Tiedonhakukäyttäytyminen vuorovaikutteisessa tiedonhakutilanteessa: Tuloslistasta katsotun dokumentin vaikutus hakuavaimen valintaan kyselyn uudelleen muotoilussa

Anna-Kaisa Hyrkkänen
Tampereen yliopisto
Informaatiotutkimuksen laitos
Pro gradu -tutkielma
Huhtikuu 2008

TAMPEREEN YLIOPISTO

Informaatiotutkimuksen laitos

HYRKKÄNEN, ANNA-KAISA: Tiedonhakukäyttäytyminen vuorovaikutteisessa tiedonhakutilanteessa: Tuloslistasta katsotun dokumentin vaikutus hakuavaimen valintaan kyselyn uudelleen muotoilussa
Pro Gradu -tutkielma, 56 s., 1 liites.
Informaatiotutkimus
Huhtikuu 2008

Tutkielmassa tarkasteltiin kyselyn uudelleen muotoiluun liittyvää tiedonhakukäyttäytymistä vuorovaikutteisessa tiedonhakutilanteessa. Vuorovaikutteisen tiedonhaun tutkimus käsittelee käyttäjän, tiedontarpeen ja tiedonhakujärjestelmän vuorovaikutteista suhdetta. Tiedonhakuprosessi nähdään vuorovaikutteisena tapahtumana, jossa vuorovaikutteisuutta ilmenee muun muassa käyttäjän muotoillessa kyselyitä, tulkitessa saatuja dokumentteja ja informaatiota käytettäessä.

Kyselyjen uudelleen muotoilun osalta oltiin erityisen kiinnostuneita lähtökyselyyn lisättyjen hakuavainten alkuperästä. Tutkielman tarkoituksena oli tarkastella tuloslistasta katsottujen dokumenttien vaikutusta uusien hakuavainten valintaan. Tutkielmassa haluttiin selvittää, kuinka suuri osa niistä jatkokyselyjen hakuavaimista, jotka eivät esiintyneet tehtäväkuvauksessa, olivat tunnistettavissa tuloslistasta katsotuista dokumenteista. Tutkielman empiirinen aineisto koostui osasta INEX 2004 -hankkeen vuorovaikutteisen tiedonhaun tutkimuslinjan aineistoa. Tehtäväkuvausten ulkopuolisten hakuavainten esiintymistä tuloslistan dokumenteissa selvitettiin tutkimalla hakijoiden suorittamien hakujen lokitietoja. Lokitiedoista tarkasteltiin kyselyitä, joissa tehtäväkuvauksen ulkopuolisia hakuavaimia esiintyi sekä dokumentteja, joita hakijat olivat katselleet ennen tällaisen kyselyn muodostamista. Jos hakuavaimen todettiin esiintyvän dokumentista, selvitettiin, millaisen relevanssiarvion se oli saanut hakijalta. Lisäksi tutkittiin, miltä kohtaa dokumenttia hakuavain oli löydettävissä ja missä muodossa hakuavain oli tunnistettavissa.

Tutkimustulokset osoittivat, että suurin osa hakuavaimista, jotka eivät olleet löydettävissä tehtäväkuvauksista, löytyivät sen sijaan hakijoiden katsomista tuloslistan dokumenteissa. Dokumentit, joista jatkokyselyjen hakuavaimia esiintyi, arvioitiin pääsääntöisesti relevanteiksi. Tällaiset dokumentit sijoittuivat hakijoiden tarkasteluissa lähelle uuden kyselyn muodostushetkeä. Vaikka suurin osa hakuavaimista oli löydettävissä dokumenttien leipätekstistä, huomattavan suuri osa hakuavaimista esiintyi keskeisemmässä kohdassa dokumenttia, kuten dokumentin otsikosta, väliotsikossa tai dokumentin alusta.

Asiasanat: vuorovaikutteinen tiedonhaku, INEX 2004, kyselyn uudelleen muotoilu, lokianalyysi

SISÄLLYSLUETTELO

1 Johdanto

Tämä tutkielma kuuluu vuorovaikutteisen tiedonhaun tutkimuksen piiriin. Vuorovaikutteinen tiedonhaun tutkimus on kiinnostunut käyttäjän, tiedontarpeen ja tiedonhakujärjestelmän vuorovaikutteisesta suhteesta. Koko tiedonhakuprosessi voidaan nähdä vuorovaikutteisena tapahtumana, jossa vuorovaikutteisuutta ilmenee muun muassa käyttäjän muotoillessa kyselyitä, tulkitessa saatuja dokumentteja ja informaatiota käytettäessä. (Rieh & Xie 2005, 753.) Vuorovaikutteisessa tiedonhakutilanteessa pyritään pääsemään mahdollisimman lähelle todellista tiedonhakutilannetta säilyttämällä kuitenkin suhteellisen kontrolloitu testiympäristö (Borlund 2000, 75).

Tutkielman aiheena on tarkastella tiedonhakijoiden kyselyjen uudelleen muotoiluun liittyvää tiedonhakukäyttäytymistä vuorovaikutteisessa tiedonhakutilanteessa. Aiemmissa aihetta koskevissa tutkimuksissa pääpaino on ollut suoritettujen kyselyjen sekä käytettyjen hakuavainten määrien tarkastelussa. Kyselyjen uudelleen muotoilua koskevissa tutkimuksissa on tyypillisesti tarkasteltu muutoksia, joita hakijat tekevät kyselyihin. Tässä tutkimuksessa ollaan kiinnostuneita jatkokyselyihin lisättyjen hakuavainten mahdollisesta alkuperästä. Syitä, joiden perusteella hakijat valitsevat uusia hakuavaimia kyselyihinsä on jokseenkin mahdotonta todistettavasti esittää, mutta oletuksia hakuavainten alkuperästä voidaan kuitenkin esittää. Tämän tutkimuksen tarkoituksena on tarkastella tuloslistasta katsottujen dokumenttien vaikutusta kyselyjen uudelleen muotoiluun. Lähtökohtana tutkimukselle on oletus siitä, että uudelleen muotoilluissa kyselyissä esiintyvät muut kuin tehtäväkuvauksen hakuavaimet ovat tunnistettavissa joko suoraan tai epäsuorasti hakijoiden tuloslistasta katsomista dokumenteista. Tutkimukselle on aihetta koska vastaavanlaisia tutkimuksia ei ole aiemmin tehty. Nurmela (2006) tarkasteli pro gradu -tutkielmassaan tehtäväkuvauksen sanojen ja hakuavainten yhdenmukaisuutta mutta vastaavanlaista tutkimusta tuloslistan dokumenttien sanojen ja hakuavainten yhdenmukaisuudesta ei ole tehty.

Tutkielman empiirinen aineisto koostuu osasta INEX 2004-hankkeen vuorovaikutteisen tiedonhaun tutkimuslinjan aineistoa. INEX-hanke, joka tulee sanoista INitiative for the Evaluation of XML Retrieval, on tiedonhaun tutkimusta ja evaluointia varten XML-ympäristössä perustettu yhteistyöverkosto. Vuorovaikutteinen tutkimuslinja on kuulunut INEX-hankkeeseen vuodesta 2004 lähtien ja sen tavoitteena on tutkia hakijoiden käyttäytymisistä kun haun kohteena on XML-dokumentteja. (Fuhr, Lalmas, Malik & Szlavik 2004, 7.)

Tutkielmassa lähdetään liikkeelle esittelemällä tutkimuksen kannalta keskeisimmät käsitteet. Kolmannessa luvussa esitellään tiedon tallennuksen ja haun tasoperiaate. Neljännessä luvussa tarkastellaan tiedonhaun vuorovaikutteista luonnetta ja esitellään simuloidun työtehtävän malli sekä relevanssin kiistelty käsite. Viidennessä luvussa kerrotaan kyselyjen muokkaamisesta ja niiden muokkaamiseen käytetyistä menetelmistä sekä tarkastellaan tutkimuksia, joita on tehty hakijoiden tiedonhakukäyttäytymistä. Kuudennessa luvussa esitellään tutkimuskysymykset, käytetty tutkimusaineisto sekä aineiston käsittelyyn käytetyt menetelmät. Ennen tutkimustuloksia esitellään Nurmelan (2006) tutkimuksen tuloksia simuloidun tehtäväkuvauksen vaikutuksesta kyselyjen muodostamiseen. Kahdeksas luku sisältää tutkimuksen tulokset ja viimeisessä eli yhdeksännessä luvussa vedetään yhteenveto ja johtopäätökset tutkimuksen tuloksista.

2 Tutkimuksen keskeisiä käsitteitä

2.1 Tiedonhaun tutkimus ja tiedonhakujärjestelmät

Tiedonhaun tutkimus (IR, Information Retrieval) kuuluu osaksi tiedonhankinnan (IS, Information Seeking) laajempaa kokonaisuutta. Tiedonhauksi voidaan nähdä se osa tiedonhankintaa, joka suoritetaan käyttämällä apuna tietokonetta. (Marchionini 1995, 8.) Tämä työ kuuluu tiedonhaun tutkimuksen piiriin.

Tiedon tallennuksen ja haun tutkimuksen päätavoite on kehittää sellaisia käsitteitä, menetelmiä ja järjestelmiä, joiden avulla tarvittava tieto, riippumatta tiedon esitysmuodosta tai -paikasta, saadaan vaivattomasti kaikkien sitä tarvitsevien ulottuville mahdollisimman hyödyllisessä ja helposti omaksuttavassa muodossa. Pääpaino tutkimuksessa on ollut pitkään tekstitiedon tallennuksen ja haun ongelmien tutkimisessa. (Järvelin 1995, 25.) Muita tiedonhaun tutkimuksen osa-alueita ovat muun muassa ääneen perustuva tiedonhaku (kuten puhe- ja musiikkitiedonhaku) ja kuvan sisällöllisiin piirteisiin perustuva tiedonhaku.

Järvelinin (1995, 20) määritelmän mukaan tiedonhakujärjestelmä on tietoyksiköiden tallentamiseen, etsintään, jälleenhakuun ja jakeluun käytettävä järjestelmä. Määritelmä kattaa kaikenlaiset tiedostot riippumatta niiden sisällöstä. Tiedonhakujärjestelmän palauttama hakutulos voi koostua muun muassa elektronisesta tekstistä, kuvista tai viitteistä.

2.2 Dokumentti

Dokumentti on tietovälineen ja siihen tallennetun tiedon muodostama, asiasisällöltään rajattu kokonaisuus (Tietohuollon sanasto 1993, 13). Dokumentti on siis kooste tekijänsä ideoista ja kehittelyistä sekä tietoväline, jossa ne esitetään. Dokumentteihin kuuluvat painetut julkaisut, kuten kirjat ja lehdet sekä digitaaliset julkaisut, joihin lasketaan kuuluviksi muun muassa äänitteet ja multimediatallenteet. (Vakkari 1999, 17.)

Järvelinin (1995, 9) mukaan dokumentti rakentuu sisällöstä, loogisesta rakenteesta (esimerkiksi kirjan pää- ja alaluvut sekä teksti, elokuvan kohtaukset, sinfonian rakenne) sekä ulkoasusta (esimerkiksi kirjasinmallit ja -koot, väri- ja mustavalkofilmi, soitettu esitys). Dokumentin tarkoitus on olla

yhtenäinen, kerralla välitettävä yksikkö. Dokumentit voidaan ryhmitellä niiden tietotyypin (teksti, kuva, ääni, multimedia), rakenteen (romaani tai tutkimusraportti, sinfonia tai iskelmä), sisällön (kirjallisuuden genre) ja käyttötarkoituksen (uutiset, tutkimus, taide, viihde, mainos) mukaan monenlaisiin ryhmiin. Yksi dokumentti voi periaatteessa sisältää useita eri tietotyyppejä tai koostua vain yhdestä tietotyypistä. (Järvelin 1995, 9–11.)

Tiedonhakujärjestelmissä dokumentit on yleensä tallennettu elektroniseen muotoon. Elektroniseen dokumenttiin kuuluu aina dokumentin sisäinen ja ulkoinen esitysmuoto. Ulkoinen esitysmuoto on tarkoitettu ihmisen aistien havaittavaksi ja tarkasteltavaksi. Sisäinen esitysmuoto on puolestaan tarkoitettu tietokoneen käsiteltäväksi ja tietokoneiden välillä siirreltäväksi. Elektronisen dokumentin sisäinen esitysmuoto kattaa sekä dokumentin sisällön, rakenteen että ulkoasun. Näin ollen dokumentin tekijän luoma alkuperäinen sisältö säilyy mahdollisimman hyvin tekijältä vastaanottajille. Samalla elektronisten dokumenttien rakenteen ja ulkoasun automaattinen muokkaaminen tulee mahdolliseksi. (Järvelin 1995, 9–11.) Yksi tapa elektronisen dokumentin sisäisen esitysmuodon määrittelemiseen on XML -kielen käyttö. XML-kielen käytöstä kerrotaan tarkemmin kappaleessa 2.5.

