonen, Antti Tani, Sakari Jokinen

8. lokakuuta 2007

1 Johdanto

Syyskuussa 2007, Netcraftin verkkotutkimus löysi Internetistä 135 166 473 eri osoitetta, kasvua edelliseen, kaksi kuukautta vanhan tutkimuksen osoitemäärään, oli noin 5,5% [Net07]. Osoitteista ainoastaan osa, noin puolet on aktiivisia, mutta silti niitä on valtavan paljon. Internetiä pidetään hyvästä syystä maailman kattavimpana tietovarastona; Internetistä löytyy tietoa lähes kaikesta maailman ja taivaan väliltä ja vähän kauempaakin. Tietoa on kuitenkin niin paljon, että normaalin käyttäjän on jo hankalaa löytää ja suodattaa etsimäänsä tietoa. Käyttäjien avuksi onkin hakukoneita kuten Google tai Altavista, joiden avulla Internetistä voi etsiä sivuja ja dokumentteja.

Ongelmana perinteisille hakukoneille on, että ne tarjoavat ainoastaan listan linkkejä dokumentteihin, joista käyttäjän pitää itse valita mitä linkkiä seurata, sekä etsiä dokumenteista haluamansa tiedot. Suosituilla hakusanoilla hakukone voi tarjota miljoonia linkkejä, joista käyttäjän pitäisi itse etsiä haluamansa tiedot. Esimerkiksi sanoilla "George Bush", Google tarjoaa yli 14 miljoonaa sivua. Tietoa on siis saatavilla normaalille käyttäjälle liikaa ja liian sekavassa muodossa. Sivustoja on liikaa myös manuaalista ja keskitettyä sivustojen luokittelua varten.

Vaikka tietoa on Internetissä valtavasti ja sivustoilla oleva tieto on suunnattu ihmisten luettavaksi, tiedon automaattista käsittelyä helpottaa esimerkiksi se, että osa tiedosta on rakenteellista. Suuri osa Webin tiedoista on talletettu erilaisiin taulukoihin, sitä on merkitty normaaleilla HTML-tageilla, sekä monelle sivulle on lisätty paljon metatietoa [ChC06].

Netissä olevan tiedon automaattista käsittelyä vaikeuttaa kuitenkin moni asia. Kuten aiemmin todettiin, suurin osa tiedosta on tarkoitettu ihmisten luettavaksi. Tieto on erittäin heterogeenista, kirjoitettu eri kielillä, osa puhekielellä, käytetty lyhenteitä yms. Vaikka tiedon esittämiseen Internetissä on tarjolla paljon standardeja, kaikki tieto ei suinkaan ole tarjolla yhdessä formaatissa. Suuri osa sivustoista on tarjolla HTML-muodossa, sen lisäksi käytössä on pelkkää tekstiä, osa tiedoista on kuvia, osa pdf-dokumentteja, osa videoita ja niin edelleen ja edelleen. Ihminen osaa käsitellä suurinta osaa näistä dokumenteista, mutta automaattinen, ohjelmallinen tiedonkäsittely on vaikeaa. Vaikka perinteisesti kaikki kirjoihin painettu teksti on ollut tutkittua ja ihmiset ovat pitäneet tietoa totena, kuka tahansa voi kirjoittaa näkemyksiään tai mielipiteitään Internetiin, eikä tekstillä ole välttämättä mitään tekemistä totuuden kanssa. Oman haasteensa verkkotiedon hyväksikäyttämiseen tuo se, että tietoa ei välttämättä löydy enää sieltä, missä se oli vielä edellisellä viikolla, eikä samansisältöisenä kuin sitä on viimeksi. Iso osa Webissä olevasta datasta on staattisina HTML-sivustoina, mutta melkein yhtä suuri määrä tietoa löytyy dynaamisten hakulomakkeiden ja sivustojen avulla, näiden läpikäyminen on vielä haastavampaa kuin staattisten sivujen [LiC04].

Tässä artikkelissa keskitytään verkon sisällön, kuten Internetistä löytyvien dokumenttien, automaattiseen etsimiseen sekä koneelliseen tutkimiseen, tulkitsemiseen ja luokitteluun. Tätä osa-aluetta tiedonlouhinnasta kutsutaan *verkon sisällön louhinnaksi* (engl. web content mining). Kappale 2 antaa yleisluontoisen kuvan tästä verkon sisällön louhinnasta. Seuraavat viisi kappaletta käsittelevät tarkemmin eräitä näistä sisällön louhinnan osa-alueista: tiedon yhdistämistä eri lähteistä (Kappale 3), tiedon kokoamista ja esittämistä hierarkkisesti jäsenneltynä (kappale 4), tiedon eristämistä WWW-sivuilta (kappale

5), mieilipiteiden louhintaa esimerkiksi tuotearvioista (kappale 6), sekä WWW-sivujen semanttisten osien tunnistamista ja osittelua (kappale 7).

2 Verkon sisällön louhinnan eri suuntauksia

Verkon sisällön louhinta liittyy vahvasti perinteiseen tiedonlouhintaan. Tiedonlouhinta käsittää tiedon eristämisen tietokannasta tai tietovarastosta, tiedon käsittelyn ja tiedon esittämisen sopivammalla tavalla. Verkon voidaan ajatella olevan maailman suurin tietokanta.

Tiedonlouhinta verkossa jakautuu kahteen osaan: *verkon käytön louhintaan* (engl. web usage mining) sekä verkon sisällön louhintaan. Verkosta voidaan louhia myös multimedia-aineistoa, mutta nyt olemme keskittyneet tekstitiedon louhintaan. Perinteiset tiedonlouhinnan menetelmät, luokittelu, assosiaatiosäännöt, klusterointi, ennustaminen ja *luonnollisen kielen käsittely* (engl. natural language processing) toistuvat, mutta verkko synnyttää lisähaasteita tekniikoiden soveltamiseen.

Verkon luonne on dynaaminen, muuttuva, ja dokumentteja on lukematon määrä. Näin ollen tarvitaan uusia ratkaisuja. Myös kyselytulosten esittäminen on tärkeä seikka. Koska hakuala on valtava, verkkokysely voi palauttaa tuloksena tuhansia verkkosivuja. Tarvitaan mielekkäitä menetelmiä esittää näin valtavia tuloksia, jotta käyttäjän olisi helpompi valita tuloksista häntä kiinnostavat sisällöt.

