Tiedon tallennus ja haku

Kalervo Järvelin ja Eero Sormunen

Tiedonhaku on oppiaine ja tutkimusala, joka sijoittuu informaatiotutkimuksen ja tietojenkäsittelyopin raja-alueelle. Netin ja sen hakukoneiden kehittyessä tiedonhausta on tullut kaikille tuttu, itsestään selvä asia. ”Googlaaminen” tuo kenen tahansa ulottuville koko verkostoituneen maailman digitaaliset tietoresurssit. Tässä artikkelissa ”googlaaminen” asetetaan tiedonhaun historialliseen viitekehykseen, valotetaan tiedonhakujärjestelmiä ja tiedonhaun ilmiöitä hiukan pintaa syvemmältä ja esitellään tiedonhaun tutkimuksen eri osa-alueita. Tekstin juonta vievät eteenpäin ja havainnollistavat nettiin liittyvät tiedonhakuesimerkit.

Esimerkkitiedonhaku

Jokaisella tiedonhakuja paljon tehneellä lienee sekä positiivisia että negatiivisia kokemuksia tiedonhausta. Google on kehittynyt suosituimmaksi hakukoneeksi – eikä varmasti aiheetta. Useimmiten se löytää hämmästyttävän nopeasti ja tarkasti Webin satojen miljoonien sivujen joukosta juuri ne sivut, joita olemme hakemassa. Esimerkiksi, jos teet haun tämän artikkelin kirjoittajien nimellä, saat hakutukoksen kärkeen kotisivumme vaikka kummankin nimi esiintyy tuhansilla Web-sivuilla. Toisaalta löytyy helposti esimerkkejä hakutarpeista, joissa hyvään hakutulokseen pääseminen ei ole helppoa. Tarkastelemme seuraavaksi esimerkkiä haastavasta hakutehtävästä, jonka avulla tiedonhakujärjestelmien toimintaperiaatetta, hakemisen mahdollisuuksia ja rajoituksia sekä niiden tutkimusta on helpompi havainnollistaa.

Olettakaamme, että ystävämme Pekka on joutunut seuraavaan tilanteeseen. Pekan irlantilainen ystävä Alan on innostunut Suomessa käydessään saunomisesta ja haluaa tietää, miten rakentaa sauna Dubliniin. Pekalla on kyllä kokemusta saunojen rakentamisesta, mutta hän ajattelee viisaimmaksi hankkia jotain “kättä pitempää” eli englanninkielisiä saunanrakennusohjeita. Netistä niitä varmaan löytyisi. Informaatiotutkimuksen termein ilmaisten Alanilla on *tiedontarve (information need)*, jonka Pekka pyrkii tyydyttämään hakemalla sopivaa aineistoa (dokumentteja, joka sisältävät tarvittua informaatiota).

Pekka on käyttänyt netin erilaisia palveluja ja tietää, että on useita vaihtoehtoisia tapoja etsiä informaatiota. Yksi mahdollisuus olisi lähteä selailemaan (surfailemaan, navigoimaan) omalta kotisivulta WWW:n *hypertekstirakenteissa* eli käyttää hyväksi dokumenttien välisiä *linkkejä*. Tuo vaihtoehto ei kuitenkaan vaikuta houkuttelevalta, koska Pekan mielessä ei ole yhtään aiheeseen liittyvää *sivustoa (site)* tai *hakemistopalvelua* (*directory service*), josta aloittaa selailu. Siksi hän turvautuu *sanahakupalveluun* (*search service*), joista hän valitsee itselleen tutuimman eli *Googlen*.

Google on yksi monista sanahakupalveluista, joista suurimmat tarjoavat hakumahdollisuuden kymmeniin miljardeihin WWW-dokumentteihin[1]. Palvelujen kate, haku- ja muut ominaisuudet vaihtelevat, mutta kaikki sanahakupalvelut perustuvat kolmen päätoiminnon varaan:

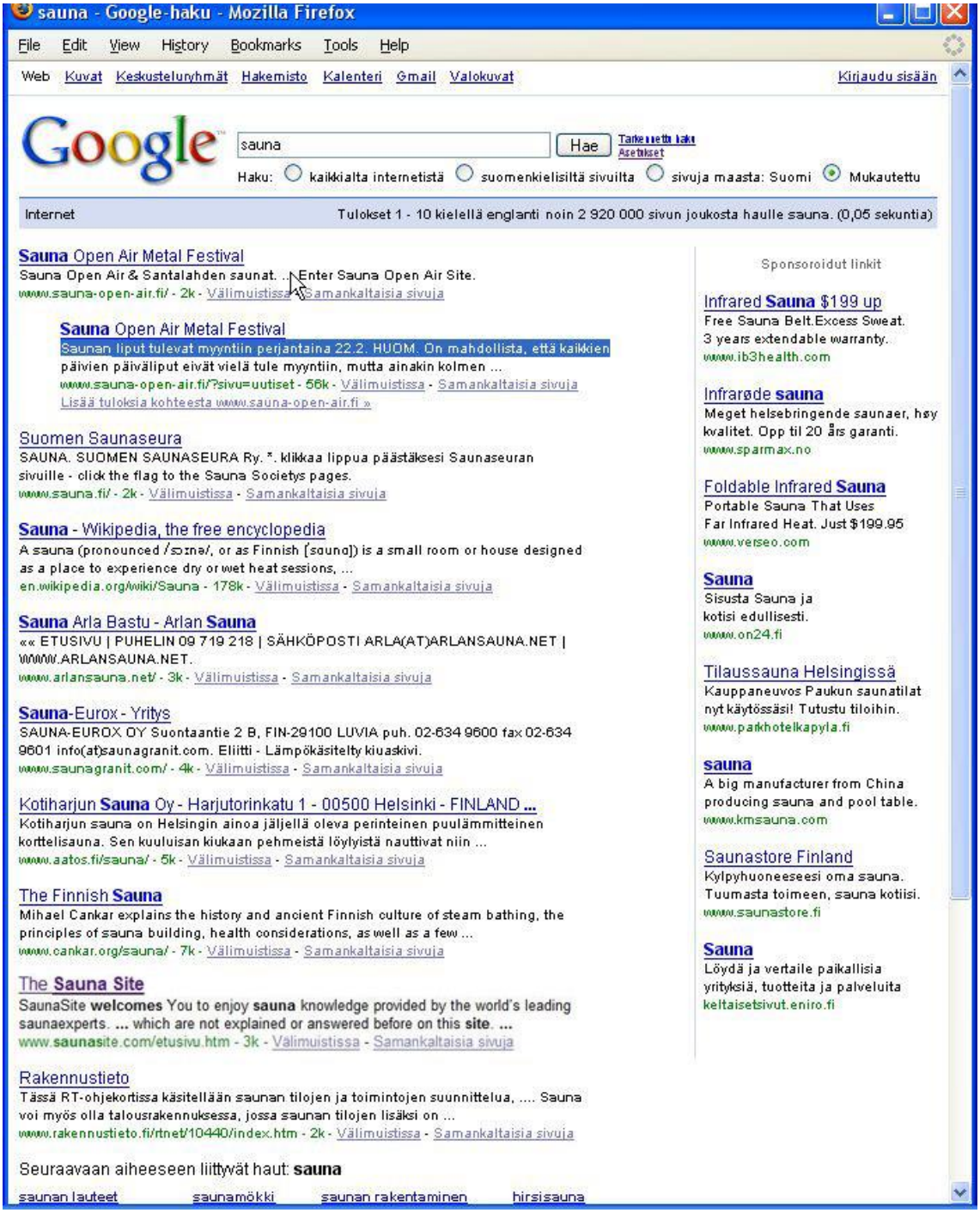
- Aineiston kartoitus ja kokoaminen. *Keräilyrobotit* (*crawlers*) noutavat verkon dokumentteja laajentaen automaattisesti keräilyaluettaan dokumenttien välisiä linkkejä seuraten.
- Tietokannan rakentaminen. Kerättyjen dokumenttien osoite- ja viitetiedot tallennetaan *tietokantaan* (*data base*) ja dokumentin sisältämät sanat tallennetaan tietokannan hakemistoon.
- Hakupalvelu. Tietokannan ylläpito perustuu tiedonhakujärjestelmään (information retrieval system), jonka avulla koottuun tietokantaan voidaan kohdistaa kyselyjä (queries). Mitä tahansa verkkodokumenttiin sisältynyttä merkkijonoa voidaan käyttää kyselyä määrittävänä *hakuavaimena* (*search key*).

Hakupalvelun tekniseen toteutukseen liittyvän tiedonhakujärjestelmän periaatetta on selvitetty tekstilaatikossa "Tiedonhakujärjestelmät" ja palvelun tuottamisen ja käytön prosessia laatikossa "Tallennus- ja hakuprosessi".

Pekka kirjoittaa Googlen hakulomakkeelle *hakusanan* “sauna” ja rajaa kyselyn englanninkielisiin dokumentteihin. Hetken kuluttua Google vastaa, että lähes 3 miljoonaa dokumenttia on löydetty hakusanalla “sauna”. Pekka ryhtyy tutkimaan saalistaan *tuloslistasta* (Kuva 1). Hän käy kaikkiaan läpi 20 ensimmäistä dokumenttilinkkiä. Pekan tarkoituksena on tunnistaa sellaiset linkit, jotka ilmeisesti

[1] SearchEngineWatchin mukaan Googlen kate oli vuonna 2007 noin 24 miljardia sivua, ks. http://blog.searchenginewatch.com/blog/071001-105927.

tarjoaisivat ohjeita saunanrakennukseen. Jos sivu sisältää saunanrakennusohjeita, hän tulostaa sivun paperille, tallentaa sen tiedostoksi tietokoneensa kovalevylle ja/tai tallentaa osoitteen selaimensa *kirjanmerkkilistaan* (bookmark).



Kuva 1. Google -palvelun hakulomake ja tuloslistan alkuosa, linkit 1-10.

Kun Pekka silmäilee löytyviä dokumentteja ja arvioi niitä saunanrakennushankkeen valossa, hän tekee dokumenteille *relevanssiarvion*. Pekan relevanssiarviossa on kyse

Alanin rakennushankkeeseen liittyvien tiedontarpeiden ja löytyvien dokumenttien vertailusta. Jos dokumentti sisältää Pekan arvion mukaan käyttökelpoista tietoa saunanrakennuksesta, dokumentin sanotaan olevan *relevantti*, muutoin *epärelevantti*.

Pekka tekee kyselyn tuloksesta muun muassa seuraavia huomioita. Ensimmäinen linkki[2] on täysi harhalaukaus: Mansen metallimusafestarien pääsivu. Toinen, kolmas ja seitsemäs linkki vaikuttavat jossain määrin hyödyllisiltä. Sieltä avautuvat saunaseuran kotisivut, Wikipedian saunasivut ja yksittäisen harrastajan saunasivut. Aineisto on kuitenkin yleisluonteista ja antaa vain taustatietoa rakentamisesta. Vasta kahdeksas linkki on selkeästi rakentajalle tarkoitettu. ”The Sauna Site” sisältää saunan rakennusohjeita mutta ei rakennepiirustuksia. Kymmenes linkki vie yksityiskohtaiseen saunan rakentamistietoon eli RT-ohjekortille. Tarjolla on mitoitusohjeita ja mallipiirustuksia mutta Pekan harmiksi RT-ohjekortti on saatavissa vain suomeksi.

Yhteenvetona 20 ensimmäisestä tuloslinkistä Pekka toteaa, että

- kaksi linkkiä johti erittäin relevanttiin dokumenttiin (toisen kieli suomi)
- kolme linkkiä vei jossain määrin relevantteihin dokumentteihin
- loput 15 (16) linkkiä olivat hyödyttömiä.

Relevantit linkit sijoittuivat tuloslistalle seuraavasti (kuvio 1).

Linkin no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevanssi	-	+	+	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-

Kuvio 1. Pekan hakutuloksen relevanssitarkastelu

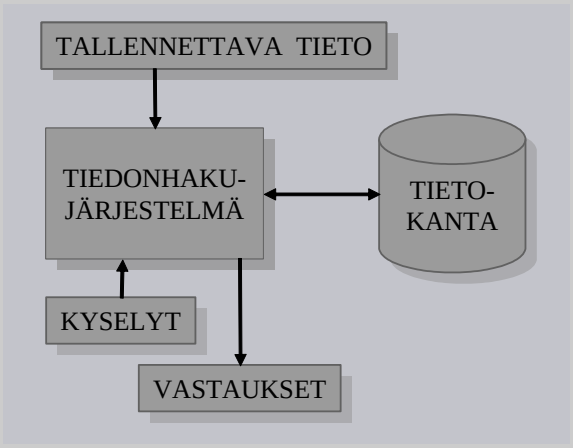
Oliko Pekan haku onnistunut? Tähän on vaikea vastata tyhjentävästi useastakin syystä. Se voidaan todeta, että sanahakupalvelu ei toiminut ideaalisesti, koska suurin osa sen tarjoamista linkeistä oli hyödyttömiä. Toisaalta parinkymmenen linkin selailu ja kiinnostavien sivujen tulostaminen oli pieni vaiva, jos saatu informaatio auttaa Alania ratkaisemaan saunanrakentamiseen liittyvät ongelmat. Tätä emme kuitenkaan pysty arvioimaan, koska emme itse tunne Alanin yksilöllisiä tiedontarpeita emmekä Pekan kykyä hahmottaa niitä.

[2] Tuloslistan kaksi ensimmäistä linkkiä lasketaan tässä samaksi, sillä sisennetty linkki on samalta sivustolta kuin ensimmäisenä listattu linkki. Vinkit alasivuille on Googlen tapa korostaa hakutulosten jotain yksityiskohtaa.

Saunan rakentamiskokemuksensa perusteella Pekka ymmärtää, mitä tietoja saunan rakentamisessa *yleensä* tarvitaan. Toisaalta hän ei välttämättä tiedä Alanin tiedontarpeiden yksityiskohtia; esimerkiksi mitkä ovat Alanin valmiudet hyödyntää teknisiä ohjeita tai mitä paikallisia erityispiirteitä saunan rakentamiseen liittyy Dublinissa ja Alanin taloyhtiössä. Pekan relevanssiarvio pystyy siis ilmaisemaan vain minkä linkkien välityksellä löytyy sauna-aiheeseen liittyviä, *potentiaalisesti* hyödyllistä tietoa.

Tiedonhakujärjestelmät

Tiedonhakujärjestelmät ovat paljon vanhempi ilmiö kuin Internet ja WWW. Perinteisiä sovelluksia ovat muun muassa kirjastojen atk-luettelot sekä eri alojen tiedonhankintaa helpottavat viite- ja tekstitietokannat. Uudempia sovelluksia ovat muun muassa digitaaliset kuva-, video- ja äänitetietokannat.