2.3 Hakuavain

Hakuavaimella tarkoitetaan avainta, jonka perusteella tietty asia, esimerkiksi tietokantahaun kohde, tunnistetaan haun kohderyhmään kuuluvaksi. Avainsana viittaa siis haun kohteena olevan tiedoston tai tietueen sisältöön (Tietotekniikan sanasto 1990, 172). Tiedonhaun kirjallisuudessa ja keskusteluissa käytetään sanoja hakusana, hakuavain ja hakutermi useissa eri merkityksissä, ilman järkevää logiikkaa. Järvelinin (1993) tekee kuitenkin selvän eron näiden sanojen käytössä. Hakusanat ovat luonnollisen kielen yksittäisiä sanoja tai yhdyssanoja. Hakutermit ovat puolestaan dokumentaatiokielen tai muun tietyn erityiskielen termejä. Ne voivat olla sanaperusteisia (esimerkiksi tietoliikenneprotokolla) tai koodiperusteisia (esimerkiksi X500 -protokolla). Hakuavain on sen sijaan yleisnimitys luonnollisen kielen hakusanoille, yleiskielessä esiintyville lyhenteille ja koodeille sekä hakutermeille silloin kun erottelua näihin luokkiin ei tarvita. (Järvelin 1993, 125–126.) Tässä tutkielmassa puhutaan jatkossa hakuavaimista koska katson sen olevan sopiva yleisnimitys tutkimusaineistossa käytetyille luonnollisen kielen hakusanoille sekä muutamille käytetyille lyhenteille.

2.4 Kysely

Kysely on tiedonhakijan mielessä olleen tiedontarpeen muotoilu sellaiseen muotoon, että tiedonhakujärjestelmä ymmärtää sen. Tiedonhakujärjestelmä lukee vain merkkejä, ei inhimillisiä ajatuksia, joten on käytettävä järjestelmän omaa hakukieltä (Alaterä & Halttunen 2002, 34–35). Hakukieli koostuu hakuaihetta kuvaavista hakuavaimista sekä mahdollisesti niitä yhdistävistä operaattoreista (Iivonen 1995, 9). Boolen perusoperaattorit: and, or ja not ovat ehkä tunnetuimpia ja käytetyimpiä operaraattoreita.

Tämän tutkimuksen kyselyissä hakijat eivät voineet käyttää erillisiä operaattoreita. Hakijoiden oli mahdollista muodostaa yksinkertaisia kyselyjä, jotka koostuivat joko yksittäisistä sanoista tai lainausmerkeillä erotetuista fraaseista. Hakijoilla oli kuitenkin mahdollista käyttää + ja - merkkejä hakuavainten edessä, halutessaan lisätä tai vähentää niiden suhteellista merkitystä kyselyssä.

Tässä tutkielmassa puhutaan lähtökyselyistä tarkoitettaessa hakijan ensimmäiseksi muotoilemaa kyselyä. Hakijan muokkaamia, lähtökyselyä seuraavia kyselyitä kutsutaan puolestaan jatkokyselyiksi. Kyselyjen muokkaamisesta kerrotaan tarkemmin luvussa 5.1.

2.5 XML-tiedonhaku

XML-kieli (eXtensible Markup Language) on joukko sääntöjä, joiden avulla voidaan suunnitella tekstiformaatteja, jotka sallivat rakenteisen tiedon esittämisen. XML-kieli perustuu tekstin rakenteen kuvaamiseen tunnisteiden (tag) ja niiden sisältämien attribuuttien avulla. (W3C 2003.) XML-dokumentti koostuu siis yhdistelmästä tekstiä ja dokumentin elementtejä määritteleviä tunnisteita. XML-dokumentti rakentuu sisäkkäisistä elementeistä, joita alku ja lopputunnisteet rajaavat. (Desmarais 2000, 11–12.) XML-dokumentti voidaan siis nähdä rakenteena, jonka osia voivat olla muun muassa otsikko, tekijämerkintä ja joukko lukuja, joilla voi olla alalukuja. Alaluvut puolestaan voivat koostua kappaleista, taulukoista ja kuvista. (Järvelin 1995, 91.) Kuviossa 1. on esitetty miltä yksinkertainen XML-dokumentti voisi näyttää.

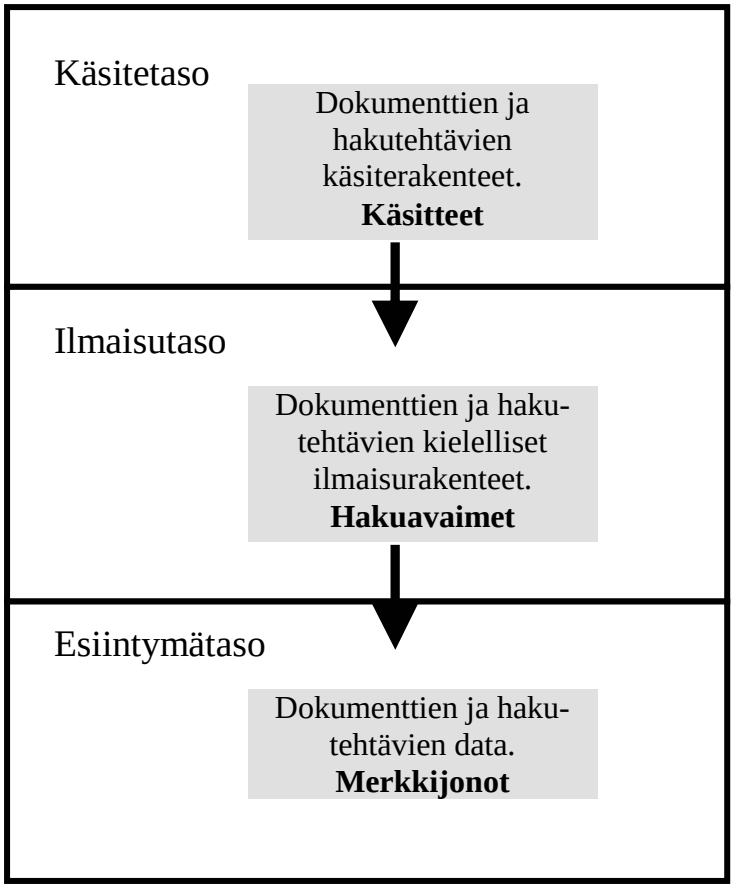
```
<?xml version=”1.0”?>
    <book>
        <author>Norman Desmarais
        </author>
        <title>The ABCs of XML
        </title>
        <place of publication>Houston
        </place of publication>
        <publisher>New Technology Press
        </publisher>
        <date of publication>2000
        </date of publication>
        <ISBN>0-9675942-0-0
        </ISBN>
    </book>
```

Kuvio 1. Esimerkki yksinkertaisesta XML-dokumentista.

Tiedonhaun tutkimuksen näkökulmasta katsottuna rakenteellisten XML dokumenttien käytöstä on hyötyä käyttäjille. XML-kielen käyttö tiedonhakujärjestelmissä mahdollistaa dokumenttien loogisen rakenteen hyödyntämisen (INEX 2004). Aikaisemmin tiedonhaun tulokset koostuivat pääsääntöisesti kokonaisista dokumenteista. XML-perustaisen tiedonhaun myötä kokonaisten dokumenttien sijaan on käyttäjille mahdollista palauttaa haun tuloksena dokumenttien osia eli elementtejä. Näin ollen järjestelmän on mahdollista palauttaa hakijalle se osa dokumenttia, joka parhaiten vastaa hänen tiedontarvetta. (Gövert, Kazai, Fuhr & Lalmas 1993, 1.) Dokumenttien rakenteisuus mahdollistaa lisäksi uudenlaisten hakutoimintojen lisäämisen tiedonhakujärjestelmiin. Haut voidaan kohdistaa dokumenttien tiettyihin elementteihin kuten dokumentin otsikkoon tai tiivistelmään (Liu, Zou & Chu 2004, 88).

3 Käsite, ilmaisu- ja esiintymätaso tiedonhaussa

Informaatiota eli kommunikoitavaksi tarkoitettua tietämyksen osaa voidaan tarkastella käsiterakenteena. Käsiterakenne koostuu käsitteistä ja niiden välisistä suhteista. Kun tämä käsiterakenne halutaan välittää eteenpäin, se pitää ilmaista luonnollisen kielen avulla tekstinä. Tämä erittely pätee myös hakutehtävien muotoiluun. Dokumentteja ja hakutehtäviä tarkastellessaan hakujärjestelmä ei pysty käsittelemään luonnollista kieltä eikä käsitteitä vaan ainoastaan dataa eli merkkijonoja. (Järvelin 1995, 68.) Dokumentteja ja tiedontarpeita voidaan siis tarkastella kolmella tasolla, käsite-, ilmaisu- ja esiintymätasolla (ks. kuvio 2). Teknisesti ajateltuna dokumentit koostuvat merkkijonoista (esiintymätaso). Kirjoitusmerkeillä esitetään luonnollisen kielen ilmaisuja (ilmaisutaso). Luonnollisen kielen ilmaisut puolestaan edustavat dokumentin käsitteellistä sisältöä (käsitetaso). Myös tiedontarpeella on käsitteellinen sisältö, joka voidaan ilmaista luonnollisella kielellä ja esittää kirjoitusmerkkien avulla. (Järvelin & Sormunen 1999, 124.)



Kuvio 2. Tiedon tallennuksen ja haun tasoperiaate (Järvelin 1995, 69).

Informaation haku suunnitellaan, ainakin periaatteessa tasoperiaatteen mukaisesti. Aluksi käsitetasolla analysoidaan hakupyynnön käsitteet ja niiden väliset suhteet käsitteelliseksi hakusuunnitelmaksi. Seuraavaksi mietitään, kuinka nämä käsitteet pystytään ilmaisemaan ilmaisutason hakusuunnitelmana. Viimeiseksi muotoillaan kysely, jossa huomioidaan kirjoitusasut, katkaisut ja läheisyysoperaatiot. (Järvelin 1993, 125.)

4 Vuorovaikutteinen tiedonhaku

4.1 Järjestelmäkeskeisestä tiedonhausta vuorovaikutteiseen tiedonhakuun

Tiedonhaun tutkimuksen alkuaikoina pääpaino oli pitkään tiedonhakujärjestelmien tehokkuuden mittaamisessa. Tiedonhaun järjestelmäkeskeisessä evaluoinnissa pääpaino on perinteisesti ollut kyselyjen ja saatujen elementtien täsmäämisen tutkimisessa sekä käytetyn järjestelmän suorituskyvyn mittaamisessa. (Rieh & Xie 2005, 752; Robertson & Hancock-Beaulieu 1992, 458.) Järjestelmien evaluointi on suoritettu tyypillisesti laboratorio-olosuhteissa ja koeasetelma on koostunut dokumenttikokoelmasta, sarjasta kyselyitä ja itsenäisistä relevanssiarvioista (Borlund 2000, 74). Järjestelmäkeskeisen tiedonhaun tutkimuksen puutteena on pidetty sitä, ettei se huomioi todellisia käyttäjiä ja eikä näin ollen anna kovinkaan realistista kuvaa tiedonhakutilanteesta. 1990 -luvulta lähtien kiinnostus on yhä kasvavassa määrin siirtynyt tiedonhaun vuorovaikutteisen luonteen tutkimiseen. Vuorovaikutteisen tiedonhaun tutkimus on perinteisen lähestymistavan lisäksi kiinnostunut käyttäjän, tiedontarpeen ja tiedonhakujärjestelmän vuorovaikutteisesta suhteesta. Koko tiedonhakuprosessi nähdään siis vuorovaikutteisena tapahtumana, jossa vuorovaikutteisuutta ilmenee muun muassa käyttäjän muotoillessa kyselyitä, tulkitessa saatuja dokumentteja ja informaatiota käytettäessä. Kaikki nämä toiminnot ovat riippuvaisia kulloisenkin käyttäjän tavoitteista, päämääristä, tietämyksen tilasta ja tiedontarpeen tilanteesta. (Rieh & Xie 2005, 753.)

Borlund (2000; 2003b) on keskittynyt tutkimuksissaan vuorovaikutteisten tiedonhakutilanteiden evaluoimiseen. Borlund pyrkii kokeillaan pääsemään mahdollisimman lähelle todellista tiedonhakutilannetta säilyttämällä kuitenkin suhteellisen kontrolloitu testiympäristö. Kontrolloitu testiympäristö on säilytettävä jotta tutkimustulokset ovat vertailukelpoisia eri testihenkilöiden sekä tiedonhakujärjestelmien välillä. Tällainen vuorovaikutteinen koetilanne on Borlundin mukaan mahdollista saavuttaa kolmen osatekijän avulla. Nämä osatekijät ovat: (1) potentiaalisten käyttäjien käyttäminen testihenkilöinä, (2) dynaamisten ja yksilöllisten tiedontarpeiden sekä (3) moniulotteisten ja dynaamisten relevanssiarvioiden käyttäminen koetilanteissa. Dynaamisella tiedontarpeella Borlund viittaa tiedontarpeen kehittymiseen ja muuttumiseen tiedonhakuprosessin edetessä. Simuloidut työtehtävät ovat Borlundin vastaus mahdollisimman realistisen tiedontarvetilanteen luomiseksi koetilanteissa. Myös käsitys relevanssista voi muuttua ajan ja tilanteen mukaan. (Borlund 2000, 72–75.)