2.1 Tekniikoita

2.1.1 Verkkosivujen luokittelu

Verkkosivun sisältö täytyy ensin analysoida, jotta voitaisiin hakea relevanttia tietoa. Sivujen luokittelu tarjoaa automaattisen tavan päättää kohteen oleellisuus. Luokittelu kohdistuu verkkosivuihin tai palvelimiin. Koska verkkosivut usein esittävät tietoa yleisemmällä tasolla (esim. koko yrityksen tiedot) ja tieto on usein esitetty usealla sivulla, tarvitaan uusia algoritmeja, jotka mahdollistavat sivuston luokittelun etusivun perusteella.

2.1.2 Keskitetty ryömintä

Keskitetty verkkoryömijä (engl. focused web crawler) aloittaa joukosta tarkoin valikoituja sivuja, jotka kuvaavat käyttäjän mielenkiinnon kohdetta. Ryömijä aloittaa annetuista sivuista ja rekursiivisesti tutkii linkitettyjä sivuja. Keskitetty ryömijä käyttää parasensin hakua, jota ohjaa käyttäjän mielenkiinto.

2.1.3 Verkkokohteiden klusterointi

Keskitetty ryömintä hakee suuren määrän oleellista dataa. Jos halutaan tarjota nopeampi ja pääsy kyselyn tuloksiin, voidaan käyttää klusterointia. *Klusterointi* kokoaa haetun informaation ryhmiksi, jotta ymmärrys tuloksista paranee. Koska verkkomateriaali on tekstin lisäksi verkkosivustoja ja kuvia, tarvitaan klusterointiin uusia algoritmeja.

2.1.4 Tiedon integrointi

Monet verkkosivut ilmaisevat samoja tai toisiinsa liittyviä asioita eri tavalla. Kun tietoa halutaan koostaa useilta eri sivuilta, tarvitaan semanttista integraatiota useista lähteistä. Tätä aihetta käsitellään laajemmin kappaleessa 3.

2.1.5 Käsitehierarkioiden rakentaminen

Koko verkkoa ei voida millään järjestää, mutta verkkoon kohdistetun kyselyn hakutulos voidaan. Hierarkista järjestämistä voidaan tehdä esimerkiksi klusteroimalla tekstiä, mikä yhdistää samankaltaiset vastaukset hierarkisiksi ryhmiksi. Toinen tapa on etsiä tärkeitä käsitteitä ja alikäsitteitä ja näiden hierarkisia suhteita. Kappale 4 käsittelee tarkemmin tätä aihetta.

2.1.6 Tiedon eristäminen

Rakenteisesti esitetty tieto on usein tärkeää. Esimerkiksi verkkokauppojen tuotteet esitellään usein jollain tavalla säännöllisessä muodossa. Rakenteinen data on suhteellisen helppo eristää. Toisaalta monet verkkosivut ovan vain tekstidokumentteja, joilla ei ole säädettyä rakennetta. Tälläisen rakenteettoman tiedon louhinta liittyy läheisesti teks-

tinlouhintaan, tiedonhakuun ja luonnollisen kielen käsittelyyn. Tämänhetkiset tiedoneristämistekniikat perustuvat pitkälti koneoppimiseen ja luonnollisen kielen käsittelyyn. Tiedon eristämistä käsitellään tarkemmin kappaleessa 5.

2.1.7 Mielipiteiden louhinta

Verkko on loistava lähde kuluttajien mielipiteen tutkimiseen. Ennen tietoa kerättiin erilaisilla gallupeilla, nyt sama tieto on vapaasti saatavilla netissä. Mielipiteitä löytyy sekä erikseen mielipiteille tarkoitetuilta saiteilta, että miltä tahansa keskustelupalstalta tai muusta sosiaalisesta sivustosta. Tällä hetkellä kehitetään tekniikoita, joiden avulla tätä mielipidetietoa saadaan tehokkaasti ja helposti kerättyä, koottua sekä esitettyä. Mielipiteiden louhintaa käsitellään laajemmin kappaleessa 6

2.1.8 WWW-sivujen osittelu ja hälyn torjuminen

WWW-sivuilla on usein eri alueita, joilla on eri merkitys: pääsisältöalue, navigaatioalue, mainostila jne. Näiden alueiden erottaminen toisistaan on hyödyllistä, jotta kyselyt ja haut voidaan kohdistaa vain haluttuun alueeseen. Tätä aihetta käsitellään kappaleessa 7.

3 Tiedon integrointi

Kaikki tietävät Webistä löytyvien sivujen lukumäärän ja sivujen sisältämän tiedon määrän kasvaneen viime vuosina erittäin nopeasti; vuonna 1999 web-sivuja arvioitiin olevan noin 800 miljoonaa kappaletta, tuo luku oli kasvanut vuoteen 2005 mennessä 19,2 miljardiin. Sivujen määrä on siis kasvanut kuudessa vuodessa noin 24 kertaiseksi. Näissä luvuissa mukana olevat sivut ovat staattista tietoa sisältäviä HTML-sivuja [ChC06].

Toinen, melkein samalla vauhdilla kasvanut tietomäärä löytyy erilaisten hakukaavakkeiden ja dynaamisten sivustojen takaa. Vuosina 2000 ja 2004 tehtiin tutkimus, jossa etsittiin dynaamisia websivustoja ja niiden sisältämiä erilaisia hakulomakkeita. Vuonna 2000 hakusivustoja löytyi 96 tuhatta kappaletta, joiden avulla löytyi jopa 550 miljardia erilaista sisältösivua, neljän vuoden jälkeen hakulomakkeita löytyi jo 1,2 miljoonaa kappaletta. Dynaamisten sivujen ja erilaisten hakusivustojen määrä kasvoi tutkimusaikana noin 3-7 kertaiseksi. Kun normaalien, staattisten, webin "pintatasolla" olevien sivustojen tiedonlouhintaa on tutkittu paljon ja Googlen, Yahoon ja muiden hakukoneiden hakuroboteille sivustojen läpikäyminen ja tulkitseminen on arkipäiväistä, dynaamisten sivustojen, tiedonhakulomakkeiden takaa löytyvien, webissä "syvemmältä" löytyvän tiedon louhinnassa on vielä paljon tutkittavaa [ChC06].