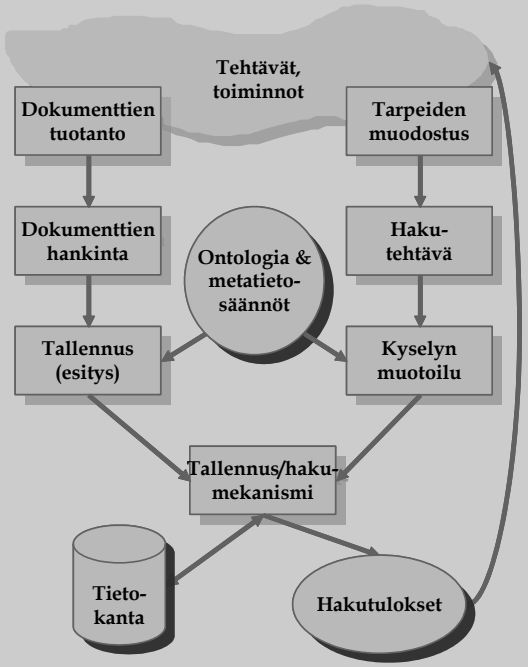
Oheinen kaavio esittää, miten tallennettava tieto, tiedontarpeita kuvaavat kyselyt ja niihin saadut vastaukset, *hakutulokset*, liittyvät tiedonhakujärjestelmän toimintaan. Uusien tietojen tallentamista tietokantaan ja siellä jo tallennettujen tietojen ajantasallapitämistä kutsutaan *päivittämiseksi*. Tiedontarpeet muotoillaan kyselykielen *kyselyiksi*, joita tiedonhakujärjestelmä pystyy käsittelemään vastauksien tuottamiseksi.

Tiedon tallennus- ja hakujärjestelmä (*information storage and retrieval system*) on tiedonhallintajärjestelmä. Lyhyttä nimitystä *(tiedon) hakujärjestelmä* käytetään usein koko nimen sijasta. Se kattaa myös tallennus-toiminnon, ellei erikseen toisin mainita. Internetissä hakujärjestelmästä käytetään usein nimitystä *hakukone* (*search engine*).

Hakujärjestelmän käyttö johtaa hakuprosessiin, jota jäsennetään tekstilaatikossa Tallennus- ja hakuprosessi.

Tiedon tallennus- ja hakuprosessi

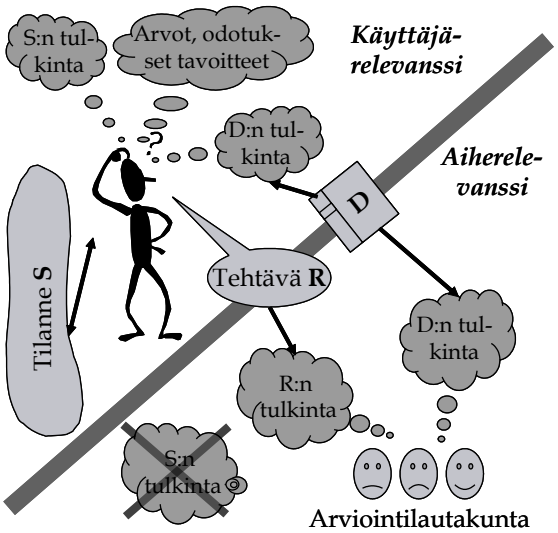
Oheinen kuva havainnollistaa tiedon tallennus- ja hakuprosessia. Yhteisössä esiintyvät toiminnot ja tehtävät ovat dokumenttien tuotannon ja tiedontarpeiden muodostuksen lähtökohtia. Dokumenttien tuottajat, kuten tutkijat, kirjailijat ja kustantajat kirjoittavat ja kustantavat dokumentit (ml. epävirallinen "verkkokustantaminen"). Tietokantojen tuottajat suorittavat tallennustyön, ts. hankkivat dokumentit ja muokkaavat ne tietokannoissa esitettäviksi. Tämä voi tarkoittaa myös pelkästään dokumenttien metadatan luomista ja indeksointipiirteiden valintaa, jolloin hakutulokset sisältävät metadataa ja linkin itse dokumentteihin. Metadatasäännöt opastavat kuvaamaan dokumentin tunnistamiseen tarvittavat tiedot kuten tekijät, nimekkeet ja julkaisuvuodet. Dokumentin sisällön kuvaamiseen voidaan käyttää ontologiaa, joka määrää, millaisin käsittein dokumenttien sisältöä annotoidaan eli kuvataan. Usein kuvaukseen kuitenkin käytetään dokumenttien sisältämiä indeksointipiirteitä, kuten sanoja, suoraan. Yleensä tietokannan tuottaja tarjoaa tietokannan haettavaksi, mutta voi myös myydä sen hakupalvelun tuottajalle, joka tarjoaa tietokannan käytettäväksi esimerkiksi webin välityksellä.



Yhteisön toiminnassa toimijoille muodostuu tiedontarpeita, joiden tyydyttämiseksi voidaan muotoilla enemmän tai vähemmän tietoisesti hakutehtävä, jota lähdetään suorittamaan. Usein tiedontarvitsija muotoilee hakutehtävän itse esimerkiksi jollekin webin hakukoneelle haettavaksi. Joissain tapauksissa tiedontarvitsija voi kääntyä välittäjän (tietopalvelun ammattilaiset kuten kirjastonhoitajat ja informaatikot) puoleen ja delegoida hakutehtävän suorituksen hänelle. Tavoitteena on löytää tietokannasta dokumentteja, jotka auttavat tiedontarvitsijan ongelmanratkaisua tai muuta toiminnan suorittamista.

Relevanssi ja hakutuloksen arviointi

Pekan hakutuloksesta tekemien relevanssiarvioiden ja Alanin tiedontarpeen vertailu johtaa mielenkiintoisiin kysymyksiin: Mitä relevanssilla tarkoitetaan ja miten tiedonhaun tai rajatummin kyselyn onnistumista mitataan? Nämä ovat *tiedonhakututkimuksessa* ja *tiedonhakujärjestelmien evaluoinnissa* (hakutehokkuuden arvioinnissa) tärkeitä periaatteellisia kysymyksiä. Borlund (2003), Ingwersen ja Järvelin (2005) sekä Saracevic (2007) tarjoavat perusteelliset katsaukset relevanssin käsitteeseen, jota seuraavassa esitellään lyhyesti.



Kuvio 2. Aihe- ja käyttäjärelevanssi

Relevanssin käsite

Tiedonhaun tarkoituksena on löytää *relevanttia informaatiota* sitä tarvitsevalle. Relevanssin määrittelyssä on kaksi pääsuuntaa: *aiherelevanssi* ja *käyttäjärelevanssi* (Kuvio 2). Aiherelevanssi tarkoittaa sitä, että dokumentti käsittelee hakutehtävän määrittelemää aihetta (esimerkiksi saunan rakentamista tai digitaalista kuvankäsittelyä). Dokumentin aiherelevanssin voi arvioida muun muassa hakutehtävän aihepiiriä tuntevat henkilöt, relevanssiarvioijat (arviointilautakunta), vertaamalla hakutehtävää kuhunkin dokumenttiin. Tiedonhakujärjestelmän evaluointi ja tutkimus on joustavampaa, kun hakija- ja tilannekohtaisia tekijöitä ei tarvitse huomioida. Esimerkissämme Pekalla oli asiantuntijana edellytykset hakutulosten arviointiin aiherelevanssin pohjalta.

Käyttäjärelevanssi huomioi dokumentin aiheen lisäksi tiedon käyttäjästä riippuvia tekijöitä — hänen arvionsa dokumenttien käyttökelpoisuudesta. Käyttäjän arvioon

vaikuttavat muun muassa tiedontarpeen aiheuttavan tehtävän luonne ja tavoitteet, tilanteen tulkinta, esitiedot, dokumentin tulkinta (käsitys sisällöstä, kieli, ulkoasu, tuttuus käyttäjälle, jne.). Esimerkissämme Alan voisi pitää amerikkalaisten suunnitteluohjeita hyödyttöminä, koska ne eivät tule saunakulttuurin alkulähteiltä. Hänen arvionsa voi myös muuttua dynaamisesti: aluksi relevanttina pidetty dokumentti voidaan arvioida hyödyttömäksi sitä olennaisesti paremman löydyttyä. Tämän luonteisia kriteereitä Pekan on vaikea ottaa huomioon. Pekalla ei myöskään ole tarkkaa tietoa Alanin tilanteesta, joten relevanssiarvio ei voi ottaa tilannetta huomioon.

Tiedontarve määritellään usein tiedontarvitsijan kokemukseksi tilanteensa ja ympäristönsä epäselvyydestä (epävarmuudesta) ja tiedon hyödyllisyydestä tilanteen hallinnassa (*sense making*). Nykytilanteen hallinta edellyttää tiedontarvitsijan ymmärtävän edesmenneet läpikäydyt tilanteet ja odotettavissa olevat tulevat tilanteet. Relevanssiarviot ovat tällöin tilannesidonnaisia ja dynaamisia. Tiedontarvitsija arvioi henkilökohtaisesti käytettävissä olevan tiedon ja oman tilanteensa suhdetta. Tältä pohjalta relevanssille voidaan antaa seuraava määritelmä:

> *Relevanssilla tarkoitetaan informaation arvioitua käyttökelpoisuutta tietyssä käyttötilanteessa ottaen huomioon käyttäjän tavoitteet, arvot ja odotukset.*

Vuosien kuluessa tiedonhakututkimuksen päähuomio on siirtynyt aiherelevanssista kohti käyttäjärelevanssia. Tämä ei kuitenkaan tarkoita sitä, että tutkimuksessa tulisi tai edes voitaisiin aina käyttää käyttäjärelevanssiin perustuvaa hakutuloksen arviointiperustaa aiherelevanssin sijasta. Monissa järjestelmäkeskeisissä tutkimusasetelmissa aiherelevanssi antaa yksinkertaisemmin toteutettavan ja riittävän pohjan hakujärjestelmän arvioinnille. Valinta riippuu tutkittavasta ongelmasta. Aiherelevanssin määrittelyksi riittää usein seuraava:

> *Dokumentti on relevantti hakutehtävän kannalta, jos dokumentissa käsitellään hakutehtävän aihetta edes jossain määrin.*

Hakutuloksen arviointi: saanti ja tarkkuus

Hakutulosten arvioimiseksi tarvitaan arviointikriteerejä, mittoja, joihin verraten tuloksesta voidaan jotakin sanoa suhteessa muihin hakutuloksiin. Tavallisimmat ja laajimmin käytetyt hakutulosta koskevat arviointikriteerit ovat *saanti* ja *tarkkuus*.

Hakutulos jakaa tietokannan dokumentit kahteen ryhmään: haussa löydettyihin ja haussa hylättyihin. Toisaalta dokumenteille tehty relevanssiarvio jakaa dokumentit periaatteessa myös kahteen ryhmään: relevantteihin ja epärelevantteihin. Tavallisesti relevanssiarviota ei kuitenkaan ole saatavissa muista kuin löydetyistä dokumenteista. Tilannetta voidaan havainnollistaa taulukon 1 avulla. Solujen dokumenttimäärien summa **a + b + c + d** edustaa koko tietokannan dokumenttien lukumäärää.

Taulukko 1. Saannin ja tarkkuuden määrittelytaulukko

	Relevanssiarvio		
Haun tulos	**Relevantti**	**Epärelevantti**	**Summa**
Löydetty	**a** osumat	**b** häly	**a + b**
Hylätty	**c** hukatut	**d** torjutut	**c + d**
Summa	**a + c**	**b + d**	**a + b + c + d**

Hakutuloksen *tarkkuus* kuvaa hakutuloksen osumien lukumäärän suhdetta kaikkien löydettyjen dokumenttien lukumäärään, ts. *tarkkuus* = **a** / (**a** + **b**). Tarkkuus esitetään tavallisesti myös desimaalilukuna välillä [0, 1] tai prosenttilukuna välillä 0 ... 100 %. Tarkkuus kuvaa sitä, kuinka suuri osuus hakutuloksesta koostui relevanteista dokumenteista.

Jos Pekan saunanrakennushaussa otetaan kaikki 5 löytynyttä relevanttia WWW-dokumenttia huomioon, on hänen hakutuloksensa tarkkuus 20 selatun dokumenttilinkin kohdalla 5/20 eli 25%. Jos hyväksytään vain erityisen hyödylliset kaksi dokumenttia (jota sisälsivät saunojen rakennuspiirustuksia), saadaan tarkkuudeksi 2/20 eli 10%.

Hakutuloksen saanti kuvaa hakutuloksen osumien lukumäärän suhdetta kaikkien relevanttien dokumenttien lukumäärään, ts. saanti = a / (a + c). Saanti esitetään yleensä desimaalilukuna välillä [0, 1] tai prosenttilukuna välillä 0 ... 100 %. Saanti kuvaa sitä, kuinka suuri osa tietokannan sisältämistä relevanteista dokumenteista kyselyllä onnistuttiin löytämään.

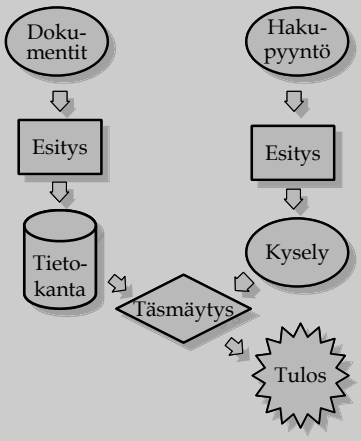
Pekan saunanrakennushaun saantia Googlen tietokannan suhteen ei voida laskea, ellei tavalla tai toisella selvitetä tietokannan kaikkien relevanttien dokumenttien lukumäärää. Käytännön tiedonhaussa tämä on lähes mahdoton tehtävä. Esimerkissämme hakukone ilmoitti löytäneensä kaikkiaan lähes kolme miljoonaa “sauna”-sanan sisältävää englanninkielistä dokumenttia. Tuon linkkijoukon selaaminen vaatisi arviolta yli kolmen vuoden yhtämittaisen työskentelyn, jos oletetaan, että yhden linkin tarkistus veisi 30 sekuntia ja työtä tehdään 150 t/kk. Ja senkin jälkeen jokin relevantti dokumentti voi ainakin periaatteessa jäädä löytymättä. Esimerkiksi dokumentin kielen tunnistava ohjelma on voinut tehdä virheen tai “sauna”-sana ei ole esiintynyt jossakin relevantissa dokumentissa (esimerkiksi saunavalmistaja on käyttänyt pelkkää tuotenimeä “Finnsauna”) tai keräilyrobotti ei ole hakenut joitakin relevantteja sivuja.

Käytännössä saannin arvioinnissa joudutaankin useimmiten tarkastelemaan hakutuloksen *suhteellista saantia* suhteessa joidenkin muiden hakujen tuloksiin tai muilla keinoin selvitettyyn tunnettujen relevanttien dokumenttien joukkoon, *saantikantaan.* Esimerkiksi voidaan tehdä useita erilaisia kyselyjä saman hakutehtävän perusteella, myös eri hakukoneilla, tulostaa vakiomäärä kunkin hakutuloksen kärkidokumentteja (esimerkiksi 20, 50, 100 tai 200 ensimmäistä), tunnistaa relevantit dokumentit näistä ja näin koota saantikanta. Yksittäistä hakua voidaan tällöin arvioida suhteessa kaikkien hakujen yhdistettyyn tulokseen. Saantikannan luontia valottaa muun muassa Sormunen (2000).

Pekan tapauksessa emme tiedä, kuinka monta linkkiä saunanrakennusta käsitteleviin dokumentteihin Googlen tietokanta sisältää. Jos niitä eri menetelmin löytyisi kaikkiaan 50, olisi Pekan saaman hakutuloksen saanti 5/50 eli 10 % 20 ensimmäisen tuloksen osalta. Googlessa kyselyn ja dokumenttien vertailu perustuu kyselyn ja dokumenttien *osittaistäsmäytykseen* ja hakutulosten *relevanssilajitteluun*. Tämän ansiosta relevantit dokumentit usein keskittyvät hakutuloksen kärkeen. Perinteiset, ns. Boolen hakujärjestelmät perustuvat kyselyn ja dokumenttien *täydelliseen täsmäytykseen* ja relevantit dokumentit hajaantuvat satunnaisesti hakutuloksessa (ks. Tekstilaatikko Täsmäytysmenetelmät ja relevanssilajittelu).