Myös INEX-hankkeen vuorovaikutteisen tiedonhaun tutkimuslinjassa on pyritty luomaan vuorovaikutteisuutta käyttäjien ja järjestelmän välillä muun muassa hakutehtävien ja relevanssiarvioiden avulla. Seuraavissa kappaleissa käsitellään vielä tarkemmin sekä simuloituja työtehtäviä että relevanssin käsitettä.

4.2 Simuloidut työtehtävät

Borlundin (2000, 80) mukaan simuloitu työtehtävä on ikään kuin kehyskertomus, joka kuvailee tiedontarvetilanteen. Samalla se tarjoaa mahdollisuuden kontrolloida yhdenmukaista tiedonhakutilannetta kaikille koehenkilöille. Tämän kontrollin ansiosta testitulokset sekä useiden koehenkilöiden että eri tiedonhakujärjestelmien välillä ovat vertailukelpoisia. Kuten todellisessa elämässä, simuloidun työtehtävän tarkoituksena on herättää koehenkilölle tiedontarve, joka täytyy tyydyttää, jotta tilanteesta voidaan siirtyä eteenpäin. Realismia lisää se seikka, että kukin koehenkilö kehittää annetun työtehtävän herättämän tiedontarpeen yksilöllisesti ja omakohtaisesti. Jokainen koehenkilö suorittaa myös omalla tavallaan tiedonhaut ja määrittelee saatujen dokumenttien relevanttiuden suhteessa heidän omaan, työtehtävän luomaan tiedontarpeeseensa. Simuloitu työtehtävä tilanne palvelee siis kahta päätarkoitusta: se sallii käyttäjän oman tulkinnan tiedonhakutilanteesta, mikä johtaa yksilöllisiin tiedontarpeen tulkintoihin ja se toimii pohjana sille, kuinka koehenkilöt arvioivat saatujen dokumenttien relevanttiutta. (Borlund 2000, 80–83.) Kuviossa 3. esitellään, miltä simuloitu työtehtävä voisi näyttää.

***Simulated work task situation**: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.*

***Indicative request:** Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.*

Kuvio 3. Esimerkki simuloidusta työtehtävästä. (Borlund 2000, 81; 2003b)

Simuloitu työtehtävä tilanne on siis melko avoin kuvaus tiedontarpeesta. Tarkoituksena on välttää tarkkaan määrättyä hakutehtävää. Tämän tiedontarpeen pohjalta koehenkilöt muodostavat kyselyt ja syöttävät ne tiedonhakujärjestelmään.

Simuloidun työtehtävän mallia on käytetty monissa vuorovaikutteisen tiedonhaun tutkimuksissa. (ks. Esim. Borlund & Ingwersen 1997; Borlund 2000; Ruthven, Lalmas & Rijsbergen 2003; Suomela & Kekäläinen 2005.) Simuloitujen työtehtävien käytön on todettu tuovan realistisuutta tiedonhakutilanteeseen. Tutkimustulokset ovat osoittaneet simuloitujen työtehtävien vastaavan hyvin hakijoiden todellisia tiedontarpeita. Myös tämän tutkimuksen käyttäjätesteissä käytettiin hyväksi simuloidun työtehtävän mallia.

4.3 Relevanssin käsitteestä

Tiedonhaun tavoitteena on löytää mahdollisimman paljon relevanttia informaatiota sitä tarvitsevalle. Useimmiten tämä tapahtuu relevanttien dokumenttien muodossa (Borlund 2000, 25). Tavoite kuulostaa yksinkertaiselta mutta kuinka voimme päättää, mikä dokumentti on relevanssi ja ennen kaikkea, kuka tämän päätöksen voi tehdä ja millä perusteilla. Tällaiset kysymykset ovat usein keskustelun aiheena tiedonhaun tutkimuksen alalla.

Relevanssi on luonteeltaan hyvin moniulotteinen ja dynaaminen. Moniulotteisuudella tarkoitetaan sitä, että eri käyttäjät sekä tutkijat arvioivat ja ymmärtävät relevanssin eri tavoin. Dynaamisuudella puolestaan viitataan siihen, kuinka käsitys relevanssista voi muuttua ajan ja tilanteen mukaan samojen käyttäjien ja tutkijoiden keskuudessa. (Borlund 2003a, 914.)

1950 -luvulta lähtien on tiedonhaun tutkimuksen ja kirjallisuuden piirissä käyty keskustelua ja väittelyä relevanssin käsitteestä (Borlund 2003a, 913). Näiden keskustelujen pohjalta on noussut kaksi pääsuuntausta relevanssin määrittelemiseksi. Nämä kaksi suuntausta ovat relevanssin jakaminen aiherelevanssiin (topical relevance) ja käyttäjärelevanssiin (user relevance). Aiherelevanssin pääajatus on, että dokumentti on relevantti, mikäli se käsittelee hakupyynnön määrittelemää aihetta. Tämä tarkoittaa käytännössä sitä, että kyselyjen hakuavaimet täsmäävät dokumentissa esiintyviin sanoihin. Aiherelevanssista käytetään usein myös nimityksiä objektiivinen tai järjestelmäkeskeinen relevanssi koska tämä relevanssin määritelmä ei huomioi käyttäjien näkemyksiä dokumenttien relevanttiudesta. Relevanssin tutkimuksen piirissä aiherelevanssi oli pitkään vallitseva suuntaus, koska hakuavainten ja sanojen täsmäämisen havainnointi ja mittaaminen on kohtalaisen helppoa. Vuosien mittaan painotus aiherelevanssista on kääntynyt kohti käyttäjärelevanssia. Käyttäjärelevanssi huomioi aiheenmukaisuuden lisäksi tiedon käyttäjästä riippuvia tekijöitä. Tällaisia tekijöitä ovat muun

muassa tiedontarpeen luonne, dokumentin kieli, ulkoasua ja tuttuus käyttäjälle. (Cosijn & Ingwersen 2000, 538–539.) Käyttäjärelevanssi siis kuvailee suhdetta informaation, käyttäjän sekä käytössä olevan järjestelmän välillä. Aluksi tiedonhaun tutkimuksen piirissä ajateltiin, että käyttäjärelevanssin mittaaminen on mahdotonta koska, tarkastelun kohteena ovat käyttäjien sisäiset ajatukset ja käsitykset, joita on hyvin vaikea tutkia. Nykyään kuitenkin uskotaan, että ainakin joitakin näkökulmia käyttäjärelevanssista voidaan mitata luotettavin keinoin. (Schamber, Eisenberg & Nilan 1990, 755)

Vaikka aiherelevanssi ja käyttäjärelevanssi ovat olleet pitkään keskipisteenä relevanssin tutkimuksessa, on tätä kahtiajakoa kuitenkin pidetty liian suppeana määritelmänä relevanssille. Relevanssin määrittelyssä on haluttu tuoda esiin useampia näkökulmia. Yksi tunnetuimmista relevanssin määritelmistä on Saracevicin (1996, 214) luoma relevanssin jaottelu viiteen eri luokkaan, nämä luokat ovat seuraavat:

1. Algoritminen relevanssi (System/Algorithmic relevance)
2. Aiherelevanssi (Topical relevance)
3. Kognitiivinen relevanssi (Cognitive relevance/Pertinence)
4. Tilannerelevanssi (Situational relevance)
5. Affektiivinen relevanssi (Motivational/Affective relevance)

Saracevicin luokista kolme ensimmäistä eli algoritminen relevanssi, aiherelevanssi ja kognitiivinen relevanssi noudattavat ajatusta relevanssin jaottelusta aihe- ja käyttäjärelevanssiin. Algoritminen relevanssi kuvailee kyselyssä esiintyvien hakuavainten suhdetta dokumentin sanoihin, järjestelmän suorittaman täsmäytyksen perusteella. Aiherelevanssi puolestaan luonnehtii kyselyn aiheen ja dokumentin aiheen suhdetta käyttäjän suorittaman arvion perusteella. Kognitiivinen relevanssi viittaa käyttäjän tiedontarpeen ja dokumentin suhteeseen. Käyttäjä arvioi muun muassa dokumentin informatiivisuutta ja laatua oman tietämyksentilansa pohjalta. Tiedonhaun tutkimuksessa mielenkiinto on viimeaikoina kääntynyt kahta jäljelle jäävää luokkaa eli tilannerelevanssia ja affektiivista relevanssia kohti. Tilannerelevanssi merkitsee, että käyttäjällä on jokin tilanne, ongelma tai tehtävä, jota hän yrittää ratkaista. Dokumentin relevanttius riippuu siitä, kuinka hyödyllinen dokumentti on tämän tehtävän tai ongelman ratkaisemiseksi. Tilannerelevanssi siis arvioi dokumentin hyödyllisyyttä käyttäjän kannalta. (Cosijn & Ingwersen 2000, 537–541.) Tilannerelevanssi on erittäin kontekstiriippuvainen ja relevanssin dynaaminen luonne tulee hyvin esiin. Tietty ongelma tai tehtävä kehittyy ja muuttuu jatkuvasti sitä mukaa, kun käyttäjän tietämys halutusta aiheesta kasvaa. Viimeisin relevanssin luokista eli affektiivinen relevanssi heijastaa käyttäjän aikomuksia, tavoitteita ja motivaatiota. Toisin sanoen aikomus, tavoite ja motivaatio ovat syy siihen, miksi käyttäjä etsii infor-

maatiota, jatkaa tiedonhakua ja arvioi saatujen dokumenttien relevanttiutta. (Borlund 2003a, 915.) Affektiivisen relevanssin näkökulmasta dokumentti on relevantti, mikäli käyttäjä tuntee tyydytystä ja onnistumista tietyn dokumentin takia (Cosijn & Ingwersen 2000, 541).

4.4 Vuorovaikutteisen tiedonhaun tutkimuksia

Perinteistä tiedonhaun tutkimusta varten on vuodesta 1992 lähtien järjestetty TREC-konferensseja (Text Retrieval Conference). TREC-hanketta organisoivat tahot ovat NIST (National Institute of Standars and Technology) sekä Yhdysvaltain hallituksen puolustuslaitoksen tutkimus ja kehitysosasto DARPA (Defense Advanced Research Projects Agensy). TREC-hankkeen päätarkoitus on ollut alun perin tukea tekstitiedonhaun tutkimusta tarjoamalla tarvittava infrastruktuuri tiedonhakumenetelmien evaluoimiseksi ja kehittämiseksi teollisuuden ja akateemisten tutkimuslaitosten tarpeisiin. TRECin tavoitteita ovat lisäksi olleet yhteistyön ja kommunikaation lisääminen eri alojen tiedeyhteisöjen keskuudessa, uusien evaluointimenetelmien kehittäminen tiedonhaun tutkimuksen tarpeisiin sekä kehitettyjen teknologioiden jalostamisen nopeuttaminen valmiiksi tuotteiksi. (Text REtrieval Conference 2000.)

TREC-konferenssit koostuvat tutkimuslinjoista (tracks), joihin on vuodesta 1997 lähtien kuulunut vuorovaikutteisen tiedonhaun tutkikimuslinja (Interactive Track). Interaktiivisen tutkimuslinjan tarkoituksena on tarkastella tiedonhaun vuorovaikutteista luonnetta, tutkimalla sekä tiedonhakuprosesseja että järjestelmän palauttamia hakutuloksia. (Text REtrieval Conference 2000.)

XML-kielen käytön yleistyminen erityisesti tieteellisen tiedon säilyttämisessä, digitaalisissa kirjastoissa ja internetissä on aiheuttanut valtaisan kiinnostuksen XML-työkalujen kehittämiseksi. Tiedonhaun tutkimusta ja evaluointia varten XML-ympäristössä on perustettu INEX-hanke, joka tulee sanoista INitiative for the Evaluation of XML Retrieval (INEX 2004). Vuosittain tehtäviä INEX-tutkimuksia on tehty vuodesta 2002 lähtien. Tiedonhaun eri osa-alueita tutkitaan INEX-hankkeessa erilaisten tutkimuslinjojen avulla, joita on vuodesta 2004 ollut viisi. Nämä viisi tutkimuslinjaa ovat Ad hoc retrieval track, Interactive track, Heterogeneous collection track, Relevance feedback track ja Natural language track. Vuorovaikutteinen tutkimuslinja otettiin INEX hankkeeseen mukaan vuonna 2004 ja sen tavoitteena on tutkia hakijoiden käyttäytymisistä kun haun kohteena on XML-dokumentteja. Tavoitteisiin kuuluu lisäksi tutkia ja kehittää menetelmiä XML-tiedonhaun tarpeisiin (Fuhr, Lalmas, Malik & Szlavik 2004, 7.) Tutkielmani empiirinen aineisto tulee koostumaan osasta

INEX 2004 aineiston vuorovaikutteisen tiedonhaun tutkimuslinjan (Interactive trac) aineistoa. Kappaleessa 6.2 esitellään tarkemmin INEX 2004 Interaktiivisen tutkimuslinjan aineisto.