"Syvästä" webistä löytyvät tiedot on usein rakenteista, mutta esimerkiksi taulukkoon tuotettu tieto, vaikkakin samaa tietoa, ei löydy samoista sarakkeista, samanlaisina merkkijonoina jne. Kuitenkin tehokasta datan hyväksikäyttöä varten tiedot pitää saada rakenteiseen tietovarastoon, josta voidaan tehdä nopeita hakuja tai laskutoimituksia.

Esimerkiksi yritysostoa suunnittelevan yrityksen johdolle olisi erittäin mielenkiintoista hakea kohdeyrityksestä löytyvää tietoa, taloudellisia lukuja, tilinpäätöstietoja yms. Usein tämmöiset tiedot julkaistaan joko yrityksen omalla kotisivulla tai eri talouslehtien artikkeleissa. Kuitenkin jo yrityksen nimi voi olla kirjoitettuna monella tapaa ja järjestelmän pitää ymmärtää, että kyseessä on saman yrityksen tietoja. Artikkeli, joka kertoo edellisen vuoden tapahtumia GENERICS KY:ssä, kertoo samaan yhtiön tietoja, kuin Generics Finland Ky:n kotisivu [May07].

Toinen esimerkki voi olla eri lentoyhtiöiden aikataulutietoja ja hintoja vertaileva yleishakukone. Melkein jokaisella lentoyhtiöllä on kotisivullaan hakulomake, jonka avulla voi selata yhtiön lentoja, aikatauluja, kohteita ja hintoja. Kaikkien lentoyhtiöiden hakulomakkeet hakevat samantyyppistä tietoa, samannäköisessä muodossa. Kun sivuja tarkastellaan lähemmin, huomataan että hakulomakkeet eroavat toisistaan, jotkut enemmän ja jotkut vähemmän. Samoin lomakkeiden tuottamat tiedot eroaa toisistaan, vähintään yhtä paljon kuin itse hakulomakkeet. Hetkeen tuskin löytyy maailmasta auktoriteettia, jolla olisi valtaa tai halua määrittää miten lentoyhtiöiden aikatauluhaut pitää toteuttaa tai miltä tulosten pitää näyttää, vaan yleisen hakukoneen toteuttajan pitää

mukautua eri järjestelmiin [Liu05].

Tietojen integroinnissa on suuria haasteita. Haettavat tietovarastot ovat kooltaan suuria, Internetin tapauksessa suorastaan valtavia ja ne on täysin hajautettuja, ympäri maailmaa. Tästä johtuen kaikkea etsittyyn asiaan liittyvää tietoa ei kannata, eikä edes voi hakea, vaan on tarpeen kehittää algoritmeja jotka hakevat tarpeellista tietoa halutuista tietolähteistä. Tietolähteet eivät ole yksittäisen suuren tahon omistamia, vaan jokaisella tietolähteellä on oma omistajansa, jolla on tiedon, tiedon esittämisen ja sen jakamisen suhteen omat intressinsä. Tietolähteet ovat, kuten jo aiemmin todettiin, täysin heterogeenisiä. Tieto voi olla tallennettu tietokantaan, tavalliseen tekstitiedostoon, tieto voi olla xml dokumenttina ja niin edelleen. Lisäksi eri sivustoilla käytetään eri ontologioita, puhutaan eri asioista samoilla sanoilla ja samoista asioista eri sanoilla. Järjestelmän pitää siis osata erottaa tekstin kontekstista, mitä tekstissä käsitellään ja mitä tietoa tekstistä voidaan käyttää hyödyksi.

Tietoja integroivan järjestelmän pitää pystyä kommunikoimaan haluttujen tietovarastojen kanssa, määrittämään kyselyn jolla saadaan tietoa haettua, linkittämään käyttäjän ontologian tietovaraston ontologiaa vastaavaksi, muuttamaan alkuperäisen kyselyn, eli hakukoneeseen syötetyn haun, tarkaksi suunnitelmaksi siitä mitä tietoa hakea ja mistä sitä pitää etsiä, sekä tuottamaan oikea vastaus muodossa, jonka järjestelmän käyttäjä ymmärtää. [Cas03]

4 Tiedon kokoaminen ja esittäminen

WWW-hakukoneet suoriutuvat hyvin pistemäisestä tiedonhausta, kuten URL:n etsimisestä, mutta WWW-haulla on vaikeampaa muodostaa yleiskuvaa annetusta aiheesta. Jos WWW-tiedonhaussa ei tyydytä pelkkiin tiedonmurusiin sieltä täältä, vaan tavoitteena on saada kattava ja koherentti kuva etsittävästä aiheesta, päästään mielenkiintoisten tiedonhaullisten menetelmien pariin. Yhteistä näille menetelmille on, että ne käyttävät hyödykseen WWW-tiedon redundanssia. Koska eri WWW-sivuita löytyy samaa tietoa, voidaan analysointiin käyttää tiedon esiintymien lukumäärää ja kontekseja. Analysoinnille on lisäksi tunnusomaista säännöllisten lausekkeiden tapaisten syntaktisten hahmojen etsiminen, ja WWW-sivujen rakenteen käyttäminen apuna.

Yleisimmin käytössä olevat hakukoneet palauttavat kyselyn tuloksena järjestetyn listan linkkejä WWW-sivuihin. Menettely on varsin tehoton, koska tuloksen koko voi olla kymmeniä tuhansia sivuja, mutta käyttäjä ei useinkaan tarkastele kuin välittömästi näkyviä muutamaa kymmentä kärkipään linkkiä. Lisäksi tulosjoukon järjestämiseen käytetty kriteeri, kuten niihin linkkaavien sivujen lukumäärä, ei useinkaan täytä käyttäjän tarpeita. Jos käyttäjä haluaa vaikkapa saada yleiskuvaa hakusanan mukaisesta aiheesta, jotkin hakusanaan liittyvät käsitteet voivat olla ilman niihin liittyviä WWW-sivuja ensimmäisten tulosten joukossa, tai johonkin alikategoriaan kuuluvat sivut voivat olla yliedustettuina tulosten kärkipäässä.