Täsmäytysmenetelmät, kyselyt ja relevanssilajittelu

Tiedonhakujärjestelmän ytimen muodostaa täsmäytysalgoritmi, joka on formaali menetelmä laskea kyselyn ja dokumentin esityksen (*representation*) samankaltaisuus. Samankaltaisuuden perusteella hakujärjestelmä päättää, kuuluuko dokumentti haun tulosjoukkoon vai ei ja mahdollisesti myös dokumentin järjestysnumeron tulosjoukon sisällä. Täsmäytysmenetelmä vertailee tietokantaan tallennettuja dokumenttien esityksiä kyselyyn oheisen kuvan mukaisesti (ks. Baeza-Yates & Ribeiro-Neto, 1999). Dokumenttien vaihtoehtoisia esitystapoja tietokannassa käsitellään kohdassa "Intellektuaalinen ja automaattinen kuvailu".



Täydellinen täsmäytys

Perinteiset Boolen logiikkaan perustuvat täsmäytysmenetelmät edellyttävät, että dokumentti täyttää kyselyn asettamat loogiset ehdot: dokumentin tulee sisältää joku kyselyn sallimista hakuavainten kombinaatioista. Jos kysely on “sauna JA (suunnittelu TAI rakentaminen TAI piirustukset)”, niin täsmäävissä dokumenteissa tulee esiintyä sana “sauna” ja ainakin yksi sanoista “suunnittelu”, “rakentaminen” ja “piirustukset”. Tulosjoukkoon tulevat kaikki loogisen ehdon täyttävät dokumentit eikä tulosjoukolla ole kyselyn ja dokumentin samankaltaisuuteen perustuvaa järjestystä, relevanssijärjestystä. Tulosjoukko esitetään esimerkiksi tallennuspäivän mukaan lajiteltuna (tuoreimmat dokumentit ensin).

Osittaistäsmäytys

Monet WWW:n sanahakupalvelut käyttävät osittaistäsmäytystä, jossa kyselyyn täsmäävän dokumentin ei välttämättä edellytetä sisältävän tiettyä hakuavainten kombinaatiota. Yksinkertaisimmat kyselyt ovat vain hakuavainlistoja, esimerkiksi “(sauna suunnittelu rakentaminen piirustukset)”. Dokumenttien indeksoinnissa niiden sisältämille sanoille annetaan dokumenttikohtaiset painot, jotka kuvaavat sanan tilastollista edustavuutta dokumentin kuvaamisessa. Vastaavasti voidaan kyselyn sisältämät hakuavaimet painottaa. Täsmäytys tapahtuu laskemalla kyselyn ja dokumentin yhteisten sanojen painojen perusteella dokumentille vertailuluku. Hakutulokseen otetaan hakijan valitsema määrä vertailuluvun perusteella parhaita dokumentteja. Dokumenttien indeksoinnissa ja vertailulukujen laskennassa voidaan käyttää monia tilastomatemaattisia menetelmiä, joiden tarkastelu jää tämän artikkelin ulkopuolelle.

Linkkirakenteen hyödyntäminen

WWW:n sanahakupalvelut käyttävät myös verkkodokumenttien keskinäistä linkkirakennetta hakutulosten järjestämiseen. Sivun tärkeyttä ja asemaa linkkirakenteessa voidaan arvioida sen saaminen ja antamien linkkien määrän perusteella. Tästä saatava painoarvo voidaan yhdistää hakuavainten tuottamaan sivun painoon sivujen lopullisen esitysjärjestyksen ratkaisemiseksi. Ks. myös alaluku Web-tiedonhaku.

Relevanssilajittelu

Relevanssilajittelussa hakutulos lajitellaan dokumenttien vertailulukujen perustella laskevaan järjestykseen. Näin kyselyä parhaiten vastaavat dokumentit tulevat hakutuloksen ensimmäisiksi. Usein nämä dokumentit ovat myös keskimäärin relevantimpia kuin tuloksessa myöhemmin esiintyvät.

Käytännön tiedonhakujärjestelmissä täsmäytysperiaatteet eivät esiinny puhtaina. Esimerkiksi Google hakee verkkosivuja ikään kuin hakuavainten välillä olisi ’AND’ – operaattori, ellei hakija toisin määrää. Hakutulos kuitenkin lajitellaan hakuavainten painojen ja linkkirakenteen perusteella. Relevanssilajittelun yksityiskohdat ovat viime kädessä kaupallisia salaisuuksia.

Pekan hakutuloksen saanti oli 10% ja tarkkuus 25% 20 löydetyn dokumentin jälkeen. Tulos voi tyydyttää Pekkaa, mutta parempikin hakutulos on usein mahdollista saada aikaan. Maksimisuoritus olisi sekä saannin että tarkkuuden osalta 100%, jos kaikki 50 relevanttia linkkiä olisivat sijoittuneet hakutuloksessa 50 ensimmäisen linkin joukkoon. Maksimisuoritukseen pääseminen on mahdollista vain harvoin.

Käytännössä tiedonhaun tutkimuksessa käytetään hakumenetelmien erojen havainnollistamiseen saanti-tarkkuuskäyrää, mutta varsinaiseen mittaamiseen saannin ja tarkkuuden sijasta johdettuja mittareita, joilla pyritään kuvaamaan hakutulosten onnistuneisuutta ottaen huomioon paitsi relevanttien dokumenttien löytyminen myös niiden sijainti hakutuloksessa. Tavallinen mittari on kyselyn keskitarkkuus (average precision), jonka laskentaa havainnollistetaan kuviossa 3.

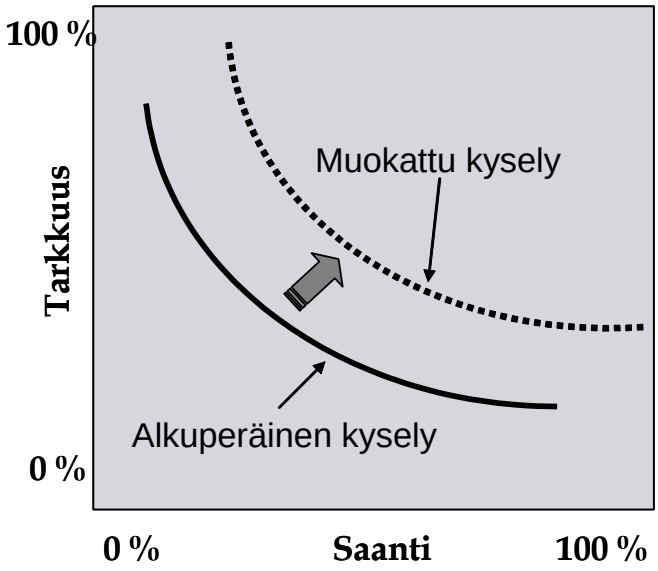
Linkin no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevanssi	-	+	+	-	-	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-
		1/2	2/3				3/7	4/8		$\frac{5}{10}$										

Kuvio 3. Keskitarkkuuden laskentaesimerkki

Keskitarkkuuden laskemiseksi selvitetään hakutuloksen tarkkuus jokaisen relevantin dokumentin kohdalla. Kuviossa 3 ensimmäinen relevantti dokumentti on järjestyksessä toinen, joten tarkkuus tässä kohdassa on 1/2. Seuraava löytyy sijalta 3, joten tarkkuudeksi saadaan 2/3. Keskitarkkuus lasketaan näiden tarkkuuslukujen keskiarvona. Jos kyselyn kaikki tunnetut relevantit vastaukset ovat kuvaan 4 merkityt viisi dokumenttia, saadaan ko. kyselyn keskitarkkuudeksi ((1/2) + (2/3) + (3/7) + (4/8) + (5/10)) / 5 = 0.52. Kun keskitarkkuuden keskiarvo lasketaan hakujärjestelmän evaluoinnissa käytettävästä hakutehtäväjoukosta muodostetuille kyselyille, saadaan kaikkein tavallisin hakujärjestelmien evaluointimittari MAP (Mean Average Precision). Sen laskennassa on tapana antaa löytymättä jääneille saantikantaan kuuluville dokumenteille tarkkuusluvuksi 0. Jos kuvion 3 esimerkissä tunnettuja relevantteja dokumentteja olisikin 50, joista vain viisi löytyi tarkasteltavaan tulokseen, saataisiin ko. hakutehtävän keskitarkkuudeksi ((1/2) + (2/3) + (3/7) + (4/8) + (5/10) + 0 + 0 + 0 + 0 + 0 + 0 +...) / 50 = 0.052.

Saannin ja tarkkuuden suhde on osoittautunut käänteiseksi: saannin parantaminen johtaa yleensä tarkkuuden huononemiseen ja päinvastoin (ks. kuvio 4). Täydellisen 100

%:n saannin saavuttaminen “normaalissa” hakutehtävässä edellyttää hyvin pitkän tuloslistan läpikäyntiä. Tällöin tarkkuus tavallisesti putoaa hyvin lähelle nollaa. Kyselyn muuntelu tai kehittyneemmän täsmäytysmenetelmän soveltaminen voi nostaa tuloksellisuuskäyrää korkeammalle tasolle, mutta suhteen käänteisyys säilyy. Maksimitulosta ei kaikissa hakutehtävissä voida saavuttaa.



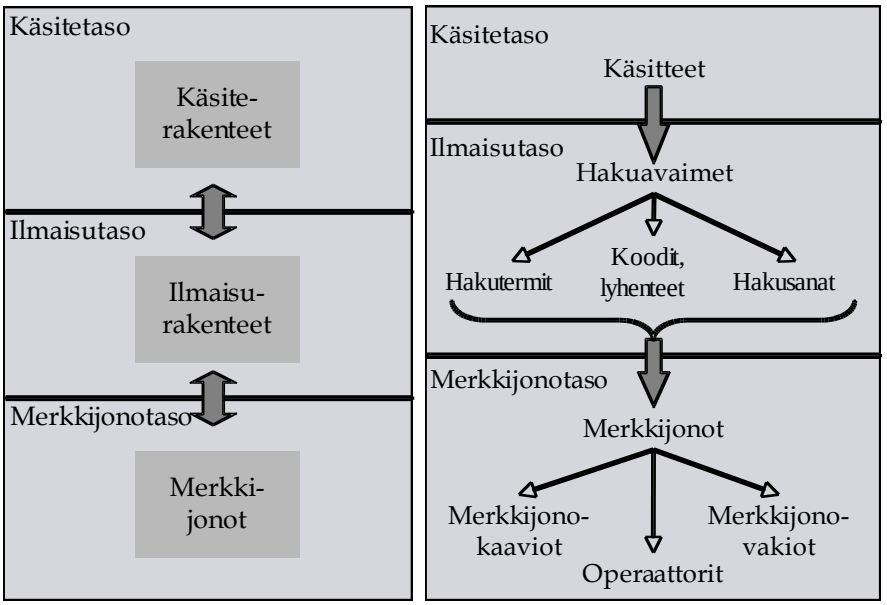
Kuvio 4. Hakutuloksen arviointi saannin ja tarkkuuden suhteen

Saannin ja tarkkuuden käänteinen suhde ja ideaalituloksen saavuttamisen vaikeus liittyy täsmäytysmenetelmien ja luonnollisen kielen ominaisuuksiin. Pelkistäen voidaan sanoa, että täsmäytysmenetelmät eivät pysty erottamaan tekstin sisältämiä merkityksiä. Täsmäytyksessä verrataan vain merkkijonojen (“sanojen”) ja näiden yhdistelmien esiintymiä dokumenteissa ja kyselyissä. Esimerkiksi hakusanoja *sauna*, *building* käsitellään Googlessa ikään kuin hakija kirjoittaisi kyselyn muotoon *sauna AND building*, siten edellyttäen molempien hakusanojen esiintymistä vastausdokumenteissa. Tämä saavutetaan, mutta kysely täsmää kuitenkin sekä tekstiin *“The sauna of the hotel is in the Main Building…”* (hotellin esittelyteksti) että tekstiin *“This document contains detailed guidelines for building a sauna…”* (saunanrakennusohje).

Tasoperiaate ja luonnollinen kieli

Tasoperiaate

Dokumentteja ja tiedontarpeita voidaan tarkastella kolmella tasolla (kuvio 5). Ensinnäkin dokumentit koostuvat *merkkijonoista.* Kirjoitusmerkkien ohella niissä on voi olla multimedia-aineistoa. Kirjoitusmerkeillä esitetään *luonnollisen kielen ilmaisuja.* Luonnollisen kielen ilmaisut puolestaan edustavat dokumentin *käsitteellistä sisältöä.* Vastaavasti tiedontarpeella on käsitteellinen sisältö, joka voidaan ilmaista luonnollisella kielellä ja esittää kirjoitusmerkkien avulla. Dokumentteja ja tiedontarpeita voidaan siten tarkastella kolmella tasolla: käsite-, ilmaisu- ja merkkijonotaso. Nämä tasot ovat periaatteessa aina mukana, kun ihmiset tarkastelevat dokumentteja ja kyselyjä tiedontarpeiden esityksinä. Näin on siitä huolimatta, ettei käsitteellistä sisältöä tai vaihtoehtoisia ilmaisuja tiedonhakutilanteissa aina tietoisesti pohditakaan. (Ingwersen & Järvelin, 2005).



Kuvio 5. Tasoperiaate

Kun dokumentti laaditaan, sen tekijällä on mielessään jokin käsitteellinen sisältö, jonka hän haluaa välittää lukijoille. Tämän käsitteellisen sisällön hän voi jäsentää monella eri tavalla ja monista näkökulmista. Jotta käsitteellinen sisältö olisi välitettävissä lukijoille, tekijä ilmaisee sen luonnollisen kielen (ja multimedian) avulla, koska käsitteellistä sisältöä ei voida suoraan välittää toisille. Tekijä voi ilmaista ajatuksensa monella tavalla, mutta hän valinnee sen, jonka odottaa olevan lukijoilleen sopivimman. Ilmaisut

hän pukee kirjoitusmerkeiksi (ja multimedia-esityksiksi). Vastaavasti lukijat tulkitsevat dokumentin kirjoitusmerkit sanoiksi ja sanat käsitteelliseksi sisällöksi omalta kannaltaan. Lukuisat arkielämän tilanteet osoittavat, ettei tulkinta aina ole tuottajan tarkoittama. (Ingwersen & Järvelin, 2005).

Jos tiedonhakija haluaa löytää useita tiedontarvettaan vastaavia dokumentteja, hänen tulee ottaa huomioon:

- eri kirjoittajien mahdolliset erilaiset käsitteelliset näkökulmat mahdollisissa relevanteissa dokumenteissa
- eri kirjoittajien mahdolliset erilaiset ilmaisut käyttämilleen käsitteille sekä
- tiedonhakujärjestelmien mahdolliset erilaiset tavat käsitellä ilmaisuja merkkijonoina.

Toisaalta hänen pitäisi ottaa huomioon tiedontarpeensa kaikki olennaiset piirteet ja niitä jäsentävät käsitteet mahdollisine ilmaisuineen. Pekan tapauksessa kyse on saunan rakentamisesta, joten tiedontarvetta jäsentää ainakin kaksi käsitettä, *[sauna]* ja *[rakentaminen]*. Pelkästään toisen käyttäminen johtaa helposti sellaisten epärelevanttien dokumenttien löytämiseen, joista puuttuu jokin keskeinen tiedontarpeen näkökulma. Kun Pekka haki saunanrakennusdokumentteja pelkästään hakuavaimella “sauna”, hän jätti rakentamiskäsitteen kokonaan ilmaisematta. Tämä näkyy hakutuloksessa dokumentteina, joissa puhutaan saunoista, muttei niiden rakentamisesta, esimerkiksi saunojen käytöstä. Pekan pitäisi myös harkita, mitä ilmaisuja hän voisi käyttää käsitteistä sauna ja rakentaminen.