5 Tiedonhakukäyttäytyminen

5.1 Kyselyjen muokkaaminen

Kyselyn muokkaaminen (Query modification/reformulation) tarkoittaa kyselyn uudelleen muotoilua muuttamalla hakuavaimia, jotta päästäisiin parempaan hakutulokseen. Ensimmäisen kyselyn muodostaminen toimii usein ensiaskeleena tiedonhakujärjestelmän pariin ja sitä seuraa luonnollisesti tulosten selailu ja alkuperäisen kyselyn muokkaaminen. Kyselyjen muotoilua ja uudelleen muotoilua on tutkittu tiedonhaun tutkimuksen piirissä paljon, sillä hyvien hakuavainten valitseminen on vaikeaa ja samalla erityisen tärkeää hyvien hakutulosten saavuttamiseksi. (Ingwersen & Järvelin 2005, 140.)

Kyselyjen uudelleen muotoilua tarvitaan Bruzan ja Dennisin (1997, 489) mukaan ainakin kahdesta eri syystä. Ensinnäkin hakijoiden mielessä olevat tiedontarpeet ovat usein hyvin täsmällisiä ja tästä syystä käyttäjillä voi olla vaikeuksia ilmaista näitä tiedontarpeita hakukielen vaatimassa muodossa. Hakijan on osattava valita, mikä ilmaisu kuvaa kaikista parhaiten tiedontarvetta suhteutettuna käytössä olevaan järjestelmään. Hakijan on usein tarpeellista täsmentää hakuavaimia, jotta haun tuloksesta pystyttäisiin pudottamaan pois sellaiset dokumentit, jotka eivät käsittele haluttua aihetta. (Bruza & Dennis 1997, 489) Esimerkiksi haettaessa tietoa hiiristä lemmikkieläiminä ja hakuavaimeksi valitaan ”mouse”, on järkevää sulkea tuloksista pois dokumentit jotka käsittelevät tietokoneiden osoitinlaitteita. Toiseksi, kyselyjen uudelleenmuotoilua tarvitaan, koska tiedontarve saattaa muuttua tiedonhakuprosessin edetessä (Bruza & Dennis 1997, 489). Esimerkiksi hakuavain ”bowling” saattaa johtaa hakijan tarkastelemaan paikkoja, joissa keilausta voi harrastaa. Tämän jälkeen hakija saattaa täsmentää kyselyään hakemalla oman alueensa keilahalleja.

Ingwersen ja Järvelin (2005, 140) toteavat, etteivät hakijoiden muodostamat kyselyt useinkaan sisällä parhaita mahdollisia ilmauksia halutusta tiedontarpeesta. Kyselyt ovat myös tyypillisesti hyvin lyhyitä. Tämän vuoksi kyselyjen muokkaamista tarvitaan. Järvelin (1995, 226) lisää vielä, että kyselyjen muokkaamista tarvitaan, koska kyselyt yleensä tuottavat joko liian vähän tai liian paljon dokumentteja. Tällöin kyselyä pitää joko laajentaa tai kaventaa.

5.1.1 Kyselyn muokkaamisen menetelmät

Useat tiedonhaun tutkijat ovat luokitelleet erilaisia muunnoksia, joita hakijat tyypillisesti tekevät kyselyjä muokatessaan. Yleisesti käytetty termi on myös hakutaktiikka, jota käyttävät muun muassa Fidel (1984, 1985), Bates (1987) ja Iivonen (1995). Iivonen (1995, 33) määrittelee hakutaktiikan seuraavasti: *”Hakutaktiikka on tiedonhaun aikana toteutettu yksi tai useampi siirto tiedonhaun jatkamiseksi eteenpäin. Tiedonhaun aikana toteutettu siirto on tunnistettavissa oleva ajatus tai toiminta, joka on osa tiedonhakua. Yksi hakustrategia voi sisältää useita hakutaktiikoita ja hakutaktiikat ovat osa hakustrategiaa.”*. Hakustrategialla tarkoitetaan tässä yhteydessä kokonaissuunnitelmaa haun suorittamiseksi.

Bates (1979, 205; 1987, 47–54) puhuu hakutaktiikoista siirtoina, joita hakijat käyttävät päästäkseen tiedonhaussa eteenpäin. Bates nimeää kaiken kaikkiaan 29 hakutaktiikkaa. Hakutaktiikoita voidaan ryhmitellä sen mukaan, mihin hakuprosessin vaiheeseen ne kuuluvat ja millaisia toimintoja ne edustavat. Osa hakutaktiikoista voi olla päätetty jo ennen haun aloittamista ja osa taktiikoista valitaan vasta hakuprosessin kuluessa. Tiedonhakuprosessin alkuvaiheessa käytettäviä hakutaktiikoita ovat muun muassa aihetta koskevien bibliografioiden ja muiden luetteloiden tutkiminen sekä mahdollisten hakuavainten etsiminen alan sanastoista. Hakuprosessin kuluessa käytettävät hakutaktiikat liittyvät usein kyselyjen muokkaamiseen tai hakuavainten tarkastamiseen. Tällaisia hakutaktiikoita ovat esimerkiksi kyselyn laajennus ja supistaminen sekä hakuavainten käyttöön liittyen synonyymien sekä hierarkkisesti laajempien, suppeampien tai rinnakkaisten hakuavainten huomioiminen. (Bates 1979, 205; 1987, 47–54.)

Bruza ja Dennis (1997) selvittivät millaisia toimenpiteitä hakijat tekevät muokatessaan alkuperäisiä kyselyitä. Bruza ja Dennis tutkivat manuaalisesti 1040 kyselyä ja jakoivat ne 11 eri luokkaan sen perusteella, millaista toimenpidettä hakijat käyttivät kyselyitä muokatessaan (ks. taulukko 1).

	Tapahtunut muutos	Esimerkki
SPL	Hakuavaimen jakaminen tai yhdistäminen	rockclimb → rock climb, centre point → centrepoint
DEL	Hakuavaimen poistaminen	malaysia electricity → malaysia
ADD	Hakuavaimen lisääminen	windows95 → windows95 help
REP	Kyselyn toisto	
SUB	Hakuavaimen korvaaminen semanttisesti läheisellä hakuavaimella	electronic commerce → electronic contract
DER	Johdos hakuavaimesta	tourism → tour
SPE	Kirjoitusasun korjaaminen	
ABR	Lyhennyksen käyttämien tai sen purkaminen	jpl → jet propulsion laboratories
PUN	Välimerkin käyttö kuten tavutuksen lisääminen tai poistaminen	hitch-hikers guide → hitchhikers guide
CAS	Tilanteen muuttuminen	
MIS	Jokin muu muutos	

Taulukko 1. Kyselyn muokkauksen lajit (Bruza & Dennis 1997, 490)

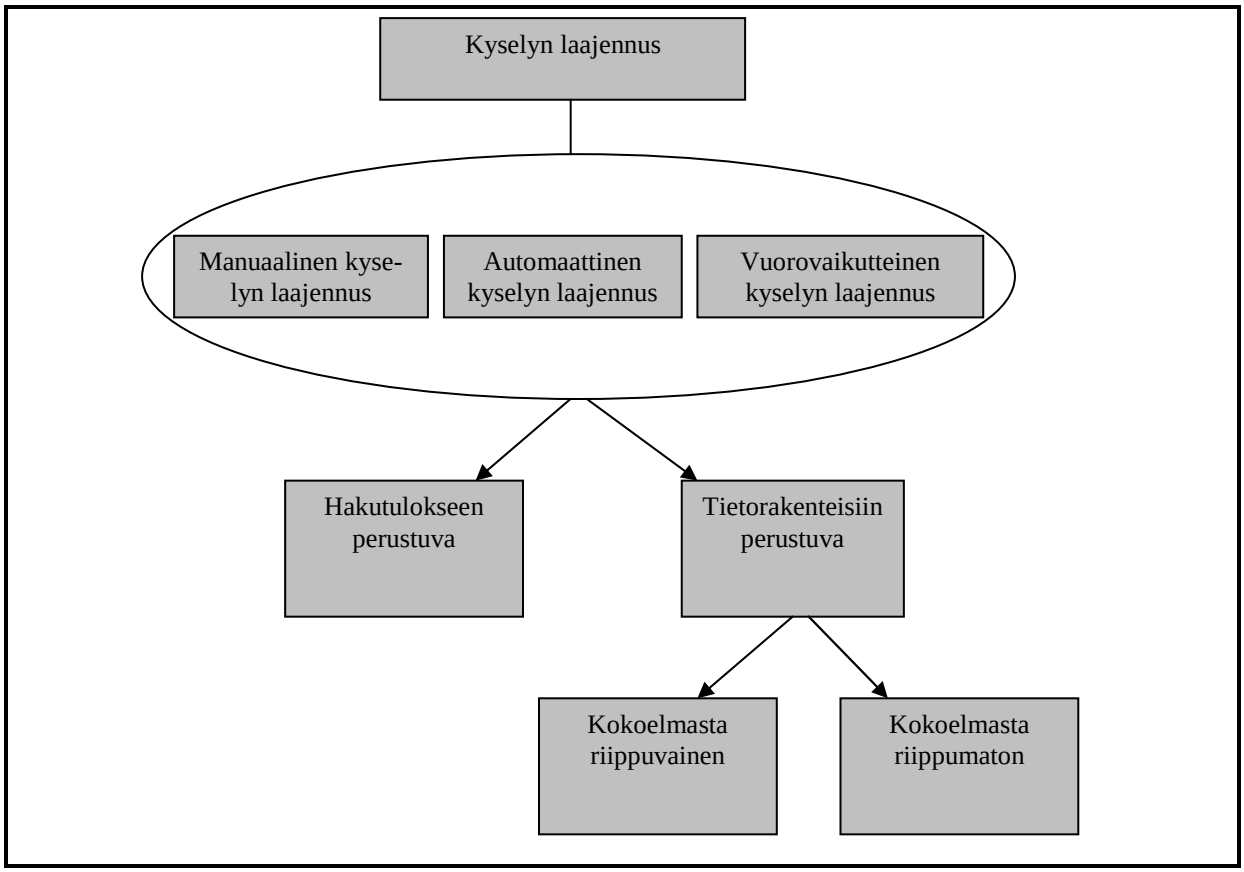
Tutkimustulokset osoittivat kaikkein käytetyimmäksi kyselyn muokkaamisen lajiksi sellaisen kyselyn toistamisen, jonka hakija on jo aiemmin suorittanut. Seuraavaksi käytetyimmät menetelmät olivat hakuavaimen korvaaminen läheisessä semanttisessa suhteessa olevalla hakuavaimella, hakuavaimen lisääminen ja hakuavaimen poistaminen. Hakuavaimen lisäämistä käytettiin, mikäli kyselyä haluttiin tarkentaa ja vastaavasti hakuavain poistettiin, jos kyselystä haluttiin yleisluontoisempi. Muita harvemmin käytettyjä menetelmiä olivat: hakuavainten jakaminen osiin tai yhdistäminen, kirjoitusasun korjaaminen, välimerkin lisääminen tai poistaminen, uusien hakuavainten johtaminen alkuperäisistä hakuavaimista sekä lyhennysten käyttäminen tai niiden purkaminen. (Bruza & Dennis 1997.)

Tämän tutkimuksen kannalta ollaan erityisen kiinnostuneita sellaisista kyselyn muokkaamisen menetelmistä, jotka liittyvät uusien hakuavainten lisäämiseen alkuperäiseen kyselyyn. Tämän vuoksi kyselyn laajennuksen menetelmiä tarkastellaan vielä yksityiskohtaisemmin seuraavassa luvussa.

5.1.2 Kyselyn laajennus

Efthimiadiksen (1996) mukaan kyselyn laajennuksesta on kyse silloin, kun alkuperäistä kyselyä täydennetään vaihtoehtoisilla hakuavaimilla hakutuloksen parantamiseksi. Kyselyä voidaan täydentää kolmella eri tavalla: manuaalisesti, automaattisesti tai vuorovaikutteisesti. Kyselyn laajennuksessa käytettävät hakuavaimet voidaan poimia joko relevanteiksi todetuista dokumenteista, jolloin uudet hakuavaimet perustuvat relevanssipalautteeseen tai ne voidaan poimia hakuprosessin ulko-

puolisista tietorakenteista. Ulkopuolinen tietorakenne voi olla joko dokumenttikokoelmaan kuuluva tai siitä täysin riippumaton. Edellä esitellyt kyselyn laajennus menetelmät on esitetty kuviossa 4. (Efthimiadis 1996)



Kuvio 4. Kyselyn laajentamisen menetelmät ja uusien hakuavainten lähteet (Efthimiadis 1996)

Manuaalisesta kyselyn laajennuksessa on kyse silloin, kun hakija lisää itse alkuperäiseen kyselyyn hakuavaimia, valitsemansa hakustrategin pohjalta. Manuaalisesta kyselyn laajennuksesta käytetään usein myös nimitystä intellektuaalinen kyselyn laajennus. Automaattisella kyselyn laajennuksella tarkoitetaan puolestaan tiedonhakujärjestelmän suorittamaa uusien hakuavainten lisäämistä. Interaktiivisessa kyselyn laajennuksessa tiedonhakujärjestelmä tarjoaa hakijalle vaihtoehtoisia hakuavaimia, joista hakija voi valita haluamiaan ja täydentää niillä alkuperäistä kyselyä. (Efthimiadis 1996.)