Usein tavoiteltava hakurakenne aihepiirin selailuun on *taksonomia*, eli hierarkinen rakenne, joka kuvaa ylikategoria-alikategoria suhteita johonkin aiheeseen liittyen). Taksonomian rakentamisen tarkoituksena on tarjota etsintämekanismi navigointiin ja selaukseen organisoimalla tietoa hierarkisiin klustereihin [Zen04], [Kri04]. Taksonomia tarjoaa yleiskatsauksen aihepiirin käsitteisiin, jakaen samalla aihepiirin WWW-sivuja näiden käsitteiden mukaisiin luokkiin. Tarkoituksena on vähentää käyttäjälle mielenkiintoisten dokumenttien kokonaishakuaikaa. Taksonomian generointialgoritmit voidaan karkeasti jakaa kahteen ryhmään sen mukaan, jakavatko ne dokumentit klustereihin yhden vai usean piirteen mukaan [Kri04]. Yhden piirteen mukaan jakavat (engl. monothetic) algoritmit soveltuvat hyvin luomaan hierarkioita hakutulosten selausta ja yhteenvetoa varten, koska niiden antama jaottelu on helposti ymmärrettävää. Kategorioiden mukaan tuloksia jakavia hakukoneita on yleisessä käytössä, esimerkki tällaisesta on Clusty[1].

Taksonomiasta voitaisiin mennä vielä paljon pidemmälle. Olisi suureksi hyödyksi, jos halutusta aiheesta voitaisiin muodostaa automaattisesti WWW-

[1] http://clusty.com/

tiedon louhinnalla pienen kirjan kaltainen esitys. Tällainen esitys on hyödyksi erityisesti aiheeseen tutustuttaessa ja oppimistarkoituksessa, sekä sellaisten uusien aiheiden kanssa, joista ei ole vielä kirjaa tai yleiskatsausta saatavilla. Tämän tavoitteen saavuttamiseen on esitetty jo joitakin menetelmiä [LCN03].

WWW-sivujen sisältö on jo valmiiksi osittain organisoitua HTML-rakenteen avulla. Olemassaolevien sivujen rakenteita voidaan käyttää hyödyksi muodostettaessa aihetta kuvaavaa semanttista rakennetta. HTML-sivun otsikot ja tekstin korostusta kuvaavat tagit tuovat esille aihepiiriin liittyviä tärkeitä käsitteitä. HTML-tagit voivat kuitenkin joskus olla huonosti perusteltuja ja HTML-merkkaus sisältää tyypillisesti runsaasti kohinaa. Tuloksia saadaan luotettavammiksi jos edellämainitun menettelyn rinnalla tai sijasta tutkitaan sanojen esiintymismääriä sivuilla. Jos jokin sana tai lause esiintyy monilla WWW-sivuilla ja toistuu sivuilla usein, on se kohtuullisen todennäköisesti tärkeä käsite tai aihetta kuvaava ilmaus. Kattavien joukkojen (engl. frequent itemset) etsinnässä käytetyt menetelmät soveltuvat tähän tehtävään joiltakin osin. Myös monimutkaisempia kielellisiä hahmoja voidaan käyttää avuksi rakennettaessa hierarkista aihejaottelua. Eräät kielihahmot ilmaisevat hierarkisia suhteita, käsitteitä tai aliaiheita. Tällaisia hahmoja ovat esimerkiksi “kuten”, “esimerkiksi” tai “sisältää”. “Lontoossa on monia tunnettuja jalkapalloseuroja, kuten Chelsea tai Arsenal”.

Tähän mennessä WWW-sivujen tietosisältö on tehty pääosin ihmisten luettavaksi. Saavutettaisiin valtavaa hyötyä jos WWW-sivujen semanttinen rakenne olisi myös tietokoneen ymmärtämässä muodossa. Tätä tavoitetta varten rakennetaan ontologioita kuvaamaan käsitteiden välisiä suhteita. Ontologioden muodostaminen käsin on hyvin työlästä, joten niiden automaattiseen muodostamiseen on kehitetty mm. klusterointiin perustuvia menetelmiä [Zen04]. Ontologiaa kuvaamaan tarvitaan formaali kieli jolla voidaan esittää ontologisia lausekkeita. WWW:n yhteydessä ehkä tunnetuin ontologiakieli on W3C:n suosittelema OWL (Web Ontology Language).

Semanttisen WWW:n ideana on, että automaattiset “agentit” voivat tarkastella WWW-sivujen semanttista tietoa ja tehdä siitä päättelyitä. Semanttisen WWW:n käyttöönottoa on hidastanut tarve tuplamerkkaukselle; eri merkkaukset ihmisiä ja koneita varten, jotka on ainakin tähän mennessä pitänyt tehdä pääosin käsin. Tavoiteltavampaa on ehkä se, että sivujen ontologiakielen mukainen merkkaus tehdään automaattisesti. Askeleita tähän suuntaan on esitetty PANKOW-menetelmässä (Pattern-based Annotation through Knowledge On the Web)[CHS04] . Menetelmä perustuu käsitteiden sekä erisnimien kategorisointiin annetun ontologian suhteen. Algoritmille annetaan syötteenä joukko WWW-sivuja, joilta etsitään lausekkeita joita todennäköisesti voitaisiin kategorisoida ontologiassa esiintyvän käsitteen ilmentymiksi, kuten erisnimiä “Helsinki”, “Matti Vanhanen”. Näin saadusta kandidaattijoukosta muodostetaan ontologian käsitteiden kanssa kielellisiä hahmoja, hypoteettisia virkkeitä kuten “Helsinki on kaupunki”, “Helsinki on hotelli”. Nämä virkkeet syötetään sitten olemassaolevalle hakukoneelle, kuten Googlelle, jonka löytämistä virkkeen esiintymien lukumääristä voidaan tehdä päätelmiä virkkeen relevanttiudesta. Jos virke saa tarpeeksi kannatusta, voidaan virkkeen erisnimi (Helsinki) merkata virkkeen käsitteen mukaisesti (kaupunki). Ylläoleva menetelmä on osaltaan askel WWW-sivun automaattiseen merkkaukseen.