Haun suorittamiseksi tiedonhakija joutuu esittämään hakutehtävänsä kyselynä, joka laaditaan tiedonhakujärjestelmän kyselykielen avulla. Pekka käytti Googlea. Koska dokumenttien yksittäiset sanat, kuten “sauna”, voivat Googlessa toimia kyselyinä, Pekan haku voitiin suorittaa. Jos hän olisi halunnut esittää kyselyssään myös käsitteen rakentaminen (englanniksi), hänen kyselynsä olisi ollut esimerkiksi *sauna, building*. Myöhemmin tarkastelemme monimutkaisempia kyselykielen ilmaisuja.

Tietokoneisiin perustuva tiedonhaku tapahtuu käytännössä aina merkkijonotasolla ja siksi tämä taso on aina läsnä tiedonhaussa. Ilmaisu- ja merkkijonotasojen välistä eroa voidaan havainnollistaa vaikka merkkijonoilla “kirjaa”, “kirjan” ja “kirjain”. Automaattisen käsittelyn kannalta kyseessä on kolme merkkijonoa, joissa viisi ensimmäistä kirjainta ovat samoja. Lukija tulkinnee kaksi ensimmäistä saman sanan

(ilmaisun) “kirja” esiintymiksi (taivutusmuodoiksi) ja viimeisen kokonaan eri käsitteeseen viittaavaksi sanaksi “kirjain”. Lukija pystyy hahmottamaan, mikä käsitteellisen tason yhteys sanoilla “kirja” ja “kirjain” on.

Tietokoneet käsittelevät vain merkkijonoja, vaikka ne voidaankin ohjelmoida toimimaan ikään kuin ne rajoitetussa mielessä ymmärtäisivät luonnollista kieltä. Kielentulkintaohjelman ja sanaston avulla tietokone pystyisi tunnistamaan “kirjaa” ja “kirjan” merkkijonot “kirja”-sanan esiintymiksi ja “kirjaimet”-merkkijonon “kirjain”-sanan esiintymäksi. Mutta arkitajuntaa edellyttävä tulkinta - *kirjassa* on lehtiä, joille voi olla painettuna *kirjaimia* - jää tietokoneelta tavoittamatta. Tiedonhakujärjestelmä kiinnittää huomiota vain merkkijonojen (parhaassa tapauksessa niistä laskennallisesti havaittujen ilmausten) esiintymiin kyselyissä ja dokumenteissa, etenkin esiintymien keskinäiseen sijaintiin ja lukumääriin tai muihin tilastollisiin tai loogisiin ominaisuuksiin.

Luonnollisen kielen ominaisuuksia

Luonnollinen kieli on alati kehittyvä ja muuttuva sosiaalinen luomus. Se on hyvin vaihtelevaa, ilmaisurikasta ja joustavaa. Luonnollinen kieli on vain osittain yhteistä saman kulttuurin jäsenille. Yhteiskunnassa esiintyy useita erilaisia kielen alakulttuureja (kielipelejä tai diskursseja). Eri osakulttuurit ilmenevät vaihteluna niin sanastossa, lauserakenteissa kuin myös taustalla olevissa käsitteissä (ks. tekstilaatikko Luonnollinen kieli tiedonhaun ongelmana). (Järvelin, 1995). Nämä erot heijastuvat sekä eri osakulttuureissa tuotettujen dokumenttien sisältöön että tapoihin ilmaista tiedontarpeita.

Luonnollisen kielen ominaisuuksien vaikutus tiedonhakuun on moninainen. Ne pitää ottaa huomioon niin dokumenttien tallennuksessa, kyselyjen muotoilussa kuin myös tulosten arvioinnissa. Suomen kielessä erityisiä ongelmia ovat (Alkula, 2000):

- Sanojen ja niiden vartaloiden taipuminen
- Yhdyssanat ja sanaliitot
- Monitulkintaisuus, erityisesti taivutusmuotohomografia ("hauissa", "puhelin")
- Johdokset

Luonnollinen kieli tiedonhaun ongelmana

Monitulkintaisuus. Monitulkintaisuus on kielen perusominaisuuksia ja sallii suuren ilmaisujoukon pienemmällä määrällä alkioita. Sitä esiintyy niin sana- kuin lausetasollakin. Esimerkiksi sanat puhe, kangas, kuusi, selkä ovat homografeja, joiden merkitys riippuu asiayhteydestä. Jokainen lukija tuo tulkintaan oman panoksensa ja näin jokainen lukukerta tuo uuden tulkinnan.

Homonymiassa kahden eri sanan kirjoitus- ja äänneasut ovat identtiset. Sen alalaji on homografia, jossa vain kirjoitusasut ovat identtiset. Esimerkiksi hauissa on ääntämisestä riippuen sanan haku tai sanan hauki taivutusmuoto. Polysemiassa samalla sanalla on kaksi tai useampia lähekkäisiä merkitysvariantteja (esim. selkä).

Synonymia. Sama käsite voidaan ilmaista monella eri sanalla. Synonymiasuhde ei aina ole suinkaan täsmällinen. Sanojen edustamat käsitteet voivat myös muodostaa perheyhtäläisen ryhmän: ne nivoutuvat yhtenäiseksi punokseksi, mutta punoksen päät ovat merkitykseltään jo varsin etäisiä.

Yhdyssanat ja sanaliitot. Yhdyssanoissa usein loppuosa ilmoittaa pääluokan ja alkuosa(t) alaluokan. Loppuosan tunnistaminen on usein tärkeää. Kielenkäyttö myös vaihtelee yhdyssanojen ja sanaliittojen kesken, esimerkiksi molempia ilmaisuja ”tiedon tarve” ja ”tiedontarve” käytetään.

Etu- ja loppuliitteet. Luonnollisissa kielissä on monia erilaisia etu- ja loppuliitteitä sanojen merkityksen muokkaamista varten. Niiden käyttö ja tärkeys eri kielissä vaihtelee. Esimerkkejä: esi-, jälki-, ylä-, ala-, etu-, taka-, epä-, ...; -inen, -lainen, ...

Sanojen taipuminen. Useimmissa kielissä yksikkö- ja monikkomuodot eroavat toisistaan. Lisäksi sanat saattavat kuulua eri sukuihin ja niiden taipuminen on tällöin suvun mukaista. Sijamuodot ilmaistaan joko taivutuspäätteillä tai pre- tai postpositioiden avulla.

Johdokset. Sanoista saadaan johtimien avulla uusia sanoja. Esimerkiksi sanasta kirja saadaan yli 60 erilaista johdosta, joihin kuuluvat: kirjo, kirjoa, kirjonta, kirjominen, kirjain, kirjasto, kirjaimisto, ... Johtimet poistamalla selviää johdoksen kantasana.

Kyselyjen muotoilu

Kyselyssä esitetään tiedontarve tiedonhakujärjestelmän kyselykielellä. Käsitteiden ja ilmaisujen pohdinta johtaa valittaviin hakuavaimiin, kyselyn sisältöön. Kyselyn avulla käyttäjä määrittelee hakujärjestelmälle minkälaisia merkkijonoja (ilmaisuja) etsittävissä dokumenteissa pitäisi esiintyä (Pekalla ”sauna”) tai minkälaisia ominaisuuksia dokumenteilla tulisi olla (Pekka rajasi englanninkieliseen aineistoon). Kysely oli pelkistetty ja tulos jäi vaatimattomaksi. Pekka ei kyselyssä viitannut lainkaan [*rakentaminen*] -käsitteeseen, joten hakujärjestelmällä ei ollut perustetta torjua

rakentamiseen liittymättömiä sauna-dokumentteja. Tämä selittää osan hakutuloksen hälystä.

Perinteisissä täydellisen täsmäytyksen hakujärjestelmissä käytetään kyselyissä Boolen algebrassa määriteltyjä operaattoreita. Operaattorit luovat kyselylle rakenteen. Rajaavia käsitteitä edustavat hakuavainryhmät yhdistellään AND- operaattorilla (konjunktio) ja kutakin käsitettä edustavat vaihtoehtoiset hakuavaimet OR-operaattorilla (disjunktio). Käsitteellinen hakusuunnitelma *[sauna] AND [rakentaminen] AND [englannin kieli]* voisi muuntua kyselyksi *”sauna AND (building OR construction OR drawings) AND kieli=English”*. Kysely määrittelee poimittavaksi englanninkieliset dokumentit, jotka sisältävät merkkijonon *”sauna”* sekä ainakin yhden sanoista ”*building*”, ”*construction*” tai ”*drawings*”. Hakijan kannalta kyselyn tulkinta on yksiselitteinen siinä mielessä, että jokaisesta dokumentista voidaan sanoa täyttääkö se kyselyn määrittelemän ehdon vai ei.

Kuten aiemmin todettiin (ks. täsmäytysmenetelmät –laatikko), osittais-täsmäyttävissä hakujärjestelmissä tavallisin kyselyn muoto on sanalista, joka ei sisällä operaattoreita. Jokainen kyselyn hakuavain antaa hakujärjestelmälle evidenssiä käyttäjän määrittelemästä tiedontarpeesta, jota se tulkitsee täsmäytysalgoritmin määrittelemällä tavalla. Kun täsmäytys on monimutkainen tilastollinen prosessi, hakijan on vaikea tietää miten hakusanalistaan tehtävät muutokset vaikuttavat hakutulokseen. Googlen tyyppisissä netin hakukoneissa täsmäytysalgoritmi huomioi lisäksi dokumenttien välisten linkkirakenteet, mikä edelleen vähentää hakijan mahdollisuuksia kontrolloida tulosjoukon ja sen järjestyksen muotoutumista. Hakijan kontrollointimahdollisuuksien vähentymisen vastapainoksi osittaistäsmäytys tuottaa yleensä täydellistä täsmäytystä parempia hakutuloksia.

Nykyaikaisissa osittaistäsmäyttävissä hakujärjestelmissä on mahdollista käyttää operaattoreita. Operaattorien käyttö kyselyn rakenteistamiseen parantaa hakutuloksia muun muassa kyselyjä laajennettaessa (Kekäläinen 1999) tai tehtäessä sanakirjakäännökseen perustuvaa kieltenvälistä tiedonhakua (Pirkola 1999). Googlekin tukee kyselyn rakenteistamista, mutta operaattorit on kätketty kyselylomakkeen taakse. Kuvassa 2 on esitetty Pekan uusi haku, joka on muotoa *”sauna AND (construction OR drawings) AND kieli=English”*. Hakusanojen syöttökenttien <u>välillä</u> sovelletaan AND-operaattoria ja kentän ”Mikä tahansa sanoista” <u>sisällä</u> oleviin sanoihin OR-operaattoria.

Kuva 2. Kyselyn rakenteistaminen Googlen tarkennetun haun lomakkeella.

Pekka huomioi täsmennetyssä kyselyssä kolme tärkeää kyselyn muodostuksen periaatetta, joiden soveltamista kannattaa harkita haastavien hakutehtävien yhteydessä:

- Huomioi kaikki olennaiset tiedontarvetta rajaavat käsitteet. Pekka lisäsi kysely käsitteen *[rakentaminen]* rajaamaan sauna-aineistoa.
- Huomioi olennaiset käsitettä edustavat vaihtoehtoiset ilmaisut. Käsitettä *[rakentaminen]* edustamaan Pekka valitsi hakuavaimet ”*construction”* ja *”drawings”*. Hän jätti käyttämättä sanan ”*building*” sillä se näytti tuottavan enemmän hälyä kuin osumia.
- Luo kyselylle tarvittaessa rakenne operaattoreita soveltamalla. Pekka käytti hyväksi Googlen tarkennetun haun lomakkeen ominaisuuksia. Vaihtoehtoiset hakuavaimet ”*construction”* ja *”drawings”* kytkettiin OR-operaattorille ja nämä elelleen hakuavaimiin *”sauna” ja ”kieli=English”*AND-operaattorilla. Tällä kertaa hakujärjestelmälle ei tarvinnut esittää fraaseja, johon kenttä ”täsmällisellä ilmauksella” antaa mahdollisuuden.

Tarkasteltuaan tulosdokumentteja Pekka havaitsi dokumentit 1, 2, 4, 8, 10 - 14 sekä 19 erittäin tai melko relevanteiksi. Loput 10 linkkiä johtivat hyödyttömiin tai arvoltaan epäselviin dokumentteihin. Tuloksen yhteenveto voidaan havainnollistaa kuvion 6 tapaan. Nyt 20 löydetyn kärkidokumentin joukossa tuloksen tarkkuus oli 50% eli kaksinkertaistui. Saanti tuplaantui myös eli nousi 20 prosenttiin. Varsin yksinkertaisin toimenpitein tulos parani olennaisesti, etenkin ensimmäisten löydettyjen dokumenttien osalta.

Linkin no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevanssi	+	+	-	+	-	-	-	+	-	+	+	+	+	+	-	-	-	-	+	-

Kuvio 6. Pekan uuden haun relevanssitarkastelu

Intellektuaalinen ja automaattinen kuvailu

Web, metadata ja ontologiat

Dokumenttien kuvailutietoja tarkoittava *metadata*-termi (tai *metatieto*) tuli käyttöön web-boomin alkuvaiheessa 1990-luvulla. Metadata on tietoa tiedosta, eli kuvailevaa ja määrittävää tietoa jostakin dokumentista tai tietovarannosta. Hakukoneet toimivat Webin alkutaipaleella erittäin huonosti. Kirjoittajien tai hakupalvelujen tuottajien ajateltiin auttavan ongelman ratkaisussa lisäämällä dokumentteihinsa metadataa turhan hälyä karsimiseksi hakutuloksista.

Tietovarantojen kuvaamista varten kehitettiin Dublin Core (DC) metadataformaatti, josta on tullut laajalti sovellettu väline tietoresurssien kuvailuun. Kuvailuformaatti standardoi metadatan esittämisen muodon niin, että tuotetut kuvaukset ovat tietokoneella käsiteltävässä muodossa. DC määrittelee viisitoista perustietotyyppiä (elementtiä), joiden avulla tietyn aineiston kuvailusovellus voidaan suunnitella. Kaikki elementit ovat tallennettaessa vapaaehtoisia eli kuvailuun voidaan valita joustavasti halutut elementit. Elementtien luettelo havainnollistaa minkä tyyppisiin dokumentin ominaisuuksiin voidaan kiinnittää huomiota: 1) Nimeke (Title), 2) Tekijä (Creator), 3) Aihe (Subject), 4) Kuvaus (Description), 5) Julkaisija (Publisher), 6) Muu tekijä (Contributor), 7) Aikamääre (Date), 8) Laji (Type), 9) Formaatti (Format), 10) Tunniste (Identifier), 11) Lähde (Source), 12) Kieli (Language), 13) Suhde (Relation), 14) Kate (Coverage) ja 15) Oikeudet (Rights).