5.1.2.1 Hakutulokseen perustuva kyselyn laajennus

Relevanssipalautteeseen pohjautuvasta kyselyjen muokkaamisesta on kyse silloin, kun muokkaaminen tapahtuu tuloslistan perusteella. Ideana on, että hakija tutkii alkuperäisen kyselyn tarjoaman tuloslistan dokumentteja ja tunnistaa niistä relevantteja ja epärelevantteja dokumentteja. Hakija

tarkkailee näissä dokumenteissa esiintyviä sanoja ja niiden painotuksia ja tätä tietoa hyödyntäen poimii uusia hakuavaimia jatkokyselyihin. Uusia hakuavaimia poimitaan pääsääntöisesti relevanteiksi todetuista dokumenteista. Tutkimustulokset (ks. esim. Efthimiadis 1996; Harman 1992) osoittavat, että ihanteellinen määrä dokumenteista poimituista hakuavaimista vaihtelee muutamasta kappaleesta aina useisiin satoihin.

Efthimiadiksen (1992, tässä Ingwersen & Järvelin 2005, 141–142) tutkimusten mukaan hakijat valitsevat kolmasosan heille tarjotuista relevanssipalautteen pohjalta poimituista hakuavaimista. Tästä hakuavainjoukosta noin kolmannes ei ollut missään yhteydessä alkuperäisiin hakuavaimiin. Lopusta kahdesta kolmasosasta uusia hakuavaimia suurin osa (70%) oli hyponyymeja, hyperonyymeja[1] (5%) tai assosiaatiosuhteessa (25%) alkuperäisiin hakuavaimiin nähden. (Ingwersen & Järvelin, 1995, 141–142.)

Tutkimus tulokset paljastavat myös eroja kokeneiden ja kokemattomien hakijoiden välillä. Kokeneet hakijat pystyvät kokemattomia paremmin ilmaisemaan tiedontarpeensa sopivien hakuavainten muodossa. Kokeneet hakijat pystyvät myös tehokkaammin parantamaan hakutuloksia interaktiivisen relevanssi palautteen pohjalta sekä hyödyntämään tehokkaammin sanastoja ja tesauruksia, niitä tarjottaessa. (Ks. esim. Magennis & van Rijsbergen 1997; Sihvonen & Vakkari 2004.)

5.1.2.2 Tietorakenteisiin perustuva kyselyn laajennus

Tietorakenteisiin perustuvasta kyselyn laajennuksesta on kyse silloin, kun poimitaan uusia hakuavaimia kyselyihin jostakin varsinaisen hakuprosessin ulkopuolisesta tietorakenteesta. Tällainen tietorakenne voi olla joko käytössä olevasta dokumenttikokoelmasta riippuvainen tai riippumaton tietorakenne. Dokumenttikokoelmasta voidaan etsiä taajaan yhdessä esiintyviä avaimia ja ryhmitellä niitä. Kyselyjä laajennetaan täten käytetyn hakuavaimen kanssa samaan ryhmään kuuluvien sanojen avulla. Efthimiadiksen (1996) mukaan todisteita hakuavainten yhteisesiintymiseen perustuvan automaattisen kyselyn laajennuksen merkittävästä tehokkuudesta ei kuitenkaan ole pystytty esittämään. Ongelmana yhteisesiintymiseen perustuvassa kyselyn laajennuksessa on se, että niistä yhteisesiintymisen perusteella poimitut hakuavaimet ovat yleensä liian yleisiä sanoja. Liian yleiset sanat eivät toimi hyvänä erottelijoina relevanttien ja epärelevanttien dokumenttien välillä, joten tällaisten sanojen käyttäminen hakuavaimina tuskin parantaa hakutulosta. (Efthimiadis 1996.)

[1] Hyponymialla tarkoitetaan sanojen hierarkkisten suhteiden määrittelyä. Alisteinen sana on hyponyymi ja yläkategorian sana on hyponyymiin nähden hyperonyymi. Esimerkiksi nisäkäs on sanan hevonen hyperonyymi ja vastaavasti hevonen on nisäkäs sanan hyponyymi. (Karlsson 1998, 221.)

Dokumenttikokoelmasta riippumattomia tietorakenteita ovat puolestaan esimerkiksi tesaurukset ja ontologiat. Perinteisellä tesauruksella tarkoitetaan esitystä, joka koostuu tietyn aihepiirin sanoista ja niiden välisistä semanttisista suhteista (Aitchison, Gilchrist & Bawden 1997, 47–58). Puhuttaessa tesauruksesta tiedonhaun alalla, sillä tarkoitetaan useimmiten tiedontallennuksen- ja haun apuvälinettä, jota on käytetty tiedonhakujärjestelmissä viimeisten 50 -vuoden ajan. Tiedonhaulle tyypillinen ongelma on haun aihetta koskevan terminologian monimuotoisuus. Samaa aihetta käsittelevät termit vaihtelevat tietokannan tekijän, indeksoijan ja käyttäjän välillä. Tiettyä aihetta voidaan lähestyä myös hyvin erilaisista näkökulmista, erilaisilta abstraktiotasoilta ja käyttäen eri ilmauksia samoista asioista. (Lykke Nielsen 2002, 9.)

Tesaurusta, jota käytetään ainoastaan tiedonhakuja tehtäessä, kutsutaan hakutesaurukseksi. Hakutesauruksen tarkoituksena on auttaa tiedonhakua kokotekstitietokannasta ehdottamalla lisätermejä, erityisesti synonyymeja ja suppeampia termejä. (Lykke Nielsen 2002, 3.) Hakutesauruksen (searching thesaurus, end-user thesaurus) tarkoituksena on toimia apuvälineenä muodostettaessa hakulauseita. Perinteisestä tesauruksesta poiketen hakutesauruksen tarkoituksena ei ole kontrolloida ja ohjata käyttämään tiettyjä hakutermejä vaan sitä vastoin tarjota käyttäjän valitsemille hakutermeille relevantteja vaihtoehtoja, joita käyttäjä ei muuten tulisi käyttäneeksi. Hakutesaurus tarjoaa yhteyksiä ja assosiaatioita käsitteiden välillä, joiden tarkoituksena on kasvattaa käyttäjän tietämystä tiedonhaun kohteena olevasta aiheesta. (Lykke Nielsen 2002, 22.)

Gruber (1993, 199) määrittelee ontologian formaaliksi, eksplisiittiseksi määrittelyksi yhteisestä käsitteistöstä. Formaalius ja eksplisiittisyys viittaavat ontologian koneelliseen tulkittavuuteen ja käsitteistön yhteisöllisyys puolestaan mahdollistaa tietämyksen jakamisen. (Gruber 1993, 199.) Dingin ja Foon (2001, 132) mukaan tesauruksen ja ontologian merkittävin ero on se, että ontologian tarkoituksena on luoda yhteys ihmisten ja tietokoneiden välille, kun sen sijaan tesaurus on tarkoitettu ihmisten keskinäisen viestinnän avuksi. Käytännössä tesaurukset ja ontologiat tarkoittavat samaa asiaa tiedonhaun alalla.

Tämän tutkimuksen kyselyitä laajennettaessa koehenkilöt laajensivat kyselyitään manuaalisesti. Tutkimuksen tavoitteena on selvittää, missä määrin on kyse hakutulokseen perustuvasta kyselyn laajennuksesta eli siitä poimitaanko uusia hakuavaimia relevanteiksi todetuista tuloslistan dokumenteista. Muista tietoresursseista peräisin olevien uusien hakukäsitteiden alkuperän selvittäminen jää tämän tutkimuksen ulkopuolelle.

5.2 Tutkimuksia tiedonhakukäyttäytymisestä

Internet on nykyään merkittävä informaationlähde monille ihmisille ympäri maailmaa. Miljoonia kyselyitä suoritetaan internetissä päivittäin. Ihmiset voivat suorittaa tiedonhakuja internetissä useiden erilaisten hakukoneiden avulla. Myös tilastoja internetin käytön määristä julkaistaan säännöllisesti. Nämä tilastot eivät kuitenkaan kovinkaan usein kerro ihmisten tiedonhakukäyttäytymisestä.

Spink, Dietmar, Jansen ja Saracevic (2001) tutkivat laajassa, yli miljoona kyselyä kattaneessa tutkimuksessa kuinka ihmiset käyttäytyvät tehdessään tiedonhakuja internet ympäristössä. Analyysi kattoi yli miljoona kyselyä, yli 200 000 käyttäjän suorittamana. Hakukoneena käytettiin ilmaista boolen menetelmään perustuvaa Excite-hakukonetta. Tutkimustulokset kertoivat seuraavaa kyselyjen määristä: 211,063 hakijaa suoritti yhteensä 1,025, 910 kyselyä yhden päivän aikana. Näistä kyselyistä 51,8% oli uniikkeja (unique queries), 38,5% oli jo aiemmin suoritettujen kyselyjen toistoja (repeat queries) ja 9,7% oli kyselyitä, jotka eivät sisältäneet yhtäkään hakuavainta (zero queries). Hakijat käyttivät keskimäärin 2,4 hakuavainta yhtä kyselyä kohti. Hakijat suorittivat keskimäärin 4,86 kyselyä yhden hakuistunnon aikana. Uniikkeja kyselyitä suoritettiin keskimäärin 2,52. Hieman alle puolet hakijoista (48,4%) suoritti ainoastaan yhden kyselyn hakuistunnon aikana, 20,8% hakijoista suoritti kaksi kyselyä ja loput 31% kolme kyselyä tai siitä ylöspäin. Koska yli puolet hakijoista suoritti enemmän kuin yhden kyselyn, Spink ym., kiinnostuivat siitä, kuinka jälkimmäiset kyselyt erosivat alkuperäisistä kyselyistä. Asiaa tarkasteltiin vertailemalla keskenään hakuavainten määriä alkuperäisissä ja myöhemmissä kyselyissä. 32,5 prosentissa muokatuista kyselyistä, hakusanojen määrä alkuperäiseen kyselyyn nähden pysyi samana vaikka muutos tapahtui yhden tai useamman hakuavaimen kohdalla. 41,6 prosentissa tapauksista hakuavainten määrä lisääntyi ja 25,9 prosentissa vähentyi. Voidaankin siis todeta hakijoiden mieluummin lisäävän uusia hakuavaimia kuin poistavan vanhoja. Tutkimuksessa lisäksi osoitettiin hakijoiden muokkaavan alkuperäisiä kyselyitä hyvin pienimuotoisesti. Noin 29,3 prosentissa muokatuista kyselyistä esiintyi vain yksi hakuavain enemmän alkuperäiseen kyselyyn nähden ja 15,5 prosentissa yksi hakuavain vähemmän. Tutkimuksessa tarkasteltiin myös, kuinka monta haun tulosta hakijat keskimäärin tarkastelivat. 28,6 prosenttia hakijoita tarkasteli ainoastaan yhtä sivua ja vain kahta sivua tarkasteli 19 prosenttia hakijoista. Tämä tarkoittaa, että lähes puolet hakijoista tarkasteli korkeintaan kahta tulossivua. (Spink ym. 2001, 226–229.)

Sama tutkimusryhmä on suorittanut myös aiemmin vastaavanlaisia tutkimuksia koskien tiedonhakukäyttäytymisestä (ks. Jansen, Spink, Bateman & Saracevic 1998; Jansen, Spink & Saracevic

2000). Vaikka nämä aikaisempien tutkimusten tulokset hieman eroavat vuoden 2001 tuloksista, joitakin johtopäätöksiä voidaan silti tehdä tuloksia vertailemalla. Hakuavainten määrät kyselyä kohden ovat hieman lisääntyneet. Kyselyjen määrät hakijaa kohden ovat vuosien varrella pysyneet pieninä (2–3 kyselyä/hakija). Myös tuloslistasta tarkasteltujen dokumenttien määrä on säilynyt pienenä.

6 Tutkimuskysymykset ja tutkimuksessa käytetyt menetelmät

6.1 Tutkimuskysymykset

Tutkielmani tarkoituksena on tarkastella hakijoiden katsomien tuloslistan dokumenttien vaikutusta kyselyjen uudelleen muotoiluun. Lähtökohtana tutkimukselleni on oletus siitä, että uudelleen muotoilluissa kyselyissä esiintyvät muut kuin tehtäväkuvauksen sanat ovat tunnistettavissa hakijoiden tuloslistasta katsomista dokumenteista. Tutkimuksen pääkysymys voidaan siis muotoilla seuraavasti:

- Kuinka suuri osa hakijoiden jatkokyselyissä käyttämistä tehtäväkuvauksen ulkopuolisista hakuavaimista on löydettävissä tuloslistasta katsotuista dokumenteista?

Tehtäväkuvauksen ulkopuolisia hakuavaimia, jotka tutkimuksessa todetaan löytyvän tuloslistasta katsotuista dokumenteista, tarkastellaan lisäksi seuraavien alakysymysten avulla:

- Miltä kohtaa dokumenttia hakuavaimet ovat tunnistettavissa?
- Missä muodossa kyselyjen hakuavaimet esiintyvät dokumenttien sanoihin nähden?
- Missä määrin hakijat arvioivat relevanteiksi ne dokumentit, joissa hakuavaimia esiintyy?
- Miltä kohtaa tuloslistaa hakuavaimen sisältävä dokumentti on löydettävissä?