5 Tiedon eristäminen

Yleensä vain murto-osa dokumentin sisällöstä on tiedon hakijalle relevanttia tietoa. *Tiedon eristämisellä* (data extraction, information extraction) tarkoitetaan olennaisen tiedon eristämistästä dokumenteista [Gri97]. Tiedon eristäminen etenee siten, että ensin järjestelmä eristää yksittäiset faktat tekstistä. Tämän jälkeen faktat yhdistellään muodostaen suurempia tosiasiakokonaisuuksia. Lopuksi faktoista muodostetaan haluttu formaatti, kuten jonkinlainen lomake, josta halutut tiedot löytyvät [Gri97]. Esimerkiksi blogikirjoituksesta voitaisiin haluta tietää otsikko, kirjoittaja, kirjoituspäivämäärä, varsinainen tekstisisältö ja kommenttien sisältö.

Eristäminen tapahtuu luomalla malleja, sääntöjä, jotka kertovat, miten haluttu tieto on esitetty tekstissä. Eristämisessä käytetään kahta sääntöä: alkuja loppusääntöä. Mistä kohde alkaa ja mihin se

päättyy [Liu06]. Sääntöjen luomisessa käytetään hyväksi sisällön rakenteisuutta tunnistamalla rakenteen tasoja ja eri osien suhteita. Rakenteiden tunnistamisessa sovelletaan esimerkiksi sanastollista analyysia, nimentunnistusta ja sanojen ryhmittelyä [Gri97]. Sanastollisessa analyysissä hyödynnetään tekstin muoto-opillisia seikkoja sekä sanastohakua. Muotoseikkojen avulla tekstistä voidaan erottaa esimerkiksi yrityksen nimiä isojen alkukirjainten ja Oy tai Inc päätteiden avulla. Kaikkien yritysten nimet eivät kuitenkaan sisällä näitä tunnisteita (esim. General Motors). Tällöin voidaan käyttää sanastohakua. Hakua varten täytyy olla olemassa sanasto, kuten luettelo tunnetuista yritysten nimistä. Nimentunnistuksessa dokumentista pyritään löytämään tiettyjä rakenteita, kuten päivämääriä ja rahayksiköitä. Nimet tunnistetaan säännöllisten lausekkeiden (engl. regular expression, regexp) avulla [Gri97]. Säännöllisillä lausekkeilla voidaan määritellä monipuolisia ehtoja eristettäviksi haluttaville kohteille. Esimerkiksi isolla kirjaimella alkavat sanat tulkitaan nimiksi. Yhteenkuuluvista sanoista on usein järkevää muodostaa ryhmiä. Esimerkiksi henkilön nimi kannattaa käsitellä kokonaisuutena “Sami Saari” ja verbit liitteineen yhtenä verbinä (“olisin laulanut”).

Tiedon yhdistelyvaiheessa pyritään tunnistamaan eri tapoja ilmaista sama asia ja päättelemään asioita dokumentista eristetyistä faktoista. Syntaktisessa eli lauseopillisessa analyysissä tunnistetaan sanojen paikkoja lauseessa ja paikkojen tuomia merkityksiä: verbiä ennen on tekijä, verbin jälkeen tekemisen kohde. Tärkeää on tulkita lause pala palalta: aikamuodot, prepositiot jne. Useiden vaikuttavien seikkojen takia lauseiden täydellinen syntaktinen ymmärtäminen on vaikeaa. Lisäksi verkko tuo omat haasteensa tiedon eristämiseen. Luonnollisen kielen käsittelymenetelmät ovat ongelmallisia verkkotekstien kanssa, sillä niiden onnistunut käyttö edellyttää kokonaisia, kieliopillisesti ehyitä lauseita [Sod97]. Verkossa teksti on paljon vapaampaa ja puhekielisempää. Verkossa menestyksekäs tiedon eristäminen on perinteisesti keskittynyt sivuihin, joilta on eristettävissä selkeitä tietorakenteita, kuten tauluja. Samat tekniikat eivät kuitenkaan ole suoraan sovellettavissa tekstiin, joka on pitkä ja kerronnallinen. Verkkotekstejä varten tarvitaan siis omia tekniikoita. Yksi tällainen on Webfoot [Sod97].

Verkon tietoa eristettäessä keskitytään sanojen ja osioiden keskinäisiin suhteisiin. Tarkoituksena on eristää kokonaisuuksia, jotka sisältävät tarpeelliset faktat, päätellä yhteenkuuluvat asiat ja näin tuottaa tietoa. Esimerkiksi verkosta löytyvä sääennustussivu saattaa sisältää useamman päivän ennustuksen usealta paikkakunnalta. Tyypillisesti luonnollisen kielen tulkkausjärjestelmä käsittelee lauseen kerrallaan ja soveltaa siihen lauseopillisia sääntöjä. Sääennustussivulta ei kuitenkaan löydy kokonaisia lauseita vaan lyhyesti esitettyä faktatietoa, josta pitäisi saada sidottua toisiinsa kuuluvat asiat yhteen. Webfoot saa syötteenä kyseisen sivun ja soveltaa siihen sivun ulkoasusta saatavia vihjeitä, kuten HTML-tageja. Webfoot jakaa tekstin loogisiin kokonaisuuksiin, jotka annetaan edelleen luonnollisen kielen menetelmiin perustuvalle eristysjärjestelmälle. Webfootin tehtävä siis on ikään kuin muokata verkkosivun sisällöstä lauseita korvaavia rakenteita [Sod97].