Elementtien lisäksi voidaan käyttää erilaisia tarkenteita, joiden avulla elementin semanttista merkitystä pystytään tarkentamaan. Esimerkiksi verkossa julkaistun kirjan artikkeliin (HTML-sivu) voi liittyä kuvaus, jossa tekijä ilmaisee oman ja artikkelinsa nimen sekä kirjan nimen:

```
<META NAME="DC.Creator.PersonalName" CONTENT="Halttunen, Kai">
<META NAME="DC.Title" CONTENT="Tiedonhankinnan ohjaamisesta">
< META NAME="DC.Relation.isPartOf" CONTENT="Vahva esitys! Vinkkejä ja
malleja esseiden ja ryhmätehtävien ohjaamiseen”>
```

NAME kertoo mistä Dublin Coren elementistä on kyse ja CONTENT ilmaisee elementin saaman arvon eli sisällön. DC-koodattu kuvaus määrittää kuvailun tarkan semantiikan: Kai Halttunen on (henkilö)tekijä nimettyyn dokumenttiin, joka on edelleen osa laajempaa nimettyä dokumenttia.

Metadata on kahta tyyppiä. *Sisäinen (embedded) metadata* tallennetaan kuvailtavaan dokumenttiin. Yllä oleva esimerkki oli sisäistä metadataa. Myös dokumentin rakenteen kuvaus jotain merkkauskieltä kuten HTML:ää tai XML:ää käyttäen on sisäistä metadataa. Sitä ovat myös linkit, jotka viittaavat kohteeseen dokumentin sisällä tai sen ulkopuolella (ilmaisevat suhteen kahden kohteen välillä). *Ulkoinen (external) metadata* tallennetaan muualle kuin kuvailtaviin dokumentteihin. Esimerkiksi kirjastojen luettelotietokannat tai Yahoon hakemistot sisältävät ulkoista metadataa.

Viime vuosien näkyvä kehittämiskohde on ollut uuden sukupolven "älykäs" informaatioympäristö - *semanttinen web*. Semanttisen webin keskeisiä tavoitteita on kehittää metakieliä ja -standardeja sekä ontologiatekniikoita, joiden avulla verkon resurssien sisällöllinen kuvaaminen ja hyödyntäminen käyvät mahdolliseksi. Eri tietojärjestelmät pystyisivät aiempaa tehokkaammin hyödyntämään toistensa tietoja ja näin tuottamaan käyttäjän kannalta laadukkaampia palveluja.

Semanttisen webin kehittäminen rakentuu useiden dokumenttien ja informaation esittämistapaa yhtenäistävien standardien varaan. XML (eXtended Mark-up Language) on HTML:ää kehittyneempi merkkauskieli dokumenttien rakenteen kuvaamiseen. Toinen tärkeä standardi on RDF (Resource Description Language), jonka avulla voidaan kuvata tietoresurssien metadatoja. Semanttiseen verkkoon liittyvät olennaisesti myös käsitemallit eli *ontologiat*, joiden esittämiseen on oma standardinsa OWL (Web Ontology Language). (Cardoso 2007.)

Semanttinen web edustaa formaalin tietojenkäsittelyn lähestymistapaa tietosisältöjen hallintaan. On esitetty epäilyjä, että tällaisen datan ekosysteemin rakentaminen on käytännössä mahdotonta vaaditun kuvailutyön määrän vuoksi. Kilpailevaksi kehityssuunnaksi on noussut sosiaalinen web, jossa formalismit korvautuvat pragmatismilla ja ontologiat folksonomioilla. Muun muassa Gruber (2008) näkee kehityslinjat toisiaan tukevina eikä kilpailevina. Aiheesta lisää alaluvussa ”Ontologinen tiedonhaku”.

Perinteinen luettelointi ja sisällönkuvailu

Metadata on nettiajan termi. Tietovarantojen kuvailua on kirjastojen ja tietopalveluorganisaatioiden piirissä tehty jo satoja vuosia yleiskäsitteiden *luettelointi* ja *sisällönkuvailu* alla. Luettelointi on perustunut standardoituihin käytäntöihin, *luettelointisääntöihin* ja *MARC-tallennusformaattiin* (vrt. Dublin Core). Sisällönkuvailussa on sovellettu ontologioiden kaltaisia *dokumentaatiokieliä*. Rajaamme seuraavassa tarkastelun pääasiassa sisällönkuvailun kysymyksiin, sillä ne ovat tiedonhaun kannalta olennaisimpia. Sisällön kuvailu ja haku on kuitenkin sidoksissa dokumentin rakenteeseen ja muihin ominaisuuksiin, joten jyrkkää erottelua ei kannata tehdä.

Hakupalvelun rakentamisessa pyritään hyödyntämään mahdollisimman paljon dokumenttien sisäistä metadataa, muun muassa HTML- tai XML-koodattua rakennetta (otsikot, kappaleet, linkit, ym.) sekä tuottamalla metadataa automaattisin menetelmin itse sisällöstä (teksti, kuvat, jne). Tiedon tallennusprosessin keskeisin kysymys on se, mitä menetelmiä käyttäen tietokannan hakemiston tiedot tuotetaan. Esimerkiksi Googlen tapauksessa sanahakupalvelun tietokannan hakemistoon tallennetaan dokumentissa esiintyneet merkkijonot, tunnistetut rakenne-elementit (esim. URL, linkit, ankkuritekstit) ja tekstianalyysillä tuotetut tiedot (esim. dokumentin kieli).

Tarkastelemme tässä lähemmin dokumentin sisältämän tekstin automaattista hyödyntämistä tietokannan hakemistoissa. Tällöin puhutaan *kokoteksti-indeksoinnista (full-text indexing)*, jonka ansiosta dokumentti on löydettävissä minkä tahansa siinä esiintyneen sanan perusteella. Perinteistä kokoteksti-indeksointia sovelletaan Boolen mallia edustavissa tiedonhakujärjestelmissä ja se suoritetaan seuraavasti: (1) Indeksointiohjelma tunnistaa dokumentista erilliset merkkijonot. (2) Kuhunkin merkkijonoon liitetään osoitetiedot, kuten dokumentin tunniste, ns. kenttätieto (esiintyikö otsikossa, tiivistelmässä, leipätekstissä, tekijänimenä, tms.), sekä sijainti kentän sisällä (esimerkiksi 3. merkkijono). (3) Merkkijonot osoitetietoineen aakkostetaan hakemistoksi eli *käänteistiedostoksi (inverted file).*

Perinteiset (Boolen mallia edustavat) hakujärjestelmät antavat mahdollisuuden kohdistaa tietokantaan kyselyjä ja asettaa hyvin monipuolisia hakusanojen esiintymiseen liittyviä ehtoja. Voidaan esimerkiksi määritellä, missä dokumentin osassa hakusanan on esiinnyttävä (kenttähaut) ja mikä on hakusanojen keskinäinen sijainti ja järjestys dokumentissa (läheisyysoperaattorihaut, fraasihaut). Perinteisen

kokoteksti-indeksoinnin heikkoutena on se, että kukin sana edustaa dokumenttia täsmälleen yhtä suurella painolla (1 - kun esiintyy; 0 - kun ei esiinny).

Painokerroin (kokoteksti-)indeksoinnin menetelmät lähtevät oletuksesta, että jotkut sanat ovat dokumentin sisällönkuvaajina tärkeämpiä kuin toiset. Painokertoimiin perustuva kokoteksti-indeksointi on kehitetty osana osittaistäsmäytysalgoritmien tutkimusta ja relevanssilajittelu perustuu painokertoimien soveltamiseen. Sanan tärkeyttä arvioidaan usein kahden piirteen – *termifrekvenssin (term frequency)* ja *käänteisen dokumenttifrekvenssin (inverted document frequency)* – perusteella. Termifrekvenssi ilmaisee yksinkertaisesti kuinka monta kertaa sana esiintyy dokumentissa. Käänteinen dokumenttifrekvenssi on verrannollinen sanan harvinaisuuteen koko tietokannassa. Dokumentissa *k* esiintyvälle sanalle *i* voidaan laskea *paino (weight)* w_i esimerkiksi kaavalla

$$\omega_{ik} = tf_{ik} \; x \; idf_i \text{ (= termifrekvenssi x käänteinen dokumenttifrekvenssi).}$$

Esimerkiksi sana *kirja* (= *i*) voisi esiintyä dokumentissa A kolme kertaa (tf_{iA}=3) ja dokumentissa B yhden kerran (tf_{iB}=1) ja koko tietokannassa kymmenessä eri dokumentissa (idf_i=0,1). Tällöin sanan *kirja* paino on dokumentissa A 0,3 ja dokumentissa B 0,1. Hakuvaiheessa täsmäytysalgoritmi laskee dokumenteille *vertailuluvut* (*score*) kyselyn eri sanojen dokumenteissa saamien painokerrointen perusteella ja lajittelee hakutuloksen vertailuluvun mukaiseen järjestykseen.

Automaattisten kuvailumenetelmien vahvuutena on niiden tyhjentävyys eli kaikki dokumentin yksityiskohdat ovat haettavissa. Kuvailu on nopea automaattinen prosessi, joten sen avulla voidaan käsitellä suuria dokumenttijoukkoja pienin kustannuksin. Ongelmana on mekaanisuus eli kuvailu perustuu dokumentin sisältämien merkkijonojen tunnistamiseen ja niiden merkittävyyden määrittelyyn tilastolliselta pohjalta. Tekstin merkityksiin liittyviä tulkintoja ei voida kuvailussa huomioida. Sen sijaan sanojen taipuminen ja muu tekstin ilmaisutason vaihtelu voidaan normalisoida. Vaihtoehtoina ovat sanamuotojen normalisointi sanavartaloiksi tai perusmuodoiksi (Kettunen 2007).

Intellektuaalinen (eli manuaalinen) sisällönkuvailu perustuu ihmisen tekemään dokumentin sisällön analyysiin ja kuvailutietojen valintaan. Se voi tapahtua joko *luonnollista kieltä* tai erityistä *dokumentaatiokieltä* käyttäen. Luonnollisella kielellä tapahtuvaa sisällönkuvailua kutsutaan *avainsanoitukseksi,* koska kuvailija poimii dokumentista tai omasta muististaan avainsanoja tai –fraaseja (keywords), jotka hänen

mielestään luonnehtivat dokumentin sisältöä. Avainsanoituksen edut ovat vähäiset, koska sitä rajoittavat samat luonnolliseen kieleen liittyvät monitulkintaisuusongelmat kuin automaattista indeksointia. Avainsanoitus rajoittuu lähinnä aineistoihin, jotka eivät sisällä elektronisessa muodossa olevaa tekstiä (mm. painetut julkaisut, audiovisuaaliset aineistot). *Tiivistelmien laatiminen* ja niiden automaattinen indeksointi on avainsanoitusta muistuttava mutta luontevampi tapa kuvailla dokumentteja. Tiivistelmien etuna on, että tiedonhakija pystyy paitsi tekemään hakuja tiivistelmän sanoilla myös arvioimaan niiden perusteella dokumenttien sisältöä.

Dokumentaatiokieli, kuten ontologiakin, pyrkii olemaan monitulkintaisuudesta vapaa, dokumenttien sisällönkuvailuun soveltuva erikoiskieli. *Luokituskaavat* ja *asiasanastot* ovat esimerkkejä dokumentaatiokielistä. Luokituskaavoja käytetään muun muassa kirjastoaineistojen kuvailuun. Niissä on usein tavoitteena universaalisuus eli pyrkimys kattaa kaikki aihealueet, joista dokumentteja on julkaistu. Luokituskaavoihin perustuvaa kuvailua kutsutaan *luokitukseksi*. Luokitustulos on tyypillisesti alfanumeerisia koodeja eli luokkatunnuksia. Esimerkiksi tietokonetekniikan kirjat sijoittuvat yleisessä kymmenluokituksessa (UDK) luokkaan ”681.3”. Automaattiseen luokitukseen liittyvää tutkimusta esitellään alaluvussa ”Dokumenttien automaattinen ryhmittely ja luokitus”.

Asiasanastot pyrkivät käyttämään luonnollisen kielen kaltaisia *asiasanoja*, joiden merkitys pyritään samaan yksiselitteiseksi dokumentaatiokielen idean mukaan. *Tesaurus* on asiasanasto, jossa asiasanojen suhteet on määritelty standardoidulla tavalla. Tesaurukset ovat yleensä aihealakohtaisia, mutta yleiskäyttöisiäkin on kehitetty kuten Yleinen Suomalainen Asiasanasto (YSA, ks. http://vesa.lib.helsinki.fi/ysa/). Asiasanastoon perustuvaa sisällönkuvailu kutsutaan *asiasanoitukseksi* tai *indeksoinniksi*.

Tesauruksessa on aakkosellinen ja systemaattinen osa. Aakkosellisessa osassa esitetään sanaston mikrorakenne kunkin asiasanan osalta – muun muassa hierarkiassa ylemmät ja alemmat asiasanat sekä muut sukulaisasiasanat. Systemaattinen osa muistuttaa luokitustaulua, koska se esittää kokonaisuutena asiasanaston hierarkkisen rakenteen. Tätä voi tukea koodijärjestelmä, jonka avulla on helppo suorittaa kyselyjä laajoillakin tesauruksen alihierarkioilla, esimerkiksi kaikilla tietokonetekniikkaan liittyvillä asiasanoilla. Tämä huomio paljastaakin, että luokitustaulujen ja asiasanastojen (erityisesti tesaurusten) välinen ero liittyy lähinnä ulkoasuun. Molemmat ovat parhaimmillaan tietyn tiedonalan käsitteistön systemaattisia, hierarkkisia

esityksiä. Tesaurusten ja ontologioiden pääasiallinen ero on siinä että ontologioissa on mahdollista esittää joustavammin, monipuolisemmin ja tarkemmin käsitteiden välisiä suhteita ja ne on esitetty automaattista päättelyä tukevalla tavalla.

Tiedonhakututkijat ovat vuosikymmeniä väitelleet automaattisten ja intellektuaalisten kuvailumenetelmien paremmuudesta. Johtopäätös tuosta väittelystä lienee se, että molemmille menetelmille löytyy järkeviä soveltamiskohteita:

1) Automaattisella kuvailulla pystytään tuottamaan hyvin tyhjentäviä, yksityiskohtiin pureutuvia kuvailuja nopeasti ja pienin kustannuksin. Automaattiset menetelmät ovat yleiskäyttöisiä ja niitä voidaan soveltaa alasta ja dokumenttityypeistä riippumatta. Rajoitukseksi muodostuu vaikeudet ratkaista luonnollisen kielen monitulkintaisuuden ja merkitysten tulkinnan ongelmat. Kokoteksti-indeksointiin perustuva automaattinen kuvailu on tuloksekkainta ja palvelee tiedonhakijaa muun muassa seuraavissa tilanteissa:
   a) Intellektuaaliseen kuvailuun ei ole riittäviä resursseja tai välineitä.
   b) Tesaurus ei sisällä tietyntyyppisiä asiasanoja (esimerkiksi nimihaut).
   c) Tesauruksen sanasto on liian yleisellä tasolla (spesifit kyselyt).
   d) Tesauruksen sanasto on vanhentunut (uusia ilmiöitä koskevat haut).
2) Intellektuaalisen kuvailun voima perustuu tesaurusten sisällön ja rakenteen räätälöintiin tietyn tiedonhakusovelluksen erityiset tarpeet huomioivana käsitemallina. Tiedonhakijoille tärkeät asiat voidaan tesauruksessa esittää heidän omalla terminologiallaan täysin riippumatta siitä, miten dokumenttien kirjoittajat asioita nimeävät. Asiasanoitus perustuu toisaalta kuvailijan kykyyn tulkita tekstin sisältämiä merkityksiä, tehdä ero olennaisen ja epäolennaisen välillä sekä heijastaa niitä tesauruksen tarjoamaan käsitemalliin. Tesaurusten rakentaminen ja intellektuaalinen kuvailu on kuitenkin kallista, joten menetelmien soveltamiseen on aina rajatut mahdollisuudet. Intellektuaalinen kuvailu parantaa hakujen tuloksellisuutta muun muassa seuraavissa tilanteissa:
   a) Dokumenteissa on luonnostaan vähän tekstiä (kuva-, äänite- tai videokokoelma).
   b) Haun aihe on laaja-alainen (aiheena esimerkiksi kokonainen tieteen tai tekniikan ala).
   c) Haetaan monikielisestä aineistosta (tieteellisen kirjaston luettelo).
   d) Dokumentin tärkeä aspekti ei ole tekstistä automaattisin menetelmin tunnistettavissa (tekstin luonne: tieteellinen – populaari).