6.2 Tutkimusaineisto

Tutkimuksessa käytetty empiirinen aineisto koostuu osasta INEX 2004 -hankkeen vuorovaikutteisen tiedonhaun tutkimuslinjan aineistoa. Vuorovaikutteinen tutkimuslinja kuuluu osaksi INEX hanketta ja sen tavoitteena on tutkia hakijoiden käyttäytymisistä, kun haun kohteena on XML-dokumentteja (Fuhr ym. 2004, 7). INEX 2004 vuorovaikutteisessa tutkimuslinjassa vuorovaikutteisuutta on lisätty luomalla hakijoille henkilökohtaisen tulkinnan mahdollisuus tiedontarpeesta, käyttämällä testitilanteessa simuloidun työtehtävän mallia sekä tarjoamalla hakijoille mahdollisuus henkilökohtaisen relevanssiarvion suorittamiseen.

6.2.1 INEX 2004 Interaktiivisen tutkimuslinjan aineisto

INEX 2004 vuorovaikutteiseen tutkimuslinjaan osallistui kymmenen ryhmää, joissa kussakin oli vähintään kahdeksan henkilöä. Ryhmät koostuivat yliopisto- ja korkeakouluopiskelijoista Aasiasta, Australiasta ja Euroopasta. INEX hankkeessa käytettiin monia erilaisia aineistonkeruumenetelmiä. Hakijoiden taustatietoja ja tietoja aiemmasta tiedonhakukäyttäytymisestä selvitettiin kyselylomakkeiden avulla, jotka täytettiin ennen varsinaista koetilannetta. Myös koetilanteen jälkeen hakijoilta kerättiin kyselylomakkeiden avulla tietoa muun muassa haun suorittamisen vaikeudesta ja hakuaiheiden ennakkotuntemuksesta. Hakijoiden käyttäytymistä havainnoitiin pitkin hakuprosessia ja koetilanteen jälkeen kaikkia osallistujia haastateltiin. (INEX 2004.) Tiedot suoritetuista hauista tallentuivat lokitietoihin. Lokitiedoista oli nähtävissä tunnistetiedot hakijoista sekä suoritetun haun tapahtuma-aika. Tiedonhakujen osalta lokitiedoista kävi ilmi kyselyt hakusanoineen, tuloslistan elementit sekä elementit, joita hakija oli katsonut. Lisäksi lokitiedoissa näkyi käyttäjien antamat relevanssiarviot katsomistaan elementeistä sekä polut näihin elementteihin. Seuraavissa kappaleissa esitellään tarkemmin INEX 2004 Interaktiivisen tutkimuslinjan testiympäristö, joka koostuu XML-dokumenttikokoelmasta, hakutehtävistä ja relevanssiarvioista.

6.2.1.1 Dokumenttikokoelma

INEX-dokumenttikokoelma koostuu 12107 kokotekstiartikkelista, jotka on kerätty IEEE Computer Sciencen 12:sta eri tieteellisen lehden julkaisuista vuosilta 1995–2002. Dokumenttikokoelma on kooltaan 494 MB ja kaikki dokumentit ovat tallennettu kokoelmaan XML muodossa (INEX 2004). XML muodossa oleva dokumentti koostuu yhdistelmästä tekstiä ja dokumentin elementtejä määritteleviä tageja. XML-dokumentti muodostuu sisäkkäisistä elementeistä, joiden alkaminen ja loppuminen ilmoitetaan alku- ja lopputunnisteiden avulla (Desmarais 2000, 11–12). Elementit edustavat eri osioita dokumenteista kuten otsikkoa, abstraktia, tekijätietoja, lähdetietoja ja itse tekstin eri lukuja. INEX-dokumenttikokoelman dokumentit elementteineen näyttävät seuraavanlaisilta:

Kuva 1. Osa INEX-dokumenttikokoelman XML-dokumentista

6.2.1.2 Hakutehtävät

Interaktiivisen tutkimuslinjan tarkoituksena on tutkia hakijoiden käyttäytymistä vuorovaikutteisessa tiedonhakutilanteessa. Tiedonhakutilanne pyrittiin luomaan mahdollisimman realistiseksi ja tämän vuoksi hakutehtävät perustuivat simuloidun työtehtävän malliin. Simuloidun työtehtävän avulla hakijoiden uskottiin pystyvän kuvittelemaan itsensä paremmin tiedonhakutilanteeseen, jolloin he olisivat motivoituneempia etsimään työtehtävän edellyttämää tietoa. Simuloidusta työtehtävästä käytetään tässä tutkimuksessa nimitystä tehtäväkuvaus, koska hakutehtävä kuuluu osaksi laajempaa kertomusta hakutilanteesta. (Tombros, Larsen & Malik 2004, 25.)

Tehtäväkuvauksia oli kaikkiaan neljä ja ne edustivat kahta eri tehtävätyyppiä, taustatietoa (Background category = B) ja vertailevaa tietoa (Comparison category = C). Taustatietoa edustavan tehtävätyypin tarkoituksena oli löytää mahdollisimman paljon yleistä tietoa koskien tehtäväkuvauksen käsittelemää aihetta. Vertailevassa tehtävätyypissä haluttiin puolestaan löytää eroavaisuuksia kahden tai useamman eri asian välillä. Erilaisten tehtävätyyppien käytöllä on osoitettu olevan vaikutusta siihen, millaisin kriteerein hakija arvioi hakutulostensa relevanssia. Lisäksi erilaisten tehtävätyyppien käytön on ajateltu vaikuttavan hakijoiden tiedonhakukäyttäytymiseen kuten selailuun ja navigointityyliin. (Tombros ym. 2004, 25.)

Kukin hakija sai valita oman mielenkiintonsa perusteella kaksi hakutehtävää, yhden kummastakin tehtävätyypistä. Vapaavalinnaisuudella pyrittiin siihen, että hakijalla oli mahdollisuus valita häntä enemmän kiinnostava aihe. Tehtäväkuvausten B ja C esittämisjärjestystä vaihdeltiin niin, ettei kumpikaan tehtäväkuvaus ollut koko ajan joko ensimmäinen tai jälkimmäinen hakijoille esitetyistä tehtäväkuvauksista. Tällä pyrittiin estämään tutkimustulosten mahdollinen vinouma, jonka aina samassa järjestyksessä esitetyt tehtäväkuvaukset olisivat voineet aiheuttaa. Aikaa yhden tehtävän tekemiseen sai käyttää maksimissaan 30 minuuttia. (Tombros ym. 2004, 25–26)

INEX 2004 Interaktiivisessa tutkimuslinjassa käytettiin web-pohjaista Baseline-käyttöliittymää, jossa hakutulokset avautuivat hakijalle yksinkertaisena listana. Käyttöliittymää varten kehitettiin myös oma hakukone HyREX. Hakijoiden oli mahdollista muodostaa yksinkertaisia kyselyjä, jotka koostuivat joko yksittäisistä sanoista tai lainausmerkeillä erotetuista fraaseista. Hakijat eivät voineet käyttää erillisiä hakuoperaattoreita. Hakujen tuloslista aukeni käyttäjille uuteen ikkunaan ja se muodostui sadasta parhaasta tuloksesta, kymmenestä tuloksesta per sivu. Jokaisen tuloksen osalta oli nähtävissä elementin sijoitus tuloslistassa sekä sen painoarvo (retrieval status value), otsikko, tekijätiedot sekä linkki elementin tarkastelua varten (ks. kuva 2).

Linkin takaa aukeni uuteen ikkunaan itse dokumentti kyseisen elementin kohdalta (ks. kuva 3). Tässä ikkunassa oli myös näkyvissä dokumentin sisällysluettelo. Kyseinen elementti oli korostettu sisällysluettelossa keltaisella ja tekstistä löytyvät hakusanat punaisella värillä. (Malik, Tombros & Larsen 2004, 264–266.) Hakijoiden oli mahdollista navigoida dokumentin eri elementtien välillä sisällysluettelon avulla tai käyttämällä selainikkunan ylälaidassa olevia seuraava- ja edellinen-painikkeita (Tombros ym. 2004, 25).

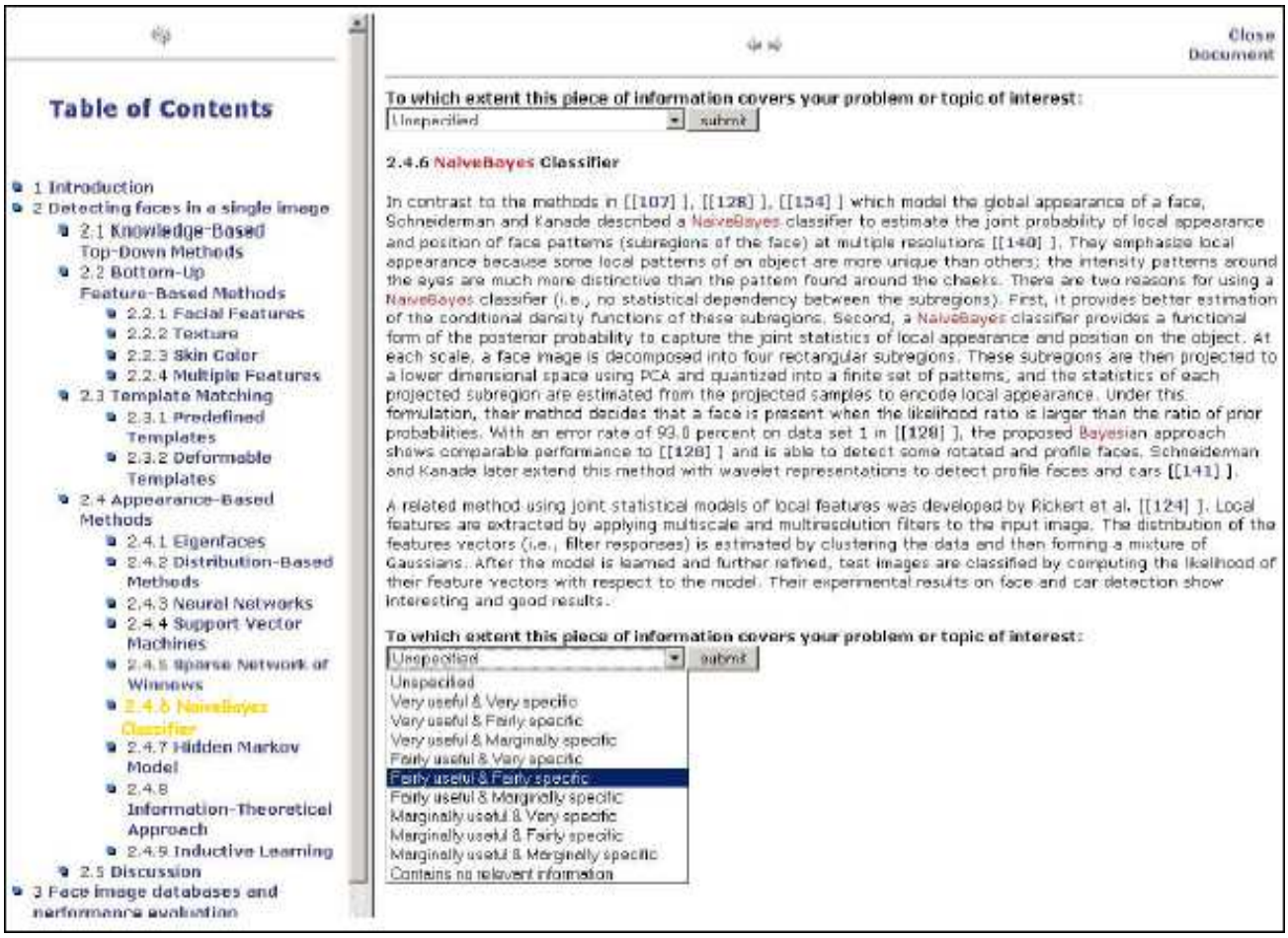
Kuva 2. Tuloslista Baseline-käyttöliittymässä.

6.2.1.3 Relevanssiarviot

Hakijoita pyydettiin tekemään tutkimilleen dokumenttien elementeille relevanssiarvio valitsemalla alasvetovalikosta mielestään sopiva arvo kymmenportaisesta asteikosta. Hakijat arvioivat elementtien relevanssia kahdesta eri näkökulmasta, käyttökelpoisuuden ja spesifisyyden mukaan. Käyttökelpoisuudella mitattiin sitä, missä määrin elementti sisälsi informaatiota, joka oli hyödyksi annetun tehtävän suorittamisessa. Spesifisyydellä puolestaan kartoitettiin sitä, missä määrin elementti käsitteli haluttua aihetta. Käyttökelpoisuudelle ja spesifisyydelle annettiin kummallekin kolme eri arvoa: erittäin, melko ja marginaalinen ja näiden kahden yhdistelmistä saatiin aikaiseksi seuraavanlainen relevanssiasteikko:

A = Erittäin käyttökelpoinen & erittäin spesifi
B = Erittäin käyttökelpoinen & melko spesifi
C = Erittäin käyttökelpoinen & jonkin verran spesifi
D = Melko käyttökelpoinen & erittäin spesifi
E = Melko käyttökelpoinen & melko spesifi
F = Melko käyttökelpoinen & jonkin verran spesifi
G = Jonkin verran käyttökelpoinen & erittäin spesifi
H = Jonkin verran käyttökelpoinen & melko spesifi
I = Jonkin verran käyttökelpoinen & jonkin verran spesifi
J = Ei sisällä relevanttia informaatiota
U = Määrittelemätön

Relevanssiasteikossa kirjain J tarkoitti, ettei dokumentin elementti sisältänyt tehtäväkuvauksen kannalta relevanttia informaatiota ollenkaan. U-kirjain puolestaan merkitsi sitä, ettei hakija ollut tehnyt katsomalleen elementille lainkaan relevanssiarviota. (Malik, Tombros & Larsen 2004, 264–266.)