Eristyssääntöjä voidaan luoda sekä manuaalisesti että automaattisesti [Liu06]. Esimerkkisivujen manuaalinen merkitseminen on työlästä, etenkin jos kohdejoukko on valtava. Toisen suuren haasteen verkkomateriaalin eristämiselle aiheuttavat verkkosivujen dynaamisuus ja tiuhaan muuttuvat rakenteet. Jos sivun rakenteessa tapahtuu sääntöihin vaikuttava muutos, jo muodostettu sääntö on hyödytön. Tällöin tarvitaan uudelleenoppimista ja usein myös uutta manuaalista merkitsemistä. Automaattisessa sääntöjen opettelussa lähdetään liikkeelle esimerkkisivusta, josta muodostettuja eristyssääntöjä hiotaan muiden esimerkkisivujen avulla. Hienosäätöä tarvitaan, kun kielioppi kahden esimerkkisivun välillä ei täsmää [Liu06]. Verkkosivujen kohdalla eräs tapa luoda eristyssääntöjä on HTML-tagien avulla määritellä polkuja halutun tiedon sisältävään alueeseen. Menetelmä on yksinkertainen, mutta haasteita tuottavat sekä verkkosivustojen rakenteiden vaihtelevuus että virheelliset tagit.

6 Mielipiteiden louhinta

Internet ja WWW ovat täynnä erilaisia mielipiteitä esittäviä tekstejä. Erilaiset sivustot jotka tarjoavat

mahdollisuuden kirjoittaa sekä lukea tuotteiden arvosteluja ovat yksi esimerkki [LHC05, YiN05]. Toisaalta blogit tai muut yleiset sivut voivat myös sisältää mielenkiintoisia mielipiteitä [YiN05].

Esimerkiksi tuotearvoisteluissa esitettävillä mielipiteillä on vaikutus tuotteen menekkiin [Ni07].Tälläisiä tuotearvosteluita voi internetissä olla valtava lukumäärä. Jos tavallisella hakukoneella hakee arvosteluja jostain tuotteesta, niin kukin hakutulos vastaa vain yhtä tekstiä. Yleiskuvan luonti, vaikkakin epäilemättä hyödyllistä niin tuotteen myyjälle kuin kuluttajallekin, on vaikeaa [YuH03].

Mielipiteiden louhinnassa (engl. opinion mining) yritetään hyödyntää tekstimassan siseltämiä mielipiteitä luokittelussa. Kwon ja muut käyttävät mielipideanalyysia artikkeleiden pääargumentin löytämiseen [Kwo07]. Zhang ja muut [ZhV06] sekä Ni ja muut [Ni07] käyttävät tekstin sisältämien mielipiteiden määrää tekstin mielenkiintoisuuden, hyödyllisyyden tai laadun määrittämiseen. Ghose ja muut [GhI07] tutkivat tuotearviointitekstin sisältämien mielipiteiden vaikutusta tekstin hyödyllisyyteen kuluttujan näkökulmasta sekä tuotteen menekkiin. Jindal ja muut [JiL06] louhivat vertailuja kahden tuotteen välillä. Kuitenkin selvästi suurin osa artikkeleista käsittelee mielipideorientaatiota.

6.1 Mielipideorientaatio

Yksinkertaisimmillaan teksti voidaan luokitella *mielipideorientaatioltaan* (engl. opinion orientation), eli subjektiiviselta asennoitumiseltaan, joko positiiviseksi tai negatiiviseksi [YuH03, Tur01]. Esimerkiksi jos on annettu kokoelma tuotearvosteluja, niin kunkin tuotearvosteluartikkelin mielipideorientaatio voidan määrittää ja esittää tämä kokoelma laskettuja mielipideorientaatioita tuotteeseen kohdistuvana yleisenä mielipiteenä.

Pelkkä koko tekstin orientaatio ei välttämättä kerro kovin paljoa kohteesta. Esimerkiksi lause “Kameran akku on riittämätön” sisältää kamera ominaisuuden “akku” sekä siihen kohdistuvan mielipiteen “on riittämätön”. Tälläinen tarkempi tieto voi olla mielenkiintoista. Yksi tapa tarkentaa annettavaa yleiskuvaa on käyttää luokittelun kohteena tekstin lauseita [DLP03]. Tällöin käyttäjä voi mahdollisesti nähdä, minkälaiset kohdetta käsittelevät lauseet ovat negatiivisia tai positiivisa. Toisaalta tekstistä voidaan yrittää erityisesti löytää syitä mielipiteelle [KiH06].

Yleisesti kohteella voi olla monia eri ominaisuuksia joihin mielipide tekstissä voi kohdistua. Yksi mahdollissuus mielipiteen louhinnassa on kiinnittää huomio erityisesti kohteen ominaisuuksiin. Tällöin on kaksi vaihtoehtoa. Joko käytetään valmiiksi annettuja ominaisuuksia [ZJZ06] tai itse tuotteen ominaisuuksia voidaan yrittää löytää automaattisesti[LHC05]. Toisaalta tuotteen ominaisuuksien löytäminen ei välttämättä ole näin helppoa[LHC05]. Edellisessä esimerkissä kameran ominaisuus “koko” sisältyi explisiittisesti lauseeseen. Tekstissä voidaan viitata myös johonkin tuotteen ominaisuuteen implisiittisesti, eli nimeämättä varsinaista viitattua ominaisuutta. Esimerkiksi voidaan todeta että “Kamera on liian iso”. Jälkimmäinen lause on kommentti kameran koosta vaikkei ominaisuutta “koko” mainita lauseessa. Tämä on yksi esimerkki haasteista, joita mielipiteiden louhinnasssa kohdataan.

6.2 Mielipidelouhinnan haasteita

Tekstin ymmärtämistä varten kaikki käsitellyt artikkelit käyttävät jotain yhdistelmää luonnollisen kielen käsittelystä sekä jostain koneoppimis -tai luokittelualgoritmista. Luonnollisen kielen käsittely on itsessään hankala ongelma. Toisaalta myös hyvän luokittelijan opettaminen on hankalaa, koska tarvitaan sopiva valmiiksi luokiteltu opetusmateriaali sekä siksi, kuten useassa käsitellyssä paperissa todetaan, että luokittelu ei ole välttämättä ylipäätään mitenkään selvä. Eri ihmiset voivat luokitella aineiston eri tavoilla [KiH06, ZJZ06, Kwo07]. Yksi tapa ratkaista, ainakin osittain, opetusmateriaalin muodostamisen hankaluus on käyttää valmiiksi arvioituja tekstejä esimerkiksi sivustoilta joiden formaattiin kuuluu tekstiarvion lisäksi annettava arvosana [DLP03, KiH06].