Tiedonhaun tutkimus- ja sovellusalueita

Tekstitiedonhaku, jossa tekstinä esitettyä hakupyyntöä (kyselyä) verrataan tekstinä esitettyihin dokumentteihin, on tiedonhaun tutkimuksen perinteisin alue. Edellä oleva tiedonhaun esittely keskittyikin juuri tekstitiedonhaun periaatteisiin, vaikka Pekan saunanrakennukseen liittyvänä esimerkkinä olikin tiedonhaku netistä, jossa dokumentit usein ovat multimediaa. Useimmiten multimediatiedonhaussakin haku perustuu tekstin, siis kielellisten hakuavainten, käyttöön. Näin ei kuitenkaan aina ole. Seuraavassa tarkastelemme tiedonhaun sovellusalueita muutamien keskeisten esimerkkien valossa. Lähes kaikissa teksti on pääosassa hakuperusteena, mutta tuomme esiin myös muita mahdollisuuksia.

Kieltenvälinen tiedonhaku

Kieltenvälinen tiedonhaku (Cross-Lingual Information Retrieval) pyrkii kehittämään välineitä monikielisyyden hallintaan tiedonhakuprosesseissa. Tiedontarvitsija joutuu usein etsimään tietoa vieraskielisistä kokoelmista. Lisäksi monet dokumenttikokoelmat ovat monikielisiä. Esimerkiksi WWW:ssä on dokumentteja sadoilla eri kielillä. Hakijan kannalta olisi yksinkertaisinta, jos hänen kyselynsä käännettäisiin automaattisesti kaikille tietokannan kielille ja niihin täsmäävät dokumentit taas hakijan äidinkielelle. Ainakaan toistaiseksi konekäännösmenetelmät eivät kuitenkaan pysty kääntämään rajoittamatonta luonnollista kieltä kelvollisiksi dokumenteiksi. Kyselyjen kääntäminen on selvästi yksinkertaisempaa. Tiedonhakija saattaa lisäksi kyetä lukemaan dokumentteja useilla eri kielillä, vaikka hakutehtävän muotoilu onkin helpointa yhdellä kielellä, erityisesti äidinkielellä.

Kyselyjen kääntäminen voidaan suorittaa käännössanakirjojen, konekäännösohjelman tai rinnakkaiskokoelman avulla. *Käännössanakirjaa* käytettäessä hakutehtävän lähtökieliset sanat käännetään sähköisen sanakirjan tarjoamiksi kohdekielisiksi sanoiksi, joista muodostetaan kysely. Sanojen monitulkintaisuus lähtö- ja kohdekielissä, sekä käännösehdokkaiden suuri määrä muodostuvat haasteiksi. *Konekäännösohjelma* kääntää lauseita kieleltä toiselle ja sopii välineeksi, jos hakutehtävät ovat kokonaisia lauseita. Näin ei kuitenkaan aina ole. Lisäksi konekäännös ei tuota sanakirjojen tapaan hakuavainten synonyymejä kohdekielellä. *Rinnakkaiskokoelmaa* käytettäessä käännössanakirja korvataan kaksikielisellä kokoelmalla, joka on rinnastettu lause-, kappale tai dokumenttitasolla erikielisiksi vastinpareiksi. Lähtökielisen sanan käännökset saadaan etsimällä niiden

esiintymät rinnakkaiskokoelmasta ja keräämällä vastinlauseista tai -kappaleista usein esiintyvät kohdekieliset sanat. Käännös on tilastollinen ja tuottaa oikeiden tarkkojen käännösten lisäksi synonyymejä ja sanan kontekstiin läheisesti liittyviä muita sanoja sekä hälyä.

Kaikissa käännösmenetelmissä tuntemattomat sanat (out-of-vocabulary words) muodostavat tärkeän käännösongelman. Usein erisnimet ja tekninen terminologia puuttuvat käännössanakirjoista ja/tai rinnakkaiskokoelmasta, mutta ne saattavat olla hakutehtävässä keskeisiä hakusanoja, joiden kirjoitusasu kohdekielessä kuitenkin muistuttaa lähtökielistä. Esimerkiksi suomen *”kemoterapia”* -sanalla on ilmeinen merkkijonotason yhteys englannin *”chemotherapy”* -sanaan. Tällöin hakusanoja voidaan kääntää sumeiden merkkijonomenetelmien avulla. Niillä tunnistetaan samantapaisia, vaikkei identtisiä merkkijonoja.

Parhailla käännös- ja kyselymenetelmillä löydetään kohdekieliset dokumentit lähes yhtä tuloksellisesti kuin suoraan kohdekielellä hakemalla. Kishida (2005), Pirkola (1999) sekä Oard ja Diekema (1998) ovat katsauksia kieltenvälisen tiedonhaun tutkimukseen. Sumeita merkkijonomenetelmiä esittelee yleisesti muun muassa Navarro (2001) ja kieltenvälisen tiedonhaun kannalta Pirkola ja kumppanit (2002).

Dokumenttien automaattinen ryhmittely ja luokitus

Dokumenttien automaattisen *ryhmittelyn* (*clustering*) tavoite on tunnistaa dokumenttien keskinäinen samankaltaisuus ja sijoittaa samankaltaiset dokumentit samaan ryhmään. Nämä ryhmät muodostuvat automaattisesti ryhmiteltävän aineiston perusteella. Kahden dokumentin samankaltaisuutta mitataan niiden sisältämien sanojen tilastollisten jakaumien samankaltaisuudella. Kun kahden dokumentin samankaltaisuus ylittää asetetun kynnysarvon, ne voidaan sijoittaa samaan ryhmään. Jos samana ryhmään sijoitetaan useita dokumentteja, eri ryhmittelymenetelmät asettavat erilaisia vaatimuksia sille, kuinka monen muun ryhmään kuuluvan dokumentin kanssa kultakin dokumentilta vaaditaan samankaltaisuutta ja kuin suurta sen tulee olla. Vähimmillään vaaditaan kynnysarvon ylittävä yhteys edes yhteen muuhun ryhmän dokumenttiin, enimmillään kaikkiin.

Dokumenttien automaattisen *luokittelun* (*text classification*) lähtökohtana on ennalta annettu luokitus, jonka luokkiin uudet dokumentit tulisi automaattisesti sijoittaa. Tilastollinen automaattinen luokittelumenetelmä opetetaan opetusaineiston avulla. Opetusaineistona toimii joukko luokituksen luokkiin ennalta ja oikein

sijoitettuja dokumentteja, joiden perusteella kullekin luokalle lasketaan kuvaaja. Kuvaaja osoittaa tilastollisesti, mitä sanoja luokan tyypilliseen edustajaan todennäköisesti kuuluu. Luokitettavia uusia dokumentteja verrataan näihin kuvaajiin ja dokumentti voidaan sijoittaa parhaiten täsmäävään luokkaan tai useampaan kynnysarvon ylittävään luokkaan.

Dokumenttien automaattisen ryhmittelyn ja luokittelun menetelmiä esittelevät Baeza-Yates & Ribeiro-Neto (1999), Belew (2000), Salton (1989) ja Sebastiani (2002).

Rakenteisten dokumenttien haku (XML-tiedonhaku)

Dokumenttien digitaalisessa esityksessä on kolme ulottuvuutta: *sisältö* (tietosisältö), looginen *rakenne* (esimerkiksi pää- ja alaluvut, teksti; sinfonian rakenne; filmin kohtaukset) sekä *ulkoasu* (palstoitus, kirjasinmallit ja -koot; partituuri tai soitettu esitys; väri- ja mustavalkofilmi). Nykyisin on tapana merkitä dokumentin rakenne dokumentin esitykseen eksplisiittisesti. Tämä tapahtuu merkintäkielen avulla, joista esimerkkejä ovat verkkotiedolle yleisesti käytetty HTML (Hypertext Markup Language), SGML (Standard Generalized Markup Language) ja XML (eXtensible Markup Language). Näitä merkintäkieliä kuvataan tarkemmin World Wide Web -konsortion sivustolla (http://www.w3.org/). Dokumentin rakenneosat merkitään dokumenttiin tunnisteiden (tag) avulla, esimerkiksi ”<otsikko>Ontologinen tiedonhaku</otsikko>” voisi olla tämän artikkelin seuraavan alaluvun otsikon merkintä.

Dokumentin rakenneosat saattavat olla tyypiltään erilaisia, kuten tekstiä, taulukoita, kuvia, videota tai ääntä. Koska dokumentit voivat myös olla pitkiä, moniaiheisia ja rakenteeltaan monimutkaisia, seuraa tiedonhakijalle haasteita halutun tiedon paikantamisessa dokumentin sisältä. Dokumenteista saatetaan etsiä vain tiettyjä osia (esim. kuvia, mp3-kappaleita tai karttoja). Tiedontarve saattaa asettaa tarkennettuja ehtoja sille, missä dokumenttien osissa hakuavaimet esiintyvät (esim. lähdeviitteessä tai matkaoppaan alaluvun otsikossa). Joskus haetaan joukkoa tietyntyyppisiä rakenneosia (esim. mp3-kappaleita) tietokannan dokumenteista hakuehdolla, joka koskee muita rakenneosia (esim. säveltäjän tai esittäjän nimi).

Rakenneosien eksplisiittinen merkintä merkintäkielen avulla tarjoaa mahdollisuuden hyödyntää dokumenttien rakennetta niin hakuehtojen kuin dokumenteista poimittavien rakenneosien määrittelyssä. Rakenteisten dokumenttien haun tutkimus (*XML Information Retrieval*) kehittää tällaisia menetelmiä. Erityisen huomion kohteena ovat XML-merkintäkieli ja hakutulosten relevanssilajittelu

osittaistäsmäyksen periaatteiden mukaisesti, mutta kuitenkin niin, että hakija saisi mahdollisimman täsmällisiä ja kattavia dokumentin osia vastaukseksi. Pyritään siis yhtäältä löytämään parhaat mahdolliset vastaukset tietokannasta (osittaistäsmäytys) sekä toisaalta välttämään sisällöltään epärelevantteja rakenneosia (täsmällisyys) ja antamaan mahdollisimman täydellisiä vastauksia (kattavuus). Kyselyn tavoitteena voi olla myös rakenneosien keskinäisten suhteiden uudelleenjärjestäminen (esim. järjestetään mp3-kappaleita sisältävä kokoelma esittäjien-ja-kappaleiden sijasta levy-yhtiöiden ja kappaleiden mukaan).

Rakenteisten (XML-) dokumenttien hakumenetelmien tutkimusta kehitetään etenkin INEX-konferenssien puitteissa (http://inex.is.informatik.uni-duisburg.de/). Pal ja Mitra (i.v.) tarjoavat käyttökelpoisen yhteenvedon tutkimuksesta.

Ontologinen tiedonhaku

Web-dokumenttien ja muiden resurssien ominaisuudet ovat usein ihmisille helposti ymmärrettäviä, mutta niitä ei ole esitetty tavalla, joka tukisi niiden luotettavaa automaattista tulkintaa. *Semanttisen webin* tarkoituksena on liittää näihin dokumentteihin ja muihin resursseihin niiden merkityksen kuvauksia, jotka tukevat automaattista päättelyä (Berners-Lee, Hendler & Lassila 2001). Ontologiat ovat keskeinen väline semanttisen webin resurssien merkityksen kuvaamisessa. Ontologiat ovat tässä yhteydessä käsitejärjestelmiä, joissa käsitteiden väliset suhteet on esitetty formaalisti, automaattista käsittelyä tukevalla tavalla. Esimerkiksi pääskysten kuuluminen lintuihin ja selkärankaisiin voitaisiin esittää seuraavasti: ”pääskynen IS-A lintu”, ”lintu IS-A selkärankainen”, jos on sovittu, että ”IS-A” on operaattori, jolla ilmaistaan alalajin ja lajin hierarkkinen suhde. Tässä ei olla kovin kaukana kirjasto- ja tietopalvelutoiminnassa käytetyistä luokituskaavoista ja tesauruksista. Erona on kuitenkin suhteiden formaali esittäminen automaattista päättelyä varten ja pyrkimys käsitteiden täsmällisempään määrittelyyn.

Ontologiat ovat antaneet uutta puhtia perinteisille sisällönkuvailumenetelmille ja haastaneet pelkkään tekstisisältöön perustuvat tekstitiedonhaun menetelmät. Ontologia-pohjaisessa tiedonhaussa (*Ontology-based Information Retrieval*) puhutaan luokittamisen ja asiasanoituksen sijasta *annotoinnista,* jossa webin resursseihin liitetään ontologioista valittuja käsitteitä kuvaamaan niiden merkitystä. Yhtäältä tällainen toiminta on tarpeellista, jos tavoitteena on sanahakupalvelujen käyttö muun kuin tekstipohjaisen median hakemisessa – tekstuaalinen metadata on välttämätöntä,

vaikkei metadatan välttämättä tarvitse olla ontologista. Yhteisöllinen luokittelu (folksonomiat) on vähemmän vaativa tapa sisällön kuvaukseen (http://en.wikipedia.org/wiki/Folksonomy). Toisaalta semanttisen webin kehittäjät uskovat, että myös tekstipohjaisen median hakeminen nopeutuu ja tarkentuu, jos dokumentit ovat huolella annotoitu. Kriitikot kuitenkin väittävät, että annotointi on kallista, kuten luokitus ja indeksointikin, eikä onnistu automaattisesti. Itse asiassa webin dokumenteista vain häviävän pieni osa sisältää toistaiseksi mitään metadatana esitettyä sisällön kuvausta, vielä harvemmin se perustuu ontologioihin. Annotoinnin johdonmukaisuus – samanlaiset resurssit kuvataan yhdenmukaisella tavalla – on myös varsin vaikea haaste, joka on jo kauan tunnistettu luokituksen ja indeksoinnin tutkimuksessa. Annotoinnin puuttuminen ei kuitenkaan estä käyttämästä ontologioita tiedon hakemisessa: hakutehtävät voidaan määritellä ontologioiden avulla ja kääntää automaattisesti tavallisiksi tekstikyselyiksi hakukoneita varten. Tämä saattaa yhtäältä helpottaa tiedontarpeen tunnistamista ja täsmentämistä. Toisaalta haun tulosten merkitys ei ole alkuperäisen idean mukaisesti tarkkaan rajattu.

Semanttisen webin teknologioihin voi tutustua Wikipedian artikkelin (http://en.wikipedia.org/wiki/Semantic_Web) ja World Wide Web –konsortion sivuston (http://www.w3.org/2001/sw/) avulla. Suomalaista alan tutkimusta esittelee FinnONTO –projektin sivusto (http://www.seco.tkk.fi/projects/finnonto/index.fi.php). Kriitikon puheenvuoron esittää muun muassa Karen Sparck Jones (2004).