Kuva 3. Dokumentin sisällysluettelo, elementin selailuikkuna ja relevanssiarvio Baseline-käyttöliittymässä.

6.3 Tutkimusaineiston rajaus

Tämä työ oli jatkoa Nurmelan (2006) pro gradu -tutkielmalle, jossa selvitettiin tehtäväkuvauksen sanojen suhdetta kyselyn sanoihin. Nurmela selvitti työssään, kuinka suuri osa kyselyjen hakuavaimista oli löydettävissä tehtäväkuvauksesta. Tässä työssä oli tarkoituksena hyödyntää Nurmelan saamia tutkimustuloksia niiden hakuavainten osalta, jotka Nurmelan tutkimuksessa todettiin olevan tehtäväkuvauksen ulkopuolelta. Tutkimuksessani käytetty aineisto oli yhdenmukainen Nurmelan tutkimusaineiston kanssa.

Tutkimukseni aineisto koostui osasta INEX 2004 aineiston vuorovaikutteisen tiedonhaun tutkimuslinjan aineistoa. Tutkimusaineisto koostui niistä hakijoista, jotka olivat taustatietoa koskevassa tehtäväkuvauksessa valinneet tehtävän B1 tai vertailevassa tehtäväkuvauksessa tehtävän C2. Hakijoista 54 oli valinnut tehtäväkuvauksen B1 ja 67 tehtäväkuvauksen C2. Tehtäväkuvaukset näyttivät seuraavanlaisilta:

> *Tehtäväkuvaus B1:*
>
> *You are writing a large article discussing virtual reality (VR) applications and you need to discuss their negative side effects.*
>
> *What you want to know is the symptoms associated with cybersickness, the amount of users who get them, and the VR situations where they occur. You are not interested in the use of VR in therapeutic treatments unless they discuss VR side effects.*
>
> *Tehtäväkuvaus C2:*
>
> *You are working on a project to develop a next generation version of a software system. You are trying to decide on the benefits and problems of implementation in a number of programming languages, but particularly Java and Python.*
>
> *You would like a good comparison of these for application development. You would like to see comparisons of Python and Java for developing large applications. You want to see articles, or parts of articles, that discuss the positive and negative aspects of the languages.*
> *Things that discuss either language with respect to application development may be also partially useful to you.*
>
> *Ideally, you would be looking for items that are discussing both efficiency of development and efficiency of execution time for applications.*

Tässä tutkimuksessa oltiin kiinnostuneita ainoastaan niistä hakijoista, jotka käyttivät kyselyissään tehtäväkuvauksen ulkopuolisia hakuavaimia. Tämän vuoksi tutkimusaineistoa rajattiin vielä lisää Nurmelan työhön nähden. Tehtäväkuvauksen B1 osalta kyselyissä esiintyi 56 erilaista hakuavainta, jotka eivät esiintyneet tehtäväkuvauksessa. Näitä 56 hakuavainta käytettiin yhteensä 93 kertaa eri käyttäjien keskuudessa. Vastaavasti tehtäväkuvauksen C2 kyselyissä käytettiin 76 erilaista tehtäväkuvauksen ulkopuolista hakuavainta, joita käytettiin yhteensä 104 kertaa. Tarkasteltavaksi saatu aineisto oli riittävän kattava tämän tutkimuksen tarpeisiin ja molemmat tehtäväkuvaukset olivat hyvin edustettuina. Lista tehtäväkuvausten B1 ja C2 ulkopuolisista hakuavaimista löytyy liitteestä 1.

6.4 Tutkimusaineiston käsittely ja analyysi

Tässä tutkimuksessa käytettiin lokianalyysia tutkimusaineiston analysoinnissa. Seuraavassa kappaleessa käsitellään ensin lokianalyysia analyysimenetelmänä yleisellä tasolla. Tämän jälkeen esitellään kuinka tutkimusaineiston analyysi tapahtui tämän työn osalta.

6.4.1 Lokianalyysi

Lokitiedot ovat tekstitiedostoja, joita kaikki Web-palvelimet pystyvät tuottamaan. Aina kun käyttäjä käyttää Web-palvelinta, hän luovuttaa huomattavan määrän tietoa kyseiselle palvelimelle. Yksittäisestä lokimerkinnästä voidaan selvittää muun muassa seuraavanlaisia tietoja: käyttäjän tietokoneen verkkotunnus (domain name) tai IP-osoite, käyttäjätunnus, mikäli kyseessä on Web-palvelin joka edellyttää sisään kirjautumista, päivämäärä ja kellonaika, jolloin käyttäjä on suorittanut tietyn pyynnön, hakijan pyytämän tiedoston nimi sekä polku tähän tiedostoon, siirretyn tiedon määrä, tieto siitä, miltä sivulta kyseiselle sivulle on siirrytty, käytössä ollut internet-selaimen tyyppi ja versio sekä käytössä ollut käyttöjärjestelmä. (Rubin 2001, 199–201)

Lokianalyysilla tarkoitetaan Web-palvelimen tallentaman lokidatan käsittelemistä ja analysointia. Lokianalyysia voidaan käyttää useisiin eri tarkoituksiin. Rubinin (2001, 197–201) mukaan lokianalyysistä on organisationaalisesti kaksi hyötyä. Lokianalyysiä voidaan hyödyntää markkinoinnin kehittämisessä sekä hakukoneiden analysoimisessa. Yleisimmin käytettyjen fraasien tunnistamisesta voi olla hyötyä markkinoinnin kehittämisessä. Web-sivujen ylläpitäjät voivat puolestaan lokianalyysin avulla muun muassa tarkastella sivustonsa rakennetta ja etsiä mahdollisia virheitä sekä kartoittaa mitä selaimia käyttäjät yleisimmin käyttävät. (Rubin 2001, 197–201.) Lokianalyysia voidaan myös hyödyntää käyttäjätietojen keräämiseen eri palvelimilta. Tiedonhaun tutkimuksessa puolestaan lokianalyysin avulla voidaan tarkastella käyttäjien tiedonhakukäyttäytymistä.

Raakaa lokidataa voidaan käsitellä hyvin taulukkolaskentaohjelmissa kuten Microsoftin Excel- ohjelman avulla. Lokianalyysia varten on kuitenkin kehitetty omia kaupallisia ohjelmia, jotka ovat erityisesti suunniteltu analysoimaan lokitietoja. Tällaisten ohjelmien avulla on muun muassa mahdollista luoda raportteja, jotka esittävät palvelimelta pyydettyjen tiedostojen määrät, viimeksi pyydetyt dokumentit sekä kokonaismäärät, kuinka monta kertaa kullakin yksittäisellä Web-palvelun sivulla on vierailtu. (Dowling 2001.)

6.4.2 Tutkimusaineiston käsittely

Tehtäväkuvausten ulkopuolisten hakuavainten esiintymistä dokumenttien elementeissä selvitettiin tutkimalla hakijoiden suorittamien hakujen lokitietoja. Lokitietoihin oli tallentunut tunnistetiedot hakijoista sekä suoritetun haun tapahtuma-aika, kyselyt hakusanoineen, tuloslistan elementit sekä elementit, joita hakija oli katsonut. Lisäksi lokitiedoissa näkyi käyttäjien suorittamat relevanssiarviot katsomistaan elementeistä sekä polut näihin elementteihin.

Tutkimuksessa lähdettiin liikkeelle tunnistamalla hakuavaimet, jotka eivät esiintyneet tehtäväkuvauksessa, sekä hakijat, jotka tällaisia avaimia käyttivät. Lista tehtäväkuvauksen ulkopuolisista hakuavaimista saatiin Nurmelan (2006) pro gradu -tutkielman aineistoa hyödyntämällä. Nurmelan tutkimuksesta saatua listaa tehtäväkuvauksen ulkopuolisista hakuavaimista muokattiin siten, että hakuavaimista poistettiin liian yleiskäyttöiset sanat. Tällaiset liian yleiskäyttöisiksi tulkitut hakuavaimet olivat tehtäväkuvauksen B1 osalta *it* -sana ja tehtäväkuvauksen C2 osalta avaimet *as-*, *by-* ja *for* -sanat. Liian yleiskäyttöisten hakuavainten vertailu dokumentin sanojen kanssa olisi voinut vääristää tutkimustulosta.

Tehtäväkuvauksen B1 osalta kyselyissä esiintyi 56 erilaista hakuavainta, jotka eivät esiintyneet tehtäväkuvauksessa. Näitä 56 hakuavainta käytettiin yhteensä 93 kertaa eri käyttäjien keskuudessa. Vastaavasti tehtäväkuvauksen C2 kyselyissä käytettiin 76 erilaista tehtäväkuvauksen ulkopuolista hakuavainta, joita käytettiin yhteensä 104 kertaa. Tutkimuksessa selvitettiin ensin, ketkä hakijoista käyttivät kyseisiä hakuavaimia. Tämän jälkeen kyseisten hakijoiden suorittamien kyselyjen lokitiedot käytiin yksi kerrallaan läpi. Lokitiedoista tunnistettiin ensin kyselyt, joissa tehtäväkuvauksen ulkopuoliset hakuavaimet esiintyivät. Tämän jälkeen tarkasteluun otettiin ne dokumenttien elementit, joita hakija oli tarkastellut ennen kyseisen hakuavaimen käyttöä. Elementtien tarkastelu aloitettiin aina sitä kyselyä edeltävästä elementistä, jossa tehtäväkuvauksen ulkopuolinen hakuavain oli esiintynyt. Mikäli hakuavain ei esiintynyt heti kyselyä edeltävässä elementissä, käytiin hakijan tarkastelemia hakuavainta edeltäviä elementtejä niin kauan läpi kunnes hakuavain joko löytyi jostakin elementistä tai hakijan tarkastelemat elementit loppuivat. Mikäli todettiin, että hakuavain löytyi dokumentin elementistä, tutkittiin lokitiedoista lisäksi seuraavia asioita: oliko nähty elementti todettu relevantiksi ja jos oli, niin millaisen relevanssiarvion se oli saanut, monennestako hakijan katsomasta elementistä käsite löytyi ja oliko kyseessä dokumentin tarkka elementti vai kokonainen dokumentti. Lisäksi elementtien osalta tutkittiin, miltä kohtaa elementtiä hakuavain löytyi ja missä

muodossa hakuavain oli löydettävissä. Kaikki tiedot kyselyjen osalta kerättiin taulukkomuotoon Excel-taulukkolaskentaohjelmaan.

Joidenkin hakujen yhteydessä hakijat olivat tutkineet kokonaisia dokumentteja, joiden pituus vaihteli muutaman luvun mittaisista dokumenteista hyvinkin laajoihin dokumentteihin. Joissakin tapauksissa hakijat olivat puolestaan tarkastelleet dokumenttien yksittäisiä elementtejä. Tiedot siitä, oliko kyseessä dokumentin yksittäinen elementti vai kokonainen dokumentti, kirjattiin ylös Excel-taulukkoon. Sillä, kuinka tarkasta kohtaa dokumenttia tehtäväkuvauksen ulkopuolinen hakuavain löytyi, katsottiin tutkimuksessa olevan merkitystä. Mitä lähempää kyselyä ja mitä suppeammalta alueelta hakuavain löytyi, sitä todennäköisempänä pidettiin sitä, että hakija oli poiminut avaimen kyselyynsä juuri katsomastaan elementistä.