Mielipidelouhinnan toimivuus voi riippua myöskin aihepiiristä, jota analysoitava teksti käsittelee[KiH06, DLP03, ZJZ06, Tur01]. Esimerkiksi arviot elektroniikkatuotteista eroavat kieleltään ravintola-arviosta [KiH06]. Siinä missä eloktroniikkatuotteilla on selvä joukko ominaisuuksia, joita arvostelut käsittelevät

niin tyypilliset arvostelut ravintoloista sisältävät mielipiteitä abstrakteista sekä vaihtelevista ominaisuuksista. Toisaalta kirjoissa, elokuvissa ja musiikissa voidaan usein käyttää mielipiteenomaisia sanontoja esimerkiksi juonilyhennelmissä, jolloin analyysi voi virheellistä tulkita nämä mielipiteiksi itse elokuvasta tai kirjasta[DLP03].

7 WWW-sivujen osittelu

Monet erityisesti kaupalliset WWW-sivut koostuvat useasta erillisestä osasta, joilla on oma merkityksensä ja erityispiirteensä [YoR02]. Osia voivat olla esimerkiksi mainospalkki, navigointipalkki, yrityksen tiedot sekä varsinainen data-alue. Jokaista osaa voidaan pitää erillisenä yksikkönä, koska ne muodostavat oman loogisen kokonaisuutensa, joka on mielekäs ilman muitakin osia. Jos WWW-sivun osat voidaan erottaa toisistaan automaattisesti ja luokitella, voidaan näiden tietojen avulla tarkentaa useita WWW-louhinnan menetelmiä. Osia voidaan poistaa kokonaan käsittelystä tai sitten niille voidaan antaa painokertoimia sen mukaan, miten oleellisia ne ovat tehtävän kannalta [Son04].

Oletetaan esimerkiksi, että hakukoneen asiakas haluaa löytää tietoa jääkiekon MM-kilpailuista ja tekee haun yksinkertaisesti sanoilla "jääkiekko MM-kilpailut". Oletetaan lisäksi, että samana vuonna on julkaistu jääkiekkopeli, joka sijoittuu MM-kilpailuihin ja peliä mainostetaan vuolaasti monilla tietokonepeleihin liittyvillä sivuilla. Jos hakukone etsii sanoja "jääkiekko MM-kilpailut" koko WWW-sivuilta, se löytää paljon tapauksia tietokonepelisivuilta, jotka mainostavat uutta peliä. Jos taas hakukone on osannut ositella sivut ja hakee sanoja vain sivujen dataosasta, jäävät tällaiset virheelliset vastaavuudet pois. Toisaalta jos hakija ilmaisee olevansa kiinnostunut nimen omaan jääkiekon MM-kilpailuihin liittyvistä tuotteista, voivat mainokset saada jopa suuremman painoarvon kuin varsinainen data.

Ositusta voidaan käyttää myös vähentämään kohinaa esimerkiksi sivujen luokittelussa tai strukturoidun tiedon louhinnassa. Tämä perustuu siihen, että eri osissa oleva data on hyvin erilaista. Jos esimerkiksi sivulta etsitään yleisiä hahmoja, voi kaikkein yleisimmäksi nousta yhtiön tietoruudun "nimi osoite puhelinnumero WWW-sivu" tyyppinen hahmo, joka voi olla täysin merkityksetön tiedonlouhinnan kannalta. Ottamalla mukaan vain merkitykselliset sivujen osat, saadaan louhittavasta datasta homogeenisempää ja enemmän merkityksellisiin asioihin keskittynyttä, kuin louhimalla kokonaisia sivuja.

Sivujen osittelusta on iloa myös tavalliselle WWW-surffaajalle. Osittelun avulla voidaan nimittäin estää tarpeettomia sivujen osia latautumasta. Jos käyttäjä ei esimerkiksi ole kiinnostunut mainoksista, voidaan koko mainospalkki jättää lataamatta, mikä nopeuttaa sivujen latautumista ja vähentää sivujen sekavuutta. Vastaavasti voitaisiin jättää myös pois vaikkapa yrityksen tietoruutu tai navigointipalkki. Jos käyttäjä hakee vain tietoa jostain asiasta, on ehkä paikallaan jättää kaikki muut osat pois paitsi varsinainen data-alue. Näin käyttäjä voi keskittyä vain siihen, mistä on oikeasti kiinnostunut.

Sivujen osat voidaan etsiä käyttämällä hyväksi HTML:n puurakennetta [YoR02]. *Sivun osa* (engl. pagelet) määritellään HTML-elementiksi, jonka lapsilla on korkeintaan k hyperlinkkiä ja jonka yksikään edeltäjä ei ole sivun osa. Tämän määritelmän on tarkoitus varmistaa, että sivu jaetaan osiin, jotka käsittelevät yhtä aihetta, mutta ovat kohtuullisen suuria. Näin määriteltyjä sivun osia voidaan etsiä automaattisesti käymällä HTML-sivun puurakennetta juuresta alkaen läpi.

Kun sivujen osat on etsitty, voidaan niiden avulla etsiä toistuvia *malleja* (engl. template) [YoR02]. Esimerkiksi kaikilla yrityksen WWW-sivuilla esiintyvä yrityksen tietoruutu ja navigaatiopalkki ovat tällaisia malleja. Mallit voidaan etsiä yksinkertaisella menetelmällä, jossa ensin etsitään jokaiselle sivun osalle sivulla esiintyvät tietyn mittaiset *alimerkkijonot* (engl. shingle, [Bro97]). Seuraavaksi sivujen osat ryhmitellään alimerkkijonojen perusteella siten, että samankaltaiset alimerkkijonot sisältävät sivujen osat päätyvät samaan ryhmään. Tällainen ryhmittely ei ole riippuvainen pienistä eroista sivujen osissa, vaan vastaa käsitystämme suunnilleen samanlaisista [Bro97, YoR02]. Löydetyt ryhmät ovat tietyn mallin toteuttavia sivun osia. Poistamalla nämä ryhmät ennen esimerkiksi tiedonhakua WWW-sivuilta voidaan

haun tarkkuutta selvästi parantaa [YoR02].