Web-tiedonhaku

Web on luonteeltaan virtuaalinen kokoelma dokumentteja sekä niihin liittyviä palveluja. Tiedonhaun tutkimuksessa on perinteisesti keskitytty kontrolloidusti rakennettuihin dokumenttikokoelmiin, jotka koostuvat sisällöltään ja muodoltaan suhteellisen homogeenista aineistoa. Webissä aineisto on *hajautetusti tuotettua, globaalia* ja *harvakseltaan linkitettyä hypermediaa.* Tiedonhaun kannalta uusia haasteita ovat muun muassa seuraavat:

- Mikä on dokumentti? Kokonaisuus, jota sen laatija kutsuu dokumentiksi, koostuu usein joukosta linkitettyjä tiedostoja. Loogista kokonaisuutta, jota kutsutaan dokumentiksi, voi olla vaikea tunnistaa automaattisesti. Dokumentti voi esiintyä useina versioina, tekijä voi muuttaa sitä tai poistaa sen milloin tahansa. Voi käydä myös niin, että dokumentti on olemassa ainoastaan hakijan

kuvaruudulla hetkellisenä koosteena, jonka osat on koottu kyselyn perusteella hajautetuista tallennuspaikoista.

- Perinteiset tiedonhakujärjestelmät käsittelevät dokumentteja itsenäisinä yksikköinä ja tiedonhakujärjestelmien tutkimus on keskittynyt sisältöön perustuviin menetelmiin. Webissä dokumenttien väliset linkit ovat muodostuneet olennaiseksi tekijäksi tiedonhaussa.
- Web on massiivinen kokonaisuus samalle alustalle tuotuja mutta sisällöltään, muodoltaan, tarkoitukseltaan ja laadultaan erilaisia aineistoja. Yleiskäyttöisen tiedonhakujärjestelmän kehittäjä joutuu sopeutumaan aineiston kirjavuuteen (esim. Google tai Yahoo); erikoispalvelun kehittäjän on ensin ratkaistava, miten rajata erikoisaineisto kokonaisuudesta (esim. Google Scholar).

Web-informaatioympäristön ominaisuuksiin, käyttäjiin ja hakemiseen kohdistuva tutkimus on luonut pohjaa varsinaiselle tiedonhakujärjestelmiin kohdistuvalle tutkimukselle. Hakujärjestelmiin liittyvä tutkimus on painottunut yhtäältä linkkirakenteen, toisaalta käyttäjistä kerättävän datan hyödyntämiseen.

Webin linkkirakenteen ominaisuuksien tunteminen on olennaista sekä keräilyrobotteja että hakualgoritmeja kehitettäessä. Yksittäisen HTML-sivun näkökulmasta tärkeitä ominaisuuksia ovat muun muassa, kuinka moni sivu viittaa ko. sivulle (in-links) ja kuinka moneen sivuun sivu itse viittaa (out-links). Dokumentit, joihin monet ulkoiset sivut ovat linkkautuneet, ovat ns. auktoriteettisivuja (authority pages) ja paljon linkkejä sisältävät dokumentit ns. napasivuja (hubs).

Tunnetuin linkkirakenteiden perusteella sivujen oletettua hyödyllisyyttä määrittävä algoritmi on Googlessa sovellettu *PageRank* (Brin & Page 1998). PageRankissä web-sivun hyvyys määräytyy kahden asian perusteella: (1) kuinka moni sivu viittaa ko. sivulle ja (2) kuinka hyviksi arvioituja viittaavat sivut ovat. Laskentakaava sivun p_i PageRankille on

$$PR(p_i) = \frac{1-d}{N} + d \cdot \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

missä p_1, $p_2, \ldots p_N$ ovat Web-sivuja, N on laskennan kohteena olevien sivujen kokonaismäärä (Webin tunnettu osa), d vaimennuskerroin, $M(p_i)$ on p_i:hin linkittyneiden sivujen joukko, $L(p_j)$ sivulta p_j lähtevien linkkien määrää (out-links) ja $PR(p_j)$ sivun p_j PageRank-arvo. Kun PR esiintyy kaavan molemmilla puolilla, laskenta

tapahtuu rekursiivisesti kaikille tarkastelussa oleville sivuille $p_1 \ldots p_N$ lähtien liikkeelle sopivasti valitusta oletusarvosta, esimerkiksi $PR(p_{1..N}) = 1$.

PageRank perustuu ajatukselle, että kukin sivulle viittaava linkki edustaa muiden sivuntekijöiden suositusta. Suositus normeerataan tekijällä $L(p_j)$, jolloin monille sivuille viittaavien tekijöiden suositukset eivät ole yhtä painavia kuin vain harvoille sivuille viittaavien. Jos viittaavat sivut $M(p_j)$ ovat saaneet nekin paljon viittauksia, niiden PageRank-arvot ovat korkeita ja tämä kumuloituu myös sivulle p_i. PageRank ei ota kantaa sivujen sisältöön vaan ainoastaan linkkirakenteen perusteella arvioituun auktoriteettiin. Sitä voidaan soveltaa sisältöperusteisiin täsmäytysalgoritmeihin yhdistettynä hakutulosten relevanssilajitteluun, edustavimpien tulossivujen valintaan hakijalle esitettäväksi sekä keräilyrobottien ohjaamiseen (Yang 2005).

Web-sivujen tekijöiden linkittämällä ilmaisemien suositusten lisäksi hakujärjestelmien tutkimuksessa on kiinnitetty kasvavaa huomiota käyttäjän tuottaman datan kuten klikkausinformaation hyödyntämiseen. Tämä voi tarkoittaa tietojen tallentamista hakijan kyselytuloksesta avaamista linkeistä ja niiden käyttöä relevanssilajittelussa, kun sama kysely toistuu myöhemmin (ks. Xue et al. 2004). Klikkausinformaation käyttö on mahdollista silloin, kun samojen kyselyjen esiintymisen todennäköisyys on suuri. Myös sosiaaliset yhteisöt kuten verkottuneet harrastuspiirit, joiden sivuilla samanhenkiset käyttäjät vierailevat, tarjoavat mahdollisuuden tutkia klikkausinformaation hyödyntämistä (Smyth 2007).

Hakukoneet kirjaavat perustiedot käyttäjän jokaisesta kyselystä ja kyselytuloksen perusteella avaamista linkeistä käyttölokiin. Lokien perusteella voidaan tutkia hakijan käyttäytymistä analysoimalla kyselyjen ja hakuistuntojen ominaisuuksia (Jansen & Spink 2006). Lokitutkimus antaa realistisen ja aidon kuvan hakijoiden tavoista hakea, mutta tutkijalla ei ole pääsyä kyselyjen taustalla oleviin käyttäjän tavoitteisiin ja tarpeisiin. Hakijan käyttäytymisen ja siihen vaikuttavien tekijöiden ymmärtäminen edellyttää käyttäjäsuuntautunutta tehtäväperustaista lähestymistapaa. Käyttäjän havainnointi ja haastattelu hänen omassa työskentely-ympäristössään auttaa ymmärtämään hakujen taustalla olevia tekijöitä (Pharo & Järvelin 2004).

Vuorovaikutteinen tiedonhaku

Tiedonhaun tutkimuksessa on vahva laboratoriotutkimuksen traditio, jossa evaluoidaan tiedonhakujärjestelmiä testitehtävien avulla parempien teknisten hakumenetelmien kehittämiseksi. Tämä tutkimus tapahtuu pääasiassa sellaisten koeasetelmien puitteissa,

joihin ei osallistu koehenkilöitä, vaan joissa hakutehtäviä suoritetaan automaattisesti yksi kerrallaan. Vaikka näin saadaankin tietoa hakumenetelmien toimivuudesta, ei ole ilman muuta selvää, että niistä olisi tiedonhakijoille hyötyä käytännössä, koska koeasetelma ei vastaa monia tavallisia tiedontarvetilanteita.

Aidossa tiedonhakutilanteessa tiedonhakijan tiedontarve voi olla epäselvä ja sen tyydyttäminen voi vaatia useita erilaisia hakutehtävän muotoiluja kyselyinä. Ensimmäinen kyselyn muotoilu voi olla kaukana optimaalisesta. Lisäksi tiedontarve voi täsmentyä tai muuttua hakuprosessin aikana ja tiedonhakija voi vaihtaa aihetta tai näkökulmaa, koska oppii hakemisen aikana. Jos haku ei tuota toivottua tai riittävää tulosta, hakija saattaa luovuttaa saamatta haluamaansa vastausta. Tämä tosin riippuu hakijan tilanteesta ja persoonasta sekä tiedontarpeen kriittisyydestä. Näitä seikkoja ei laboratoriotutkimuksessa voida ottaa huomioon, jos hakutehtävästä muotoillaan vain yksi kysely automaattisesti.

Vuorovaikutteisen tiedonhaun tutkimuksessa (*Interactive Information Retrieval*) koehenkilöt, tiedonhakijat, suorittavat hakutehtäviä ja voivat suorittaa yhden istunnon kuluessa useita kyselyjä. Tutkimuksen tavoitteena on perehtyä tiedonhakijoiden toimintaan ja kohtaamiin ongelmiin hakuprosessin aikana ja kehittää keinoja hakijoiden tueksi. Nämä keinot voivat koskea hakijan avustamista (1) tunnistamaan ja esittämään tiedontarpeensa, (2) valitsemaan sopivimmat tietokannat haun kohteeksi, (3) valitsemaan sopivat vuorovaikutusmenetelmät (haku, navigointi, selailu), (4) kyselynsä muotoilussa ja moniselitteisyyden ratkonnassa, (5) saatujen hakutulosten tutkimisessa ja järjestämisessä, (6) kyselyn parantamisessa palautteen avulla, (7) jo aikaisemmin löydettyjen dokumenttien löytämisessä uudelleen, (8) dokumenttien sopivimpien rakenneosien tarkastelussa. Vuorovaikutteisen tiedonhaun tutkimuksessa voidaan keskittyä luonteeltaan avoimiin, rajaamattomiin tiedonhakutilanteisiin, kun yleisessä verkkotiedonhaussa, tai tiedontarpeen luonteen, aihealueen tai henkilön tehtävän suhteen rajattuihin tilanteisiin.

Ian Ruthvenin (2007) katsaus johdattaa vuorovaikutteisen tiedonhaun tutkimukseen ja Pertti Vakkarin (2002) katsaus tehtäväperusteisen tiedonhaun tutkimukseen.

Kuvatiedonhaku

Kuvat ovat esimerkkejä dokumenteista, joissa keskeinen sisältö ei olekaan tekstiä vaan ei-kielellistä informaatiota (bittivirtaa). Kuvatiedonhaussa voidaan soveltaa sekä sanalliseen metadataan että visuaaliseen sisältöön perustuvia hakumenetelmiä.

Käytännön kuvahakujärjestelmät perustuvat tyypillisesti sanalliseen metadataan kohdistuviin kyselyihin ja pienikokoisten korvikekuvien (thumbnails) selailuun. Kuvatiedonhaku onkin enemmän selailuun perustuvaa kuin tekstidokumenttien haku. Kuvien silmäily yleisvaikutelman saamiseksi on nopeaa. Metadataa on tyypillisissä sovelluksissa niukasti joten sekin pakottaa hakijan käyttämään selailua.

Kuvassa 3 on esimerkki Flickr-verkkopalvelusta haetusta kuvasta[3]. Flickr-palvelun avulla kuka tahansa voi julkaista Web:ssä omia kuviaan, liittää niihin kuvailutietoja ja kommentoida muiden kuvia. Kyse on siis sosiaalisen Webin palvelusta, jossa yhteisö tuottaa aineistoon metadataa (folksomia-periaate). Esimerkin kuvassa on kirjattu kuvan kohde yleisellä tasolla (Take your sauna everywhere you go!) ja kuvaan liittyvä tapahtuma (Mobile Sauna Festival), kuvauspaikka (Teuva, Finland), kuvausaika (August 12, 2007), kuvaaja (Andrea Vascellari) sekä avainsanoja (mobile, sauna, vascellari, teuva, finland, andreavascellari). Kuva on liitetty myös ”Finland”-nimiseen joukkoon (set) ja alajoukkoon (pool) ”Suupohja” ja se on kirjattu käyttöoikeudeltaan julkiseksi (public). Esimerkissä näkyy sanallisen kuvailutiedon niukkuus, mutta toisaalta kuvaan liittyy monia ”kahvoja”, joilla poimia uusi kuvarypäs selailtavaksi.

[3] Kiitämme Andrea Vascellaria (www.andreavascellari.com) eimerkkikuvasta.

Kuva 3. Esimerkki Flickr-kuvahakupalvelun hakutuloksesta kyselyllä ”sauna” (ks. http://www.flickr.com/)

Vanha vertaus: ”Kuva kertoo enemmän kuin tuhat sanaa”, luonnehtii sattuvasti kuvien sanallisen kuvailun ongelmaa. Kuva sisältää aina enemmän yksityiskohtia ja tulkintamahdollisuuksia kuin sanallisessa kuvailussa on mahdollista tuottaa. Visuaaliset vaikutelmat ovat myös hankalia yhdenmukaisilla kielellisillä ilmauksilla esitettäviksi. Monet kuvan ominaisuudet on järkevintä jättää hakijan itsensä selausvaiheessa tulkittaviksi.

Digitaaliseen kuva-analyysiin perustuvat *piirrepohjaiset* (*faeture-based*, *content-based*) *indeksointi- ja hakumenetelmät* ovat hyvin aktiivisen tutkimuksen kohteena. Kuvasignaalin analyysi tuottaa kuvan visuaalista sisältöä edustavat *piirrevektorit* (*feature vectors*). Piirrevektorit mittavat esimerkiksi kuvan keskimääräistä värisisältöä, pintarakenteen (texture) ominaisuuksia ja tunnistettavia muotoja. Kyselynä voidaan antaa mallikuva, josta laskettuja piirrevektoreita täsmäytetään tietokannan kuvien

vastaaviin vektoreihin. Kysely palauttaa tietokannasta kuvia visuaalisen samankaltaisuuden mukaan lajiteltuna.

Piirrepohjaiset algoritmit pystyvät toistaiseksi tunnistamaan kuvasta vain hyvin matalan abstraktiotason piirteitä. Kuvassa esiintyvien objektityyppien tunnistaminen – puhumattakaan nimetyn henkilön hakemisesta mallikuvan perusteella - on vielä ylivoimainen tehtävä. Piirrevektorien soveltaminen onkin levinnyt vain kuvahaun erityissovelluksiin kuten lääketieteen kuvajärjestelmiin ja kasvokuvarekistereihin. (Enser 2007).

Videotiedonhaku

Videodokumentti on erityinen siinä mielessä, että se sisältää aidosti montaa rinnakkaista mediaa: liikkuvaa kuvaa, ääntä ja tekstiä. Kirjastot ja videovuokraamot ylläpitävät perinteisiä videokokoelmia, jossa videodokumentteja käsitellään erillisinä fyysisinä tallenteina, videonauhoina tai DVD-levyinä. Näissä sovelluksissa tiedonhaku perustuu erilliseen, sanallista metadataa sisältävään viitetietokantaan eikä hakija pääse välittömästi selailemaan dokumentin sisältöä. Integroidut videoarkistojärjestelmät ovat vähitellen yleistymässä. Näissä järjestelmissä voi tehdä sekä kyselyjä ja selailla itse dokumentteja.