Viitteet

[Bro97] Broder, A. Z., Glassman, S. C., Manasse, M. S. ja Zweig, G., Syntactic clustering of the web. *Selected papers from the 6th int. conf. on World Wide Web*, Essex, UK, 1997, Elsevier Science Publishers Ltd., sivut 1157–1166.

[YoR02] Bar-Yossef, Z. ja Rajagopalan, S., Template detection via data mining and its applications. *WWW '02: Proc. of the 11th int. conf. on World Wide Web*, New York, NY, USA, 2002, ACM Press, sivut 580–591.

[ChC06] Chang, K. C.-C. ja Cho, J., Accessing the web: from search to integration. *SIGMOD '06: Proc. of the 2006 ACM SIGMOD int. conf. on Management of data*, New York, NY, USA, 2006, ACM Press, sivut 804–805.

[CHS04] Cimiano, P., Handschuh, S. ja Staab, S., Towards the self-annotating web. *Proc. of the 13th World Wide Web Conference*, 2004.

[Cas03] Castillo, J., Silvescu, A. ja D. Caragea, J. Pathak, V. H., Information extraction and integration from heterogeneous, distributed, autonomous information sources - a federated ontology-driven query-centric approach. *Information Reuse and Integration, 2003. IRI 2003. IEEE int. conf.*, 2003, sivut 183 – 191.

[DLP03] Dave, K., Lawrence, S. ja Pennock, D. M., Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW '03: Proc. of the 12th int. conf. on World Wide Web*, New York, NY, USA, 2003, ACM Press, sivut 519–528.

[GhI07] Ghose, A. ja Ipeirotis, P. G., Designing novel review ranking systems: predicting the usefulness and impact of reviews. *ICEC '07: Proc. of the 9th int. conf. on Electronic commerce*, New York, NY, USA, 2007, ACM Press, sivut 303–310.

[Gri97] Grishman, R., Information extraction: Techniques and challenges. *Lecture Notes in Artificial Intelligence*. Springer, 1997.

[JiL06] Jindal, N. ja Liu, B., Identifying comparative sentences in text documents. *SIGIR '06: Proc. of the 29th annual int. ACM SIGIR conf. on Research and development in information retrieval*, New York, NY, USA, 2006, ACM Press, sivut 244–251.

[KiH06] Kim, S.-M. ja Hovy, E., Automatic identification of pro and con reasons in online reviews. *Proc. of the COLING/ACL on Main conf. poster sessions*, Morristown, NJ, USA, 2006, Association for Computational Linguistics, sivut 483–490.

[Kri04] Kummamuru, K., Lotlikar, R., Roy, S., Singal, K. ja Krishnapuram, R., A hierarchical monothetic document clustering algorithm for summarization and browsing search results. *WWW '04: Proc. of the 13th int. conf. on World Wide Web*, New York, NY, USA, 2004, ACM Press, sivut 658–665.

[Kwo07] Kwon, N., Zhou, L., Hovy, E. ja Shulman, S. W., Identifying and classifying subjective claims. *dg.o '07: Proc. of the 8th annual int. conf. on Digital government research*. Digital Government Research Center, 2007, sivut 76–81.

[LiC04] Liu, B. ja Chen-Chuan-Chang, K., Editorial: special issue on web content mining. *SIGKDD Explor. Newsl.*, 6,2(2004), sivut 1–4.

[LCN03] Liu, B., Chin, C. ja Ng, H., Mining topic-specific concepts and definitions on the

web. *Proc. of the 12th Int. World Wide Web Conference*, 2003.

[LHC05] Liu, B., Hu, M. ja Cheng, J., Opinion observer: analyzing and comparing opinions on the web. *WWW '05: Proc. of the 14th int. conf. on World Wide Web*, New York, NY, USA, 2005, ACM Press, sivut 342–351.

[Liu05] Liu, B., Tutorial on web content mining, 2005.

[Liu06] Liu, B., *Web Data Mining*. Springer, 2006.

[May07] Natural language technology for information integration in business intelligence. *10th int. conf. on Business Information Systems*, 2007.

[Net07] Netcraft; web server survey. URL `http://news.netcraft.com/archives/web_server_survey.html`. 2007.

[Ni07] Ni, X., Xue, G.-R., Ling, X., Yu, Y. ja Yang, Q., Exploring in the weblog space by detecting informative and affective articles. *WWW '07: Proc. of the 16th int. conf. on World Wide Web*, New York, NY, USA, 2007, ACM Press, sivut 281–290.

[Son04] Song, R., Liu, H., Wen, J.-R. ja Ma, W.-Y., Learning block importance models for web pages. *WWW '04: Proc. of the 13th int. conf. on World Wide Web*, New York, NY, USA, 2004, ACM Press, sivut 203–211.

[Sod97] Soderland, S., Learning to extract text-based information from the world wide web. *Proc. of 3rd Int. Conf. of Knowledge Discovery and Data Mining*, 1997.

[Tur01] Turney, P. D., Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *ACL '02: Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 2001, Association for Computational Linguistics, sivut 417–424.

[YuH03] Yu, H. ja Hatzivassiloglou, V., Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *Proc. of the 2003 conf. on Empirical methods in natural language processing*, Morristown, NJ, USA, 2003, Association for Computational Linguistics, sivut 129–136.

[YiN05] Yi, J. ja Niblack, W., Sentiment mining in webfountain. *ICDE '05: Proc. of the 21st Int. Conf. on Data Engineering (ICDE'05)*, Washington, DC, USA, 2005, IEEE Computer Society, sivut 1073–1083.

[Zen04] Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y. ja Ma, J., Learning to cluster web search results. *SIGIR '04: Proc. of the 27th annual int. ACM SIGIR conf. on Research and development in information retrieval*, New York, NY, USA, 2004, ACM Press, sivut 210–217.

[ZJZ06] Zhuang, L., Jing, F. ja Zhu, X.-Y., Movie review mining and summarization. *CIKM '06: Proc. of the 15th ACM int. conf. on Information and knowledge management*, New York, NY, USA, 2006, ACM Press, sivut 43–50.

[ZhV06] Zhang, Z. ja Varadarajan, B., Utility scoring of product reviews. *CIKM '06: Proc. of the 15th ACM int. conf. on Information and knowledge management*, New York, NY, USA, 2006, ACM Press, sivut 51–57.