Liikkuvaan kuvaan liittyvät tutkimuskysymykset ovat huomattavasti haasteellisempia kuin still-kuviin liittyvät. Liikkuvalla kuvalla on ajallinen rakenne, joka koostuu yksittäiskuvista (frames), otoista (shots) ja kohtauksista (scenes). Ensimmäinen haaste on tunnistaa ja merkata tuo rakenne. Oton rajat ovat helpoiten automaattisesti tunnistettavissa, mutta yleensä se on liian pieni yksikkö manuaalisesti kuvailtavaksi. Se auttaa kuitenkin videon laajempien kerronnallisten osien tunnistamisessa ja ns. videotiivistelmien luomisessa.

Videotiedonhauntutkimus ei ole johtanut läpimurtoihin videokuvan sisällön automaattisen indeksoinnin alueella. Piirrepohjaiset menetelmät ovat rajoittuneita myös videon osalta vaikka liikkuvan kuvan erityispiirteet (kameran ja kohteen liike) antavatkin uusia mahdollisuuksia. Ehkä hedelmällisimmiksi tutkimusalueiksi ovatkin muodostuneet videotiivistelmien automaattiseen tuottamiseen, tiivistelmien selailua tukeviin käyttöliittymätekniikkoihin ja videodokumentin äänen, erityisesti puheen hyödyntämiseen (ks. puhetiedonhaku) liittyvä työ. (Smeaton 2004.)

Puhetiedonhaku

Erityisesti radio- ja televisiotoiminnassa syntyy suuria sähköisiä arkistoja, jotka sisältävät puhuttua luonnollista kieltä. Silloin kun niiden tallennus perustuu digitaalisiin menetelmiin, voidaan myös tällaisia dokumentteja hakea niiden puhutun sisällön perusteella. Tällöin puheena tai kirjoitettuna esitetyn hakutehtävän ja puhedokumenttien valikoidut äännepiirteet esitetään yhteensopivalla tavalla täsmäytystä varten. Allan (2002), Ng, Wilkinson & Zobel (2000) sekä Pylkkönen & Kurimo (2004) tarkastelevat puhehaun menetelmiä.

Puhehakuun (speech retrieval) on kaksi päälähestymistapaa. Yhtäältä voidaan tunnistaa puhe kirjoitetuiksi sanoiksi käyttäen *jatkuvan puheen puheentunnistinta*. Tunnistusvirheitä voidaan vähentää muun muassa käyttämällä useita erilaisia tunnistimia rinnakkain ja tunnistamalla sekä erillisiä sanoja että sanaliittoja. Tunnistimelle tuntemattomat puhedokumentin sanat tunnistuvat väärin ja aiheuttavat siten kaksinkertaisen virheen. Väärintunnistus voi myös vaikuttaa muiden sanojen tunnistamiseen ja virhe voi siten levitä. Haasteita aiheuttavat myös sanojen taivutusmuodot joissakin kielissä, kuten suomessa, joissa tunnistettavan sanaston koko saattaa kasvaa nopeasti taivutusmuotojen takia. Toinen lähestymistapa perustuu sanaa pienempien yksikköjen tunnistamiseen, esimerkiksi *foneemien tunnistamiseen*. Tällöin vältetään tuntemattomien sanojen ja taivutusmuotojen aiheuttamat ongelmat – sanoja ei edes yritetä tunnistaa – mutta menetetään mahdollisuus tarkastaa sanakirjan avulla tunnistuksen mielekkyys. Foneemitason menetelmissä sekä hakutehtävä että puhedokumentti käännetään äänneasuesityksiksi, joita sitten vertaillaan. Vertailu tapahtuu tavallisesti sumeiden merkkijonomenetelmien avulla (ks. kieltenvälinen tiedonhaku). Haasteita puhehaun tutkimukseen tarjoaa myös puhujariippumattomien hakumenetelmien kehittämisen tarve.

Puheen ja tekstin käyttö voidaan yhdistää tiedonhaussa periaatteessa monella tapaa riippuen dokumenttien käyttötilanteesta ja tietokannan luonteesta. Esimerkiksi matkustettaessa voidaan pyytää matkapuhelimen avulla uutisia (tai muita äänidokumentteja) kiinnostavasta aiheesta. Jos käytettävä tietokanta on tekstitietokanta, voidaan puhuttu hakupyyntö kääntää kirjoitetuksi ja suorittaa tavallinen tekstihaku. Tulosdokumentit voidaan jälleen kääntää synteettiseksi puheeksi joko palvelimella tai matkaviestimessä. Myös tekstinä esitetty hakupyyntö voidaan esittää automaattisesti kyselynä siten, että puhedokumenttien haku onnistuu.

Musiikin tiedonhaku

Perinteinen musiikkikirjastotoiminta luetteloi ja luokittaa musiikkidokumentit (äänilevyt ja - kasetit, sävellykset) tekstiin perustuvina kuvauksina, joita voidaan hakea tekstitiedonhaun menetelmin. Haku perustuu tällöin sekä teoksiin liitettyyn metadataan että joskus myös teoksiin sisältyviin sanoituksiin. Kolmantena vaihtoehtona ovat suosittelut, joita teoksiin on voitu liittää suosittelujärjestelmissä (recommender systems). Varsinaista musiikkia ei kuitenkaan voida tällöin suoraan hakea ”sisältönsä” perusteella. Musiikille on kuitenkin vakiintunut esitystapa nuotteina ja useita digitaalisia esitystapoja, jotka esittävät musiikin äänenä. Nämä tarjoavat kolme musiikin melodiapohjaista hakutapaa: musiikin hakua nuoteilla, musiikin hakua melodiahahmolla ja musiikin hakua hyräilemällä. Nuotteihin perustuvat musiikin hakutavat kuitenkin edellyttävät, että hakija osaa tuottaa riittävän sävelhahmon joko nuottien avulla tai säveltapailulla (Rikkilä 2008). Edellä tarkastellut sumeat merkkijonomenetelmät tulevat tässä kuitenkin apuun. Sisältöperusteiset menetelmät ovat joka tapauksessa tarpeellisia, koska:

- hakijalla voi olla puutteelliset luettelointi- ja luokitustiedot hakemastaan sävellyksestä
- hän kenties tuntee vain osan sävellyksen melodiasta
- tekijänoikeuden valvonta: halutaan kartoittaa tietyn sävelidean käyttö tietokannan eri teoksissa.

Muun muassa MIDI -tiedostoesityksen perusteella voidaan rakentaa yksinkertaistettu sävellyksen esitys, joka kuvaa melodian, joka lienee tavallisesti helpoimmin muistettava sävellysten piirre, ja joka (Salosaari & Järvelin 1998):

- sallii kyselyt, jotka ovat musiikkia
- yksinkertaistaa nuottien ja tulkinnan monimutkaisen MIDI -esityksen
- sallii sumean haun - siis kyselyt, jotka ovat musiikillisesti “sinne päin“, esimerkiksi osittain virheellisiä sävelkorkeuden- ja keston tai säveljärjestyksen suhteen.

Musiikin tiedonhausta (*Music Information Retrieval*) kirjoittavat mm. Lemström (2000), McNab ja kumppanit (1997), Salosaari (1998) ja Rikkilä (2008). Aiheesta järjestetään vuosittain kansainvälinen symposium ISMIR. Vuoden 2007 konferenssin julkaisut löytyvät osoitteesta http://ismir2007.ismir.net/.

Monialainen tiedonhaku

Tiedonhaun kehittäminen vaatii monialaista osaamista. Luonnollisen kielen (kirjoitettuna tai puhuttuna) hallinta asettaa monia haasteita hakijan mielikuvitukselle ja järjestelmien käsittelykyvylle. Dokumentit itsessään voivat palvella hyvin monenlaisia tiedonvälitystehtäviä ja nämä vaikuttavat niin dokumenttien sisältöön, ulkoasuun kuin rakenteeseenkin, jotka kaikki vaikuttavat tiedonhaun onnistumiseen. Tietotekniikka on tuottanut kirjon menetelmiä tekstinä ja muina medioina tuotettujen dokumenttien ja hakupyyntöjen esittämiseen ja täsmäyttämiseen. Tekniikaltaan erilaiset tiedonhakuympäristöt, kuten avoin tietoverkko, organisaatioiden sisäiset tietokannat tai erikoisalojen tietokannat tarjoavat eri tavoin valikoituja ja suojattuja dokumentteja. Tiedonhakuympäristöt säätelevät niin dokumenttien valintaa kuin niiden esitys- ja käyttötapojakin. Eri työympäristöissä toimii myös sangen monenlaisia tiedonhakijoita, joiden tiedontarpeet, tiedonhankinnan käytännöt ja tiedonhakutaidot vaihtelevat. Kaikki nämä seikat tulee ottaa huomioon tiedonhakua kehitettäessä. Siksi ala tarjoaakin mahdollisuuksia niin tietotekniikasta kiinnostuneille kuin myös humanisteille ja yhteiskuntatieteilijöille — edellyttäen, että heillä on kiinnostusta tulla toimeen muiden mainittujen osaamisalueiden kanssa.

Lähteet

Alkula R (2000). Merkkijonoista suomen kielen sanoiksi. Väitöskirja. Acta Electronica Universitatis Tamperensis, 51. Tampere, TUP, URL: http://acta.uta.fi/pdf/951-44-4886-3.pdf

Allan, J. (2002). Perspectives on Information Retrieval and Speech. Teoksessa: Coden, A.; Brown, W. & Srinivasan, S. (toim.) Information Retrieval Techniques for speech Applications. Springer, 1–10.

Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Boston, MA, Addison-Wesley Longman.

Belew, R.K. (2000). Finding Out About. Cambridge University Press, Cambridge.

Berners-Lee, T. & Hendler, J. & Lassila, O. (2001). The semantic Web. Scientific American 284(5): 34-45.

Borlund, P. (2003). The concept of relevance in IR. Journal of the American Society for Information Science and Technology, 54(10): 913-925.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Proceedings of the 7th International WWW Conference, 107-117.

Enser, P. (2007). Visual image retrieval. Teoksessa: Cronin, B. (toim.). Annual review of information science and technology (ARIST 42). New York: American Association for Information Science and Technology. s. 3-42.

Cardoso, J. (toim.) (2007). Semantic Web Services. Hershey, Information Science Reference.

Glavitsh, U. & Schäuble, P. (1992). A System for retrieving speech documents. Teoksessa: Belkin, N. & Ingwersen, P. & Mark Pejtersen, A.-L. (toim.). Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Special Issue of SIGIR Forum, June 21-24, 1992, 168-176.

Gruber, T. (2008). Collective Knowledge Systems: Where the Social Web meets the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 6(1), 4-13.

Ingwersen, P. & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context. Dortrecht, Springer.

Jansen, B.J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management 42(1): 248-263.

Järvelin, K. (1995). Tekstitiedonhaku tietokannoista. Espoo, Suomen ATK-kustannus.

Kekäläinen, J. (1999). The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Väitöskirja. Acta Universitatis Tamperensis 678. Tampere, TAJU.

Kettunen, K. (2007). Reductive and Generative Approaches to Morphological Variation of Keywords in Monolingual Information Retrieval. Väitöskirja. Acta Electronica Universitatis Tamperensis 654. Tampere, TAJU.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. Information Processing & Management, 41(3): 433-455.

Lemström, K. (2000). String Matching Techniques for Music Retrieval. Helsinki, Helsinki University Press.

McNab, R. J. & Smith, L.A. & Bainbridge, D. & Witten, I. H. (1997). The New Zealand Digital Library MELody inDEX. D-Lib Magazine, May 1997. URL: http://www5.cnri.reston.va.us/dlib/may97/meldex/05witten.html

Navarro, G. (2001). A guided tour to approximate string matching. ACM Computing Surveys 33 (1), 31-88.

Ng, C., Wilkinson, R. & Zobel, J. (2000). Experiments in spoken document retrieval using phoneme n-grams. Speech Communication 32(1–2): 61–77.

Oard, D. & Diekema, A. 1998. Cross-Language Information Retrieval. Martha Williams (Toim.), Annual Review of Information Science (ARIST), Vol. 33, Medford, NJ, Information Today, 223-256.

Pal, S. & Mitra, M. (i.v.). XML Retrieval: A Survey. URL: http://www.isical.ac.in/~sukomal_r/survey.pdf.

Pharo, N. & Järvelin, K. (2004). The SST method: a tool for analysing Web information search processes. Information Processing & Management 40(4): 633-654.

Pirkola, A. (1999). Studies on linguistic problems and methods in text retrieval: the effects of anaphor and ellipsis resolution in proximity searching and translation and query structuring methods in cross-language retrieval. Väitöskirja. Acta Universitatis Tamperensis, 672. Tampere. TUP.

Pirkola, A. & Keskustalo, H. & Leppänen, E. & Känsälä, A-P. & Järvelin, K. (2002). Targeted S-Gram Matching: A Novel N-Gram Matching Technique for Cross- and Monolingual Word Form Variants. Information Research, 7(2). URL: http://InformationR.net/ir/7-2/paper126.html.

Pylkkönen, J. & Kurimo, M. (2004). Duration Modeling Techniques for Continuous Speech Recognition. Proc. 8th International Conference on Spoken Language Processing (Interspeech), 385–388.

Rikkilä, J. (2008). Musiikin melodiapohjaiset digitaaliset tiedonhakupalvelut: Nuotti-, melodiahahmo- ja hyräilyhaun käyttäjälähtöinen evaluointi. Tampereen yliopisto, Informaatiotutkimuksen laitos, Pro gradu –tutkielma.

Ruthven, I. (2007). Interactive information retrieval. Teoksessa: Cronin, B. (Toim.) Annual Review of Information Science and Technology: Volume 42 (ARIST 42). Medford, NJ. Information Today,. 43-91.

Salosaari, P. (1998). Nimikirjoitustiedostoihin perustuva musiikin tiedonhakumalli. N-grammit melodian kuvausmenetelmänä. Tampereen yliopisto, Informaatiotutkimuksen laitos, Pro gradu -tutkielma.

Salosaari, P. & Järvelin, K. (1998). MUSIR — A Retrieval Model for Music. Tampere, Finland: University of Tampere, Department of Information Studies, Research Note RN-1998-1. 14 p.

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Reading, MA: Addison-Wesley.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. Journal of the American Society for Information Science and Technology, 58 (13), 1915-1933.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47.

Smeaton, A. (2004). Indexing, browsing, and searching of digital video. Teoksessa: Annual Review of Information Science and Technology (ARIST), Vol. 38, 371-407. Medford: Information Today.

Smyth, B. (2007). A community-based approach to personalizing web search. Computer 40(8): 42-50.

Sormunen, E. (2000). A method for measuring wide range performance of Boolean queries in full-text databases. Väitöskirja. Acta Electronica Universitatis Tamperensis, URL: http://granum.uta.fi/pdf/951-44-4732-8.pdf. Tampere, TUP.

Sparck Jones, K. (2004). What's new about the Semantic Web?: some questions. SIGIR Forum, 38(2), 18-23. URL: http://portal.acm.org/citation.cfm?doid=1041394.1041398.

Vakkari, P. (2002). Task based information searching. Teoksessa: Cronin, B. (Ed.) Annual Review of Information Science and Technology (ARIST). Vol. 37. Medford, NJ: Information Today, 413-464.

Xue , G.R., Zeng , H-J., Chen, Z. & kumpp. (2004). Optimizing web search using web click-through data. Teoksessa: Proceedings of the thirteenth ACM international conference on Information and knowledge management. New York: ACM, 118 – 126.

Yang, K. (2005). Information Retrieval on the Web. Teoksessa: Annual Review of Information Science and Technology (ARIST). Vol. 40, Medford: Information Today, 33-80.