Hakuavaimen löytymiskohta jaettiin neljään luokkaan sen perusteella, kuinka keskeisestä kohdasta dokumenttia hakuavain oli tunnistettavissa. Nämä neljä luokkaa olivat *dokumentin otsikko*, *dokumentin alku*, *dokumentin viitetiedot* ja *dokumentin muu varsinainen leipäteksti*. Otsikko-luokka oli kyseessä jos hakuavain löytyi dokumentin pääotsikosta tai dokumentin kirjoittaja kentästä riippumatta siitä, oliko kyseessä kokonainen dokumentti vai yksittäinen elementti. Tämä siksi, että käyttäjän tehdessä kyselyä, tuloslistassa oli aina näkyvissä nämä tiedot (ks. kuva 2). Otsikko-luokka sisälsi lisäksi dokumenttien väliotsikot siinä tapauksessa, että kyseessä oli dokumentin yksittäinen elementti. Alku-luokka oli kyseessä silloin, kun hakuavain esiintyi kokonaisen dokumentin abstraktissa tai jos abstraktia ei ollut dokumentin ensimmäisessä kappaleessa. Dokumentin ensimmäisen kappaleen ollessa yli kymmenen riviä pitkä, alku-luokkaan laskettiin tapaukset, joissa hakuavain esiintyi tällaisen pitkän aloituskappaleen viidellä ensimmäisellä rivillä. Jos kyseessä oli dokumentin yksittäinen elementti, alku-luokaksi laskettiin, mikäli hakuavain esiintyi elementin ensimmäisessä luvussa tai jos kyseessä oli yli kymmenen rivin mittainen aloitusluku, huomioitiin kappaleen viisi ensimmäistä riviä. Jos hakuavain oli löydettävissä dokumentin viitetiedoista, kyseessä oli viite-luokka. Tämä luokka oli kyseessä vain niissä tapauksissa, joissa kyseessä oli kokonainen dokumentti, koska viitetiedot eivät olleet näkyvissä yksittäisten elementtien osalta. Jos hakuavain oli löydettävissä jostakin muualta kohtaa dokumenttia tai dokumentin elementtiä kuin otsikko, alku tai viite-luokista, kyseessä oli teksti-luokka. Tällöin hakuavain oli löydettävissä dokumenttien varsinaisesta leipätekstistä. Mitä keskeisemmässä kohdassa hakuavain esiintyi sitä todennäköisempänä sen päätymistä hakijan kyselyyn, juuri hakijan katsomasta dokumentista pidettiin. Hakuavaimen esiintymisellä otsikko-luokassa, katsottiin olevan kaikkein suurin merkitys ja vastaavasti hakuavaimen esiintymisellä teksti-luokassa oli vähiten arvoa.

Hakuavaimet jaettiin lisäksi viiteen eri luokkaan niiden ilmenemismuodon perusteella. Nämä viisi luokkaa olivat; hakuavaimesta löytyi *yhtenevä ilmaus* dokumentin sanaan nähden, hakuavaimesta löytyi *yksikkö- tai monikkovariantti*, hakuavaimesta löytyi *aikamuotovariantti*, hakuavaimessa oli *kirjoitusvirhe* dokumentissa esiintyvään sanaan nähden ja hakuavain oli *johdettu* tai oli *lyhyempi ilmaisu* dokumentin sanasta. Jos hakuavain esiintyi täsmälleen samassa muodossa kyselyssä ja dokumentissa, se kuului luokkaan yhtenevä ilmaus (analysis–analysis). Hakuavaimen varianteiksi puolestaan laskettiin tapaukset, joissa dokumenttien substantiivit olivat eri yksikkö- tai monikkomuodoissa tai verbit, jotka olivat eri aikamuodossa kuin hakuavaimet sekä tapaukset, joissa hakuavaimessa oli selvästi tapahtunut kirjoitusvirhe ja dokumentista oli tunnistettavissa kyseisen hakuavaimen virheetön ilmaus (disorder–disorders, blur–blured, quasy–queasy). Johdoksiin kuuluivat hakuavaimet, jotka olivat johdoksia dokumentissa esiintyneestä kantasanasta (affecting–affect). Johdoksiin laskettiin kuuluviksi myös tapaukset, joissa hakuavain oli lyhyempi ilmaus dokumentissa esiintyneestä sanasta. Lyhyempi ilmaus saattoi olla joko kantasana dokumentissa esiintyneestä sanasta tai dokumentissa esiintyneen yhdyssanan toinen osa (sick–sickness, motion–motion-related).

7 Tuloksia simuloidun tehtäväkuvauksen vaikutuksesta hakuavainten valintaan ja kyselyn muokkaamiseen INEX 2004 -hankkeessa

Tämä luku perustuu pitkälti Nurmelan (2006) pro gradu -tutkielmaan, jossa selvitettiin tehtäväkuvauksen vaikutusta hakuavainten valintaan ja kyselyjen muodostamiseen. Nurmela vertaili tutkimuksessaan hakijoiden käyttämiä hakuavaimia tehtäväkuvauksen sisältöön tiedon tallennuksen ja haun tasoperiaatteiden mukaisesti kolmella eri tasolla: käsite, ilmaisu- ja esiintymätasoilla. Tiedon haun ja tallennuksen tasoperiaate esiteltiin tarkemmin luvussa 3. Nurmelan tarkoituksena oli selvittää, kuinka suuri osa kyselyjen hakuavaimista on peräisin tehtäväkuvauksesta kunkin tason periaatteiden mukaisesti. Esiintymätaso oli kyseessä mikäli hakuavainten ja tehtäväkuvauksen merkkijonot vastasivat täysin toisiaan (symptoms–symptoms). Ilmaisutasolla hakuavaimia ja tehtäväkuvauksen sanoja vertailtiin väljemmin. Yhdenmukaisiksi katsottiin hakuavaimet, jotka olivat niin sanotusti lähellä tehtäväkuvauksen sanoja. Tällaiset sanat olivat joko yksikkö-monikko tai aikamuoto variaatioita tehtäväkuvauksen sanoista tai ne olivat sanoja, joissa esiintyi kirjoitusvirhe tehtäväkuvauksen sanaan nähden (treatment–treatments, develop–developed, lanuage–language). Käsitetasolla huomioitiin edellisten tasojen lisäksi johdokset ja synonyymit, jotka edustivat tehtäväkuvauksen käsitteitä (therapy–therapeutic, advantage–benefit).

Nurmela tutki työssään myös hakijoiden käyttämiä, tehtäväkuvauksen ulkopuolisia hakuavaimia ja erityisesti niiden edustamien käsitteiden suhdetta tehtäväkuvauksen käsitteisiin. Nurmelan mukaan hakuavain määriteltiin tehtäväkuvauksen ulkopuoliseksi, mikäli se ei ollut ilmaisutason kriteerien mukaan yhdenmukainen tehtäväkuvauksen sanojen kanssa. Tehtäväkuvauksen ulkopuolisten hakuavainten edustamat käsitteet jaettiin Nurmelan tutkimuksessa viiteen eri luokkaan. Nämä luokat olivat synonymia, ylä- tai alakäsite, assosiaatiosuhde ja hakuavaimet, jotka eivät olleet käsitteellisessä suhteessa tehtäväkuvauksen käsitteisiin nähden. (Nurmela 2006, 32–35.)

Luvut 7.1 ja 7.2 esittelevät Nurmelan tutkimuksen tuloksia. Näiden tutkimustulosten esittely on oleellista oman tutkimukseni kannalta, koska tämä työn tulokset ovat jatkoa Nurmelan saamille tuloksille. Tässä työssä ollaan kiinnostuneita juuri siitä hakuavainten joukosta, jotka Nurmelan tuloksissa todettiin olevan simuloidun tehtäväkuvauksen ulkopuolelta.

7.1 Yleisiä tuloksia kyselyjen muodostamisesta

Kaiken kaikkiaan 54 hakijaa suoritti kyselyitä tehtäväkuvauksen B1 osalta. Kyselyitä muodostettiin 292 kappaletta ja hakuavaimia näissä kyselyissä käytettiin yhteensä 933 kappaletta. Näin ollen hakijat muodostivat keskimäärin 5,4 kyselyä ja käyttivät näissä kyselyissä keskimäärin 3,2 hakuavainta. Tehtäväkuvauksen C2 osalta 67 hakijaa suoritti yhteensä 460 kyselyä. Hakuavainten kokonaismäärä oli 1538 kappaletta. Hakijat muodostivat siis keskimäärin 6,9 kyselyä ja käyttivät keskimäärin 3,3 hakuavainta. Hakijoiden määrä ja näin ollen myös suoritettujen kyselyjen sekä käytettyjen hakuavainten kokonaismäärä oli suurempi kuin tehtävän B1 osalta. Kuitenkin keskimääräiset hakuavainmäärät kyselyä kohden pysyivät lähestulkoon samana (ks. taulukko 2).

Taulukko 2. Yleisiä tuloksia kyselyistä (Muokattu teoksesta Nurmela 2006, 39)

Tehtävä-kuvaus	Hakijat	Kyselyt	Hakuavainten kokonaismäärä	Hakuavaimet fraaseineen	Uniikit ha-kuavaimet	Kyselyt/ hakija	Hakuavaimet/ kysely
B1	54	292	933	851	112	5,4	3,2
C2	67	460	1538	1438	181	6,9	3,3
Yhteensä	121	752	2471	2289	293	6,2	3,3

Tehtäväkuvauksen B1 hakuavaimista uniikkeja hakuavaimia oli 112 kappaletta ja tehtäväkuvauksen C2 hakuavaimista 181 kappaletta. Uniikeilla hakuavaimilla tarkoitetaan tässä avaimia, joiden esiintyminen hakulausekkeissa on laskettu vain kertaalleen. Hakuavainten kokonaismäärä saatiin laskemalla kaikki hakulausekkeissa esiintyvät hakuavaimet. Fraaseissa esiintyneet hakuavaimet laskettiin kukin erikseen. Mikäli fraasit laskettiin yhdeksi hakuavaimeksi, saatiin hakuavainten kokonaismääräksi tehtäväkuvauksen B1 osalta 851 hakuavainta ja tehtäväkuvauksen C2 osalta 1438 hakuavainta. (Nurmela 2006, 39–40.)

7.2 Hakuavainten yhdenmukaisuus tehtäväkuvauksen sanojen kanssa

Tarkastelemalla hakuavainten ja tehtäväkuvauksen sanojen yhdenmukaisuutta saatiin selville, kuinka paljon käyttäjät hyödynsivät tehtäväkuvauksen tarjoamia sanoja kyselyitä muodostaessaan. Vastaavasti selville saatiin myös tehtäväkuvauksen ulkopuolisten hakuavainten osuus. Taulukossa 3. esitellään kyselyjen ja tehtäväkuvauksen keskinäistä yhdenmukaisuutta esiintymä-, ilmaisu- ja käsitetasolla kaikkien kyselyjen osalta.

Taulukko 3. Kyselyjen ja tehtäväkuvauksen keskinäinen yhdenmukaisuus esiintymä-, ilmaisu- ja käsitetasolla (Muokattu teoksesta Nurmela 2006, 40)				
Tehtäväkuvaus	**Kyselyt**	**Esiintymätaso**	**Ilmaisutaso**	**Käsitetaso**
B1	292	0,75	0,8	0,81
C2	460	0,82	0,86	0,9
Yhteensä	752	0,79	0,84	0,87

Molempien tehtäväkuvauksien kaikkien kyselyjen osalta, keskimäärin 80–90% hakuavaimista oli löydettävissä tehtäväkuvauksista. Kaikkein tiukimpienkin eli esiintymätason kriteerien mukaan laskettuna keskimäärin 79% hakuavaimista esiintyi tehtäväkuvauksissa. Tämä tarkoittaa, että vajaa 80% käytetyistä hakuavaimista esiintyi täsmälleen samassa muodossa tehtäväkuvauksessa. Ilmaisutason kriteerien mukaan laskettuna tehtäväkuvauksesta löytyvien hakuavainten määrä kasvoi viisi prosenttiyksikkö ja käsitetason kriteerien mukaan laskettuna vielä kolme prosenttiyksikköä lisää. (Nurmela 2006, 42–43.)

Vaikka selvä enemmistö hakuavaimista esiintyi tehtäväkuvauksissa, silti keskimäärin 10–20% hakuavaimista oli tehtäväkuvausten ulkopuolisia. Lisäksi, kun yhdenmukaisuus vertailu kyselyjen ja tehtäväkuvausten sanojen välillä suoritettiin erikseen ensimmäisten ja viimeisten kyselyjen osalta, tehtäväkuvauksen ulkopuolisten hakuavainten osuus kasvoi selvästi viimeisten kyselyjen kohdalla (ks. taulukko 4).

Taulukko 4. Ensimmäisten sekä viimeisten kyselyjen ja tehtäväkuvauksen keskinäinen yhdenmukaisuus esiintymä-, ilmaisu- ja käsitetasolla. (Muokattu teoksesta Nurmela 2006, 42–43)						
Tehtäväkuvaus	**Esiintymätaso**		**Ilmaisutaso**		**Käsitetaso**	
	Ensimmäinen kysely	**Viimeinen kysely**	**Ensimmäinen kysely**	**Viimeinen kysely**	**Ensimmäinen kysely**	**Viimeinen kysely**
B1	0,88	0,67	0,93	0,71	0,93	0,76
C2	0,88	0,82	0,92	0,83	0,95	0,87
Yhteensä	0,88	0,75	0,92	0,78	0,94	0,82

Koska tämän tutkimuksen tarkoitus oli tarkastella nimenomaan tehtäväkuvausten ulkopuolisia hakuavaimia siitä näkökulmasta, että hakijat poimivat uusia hakuavaimia näkemistään dokumenteista, eivät ensimmäisten kyselyjen tarjoamat yhdenmukaisuuslukemat olleet kovinkaan informatiivisia. Vasta jatkokyselyjen kohdalla hakijan oli mahdollista hyödyntää dokumenteissa käytettyjä sanoja uusia hakuavaimia valitessaan. Mielenkiintoista oli huomata, kuinka paljon vähemmän hakijat turvautuivat tehtäväkuvauksen sanoihin viimeistä kyselyä muodostaessaan. Erot yhdenmukaisuusluvuissa tehtäväkuvausten sanojen sekä ensimmäisten ja viimeisten kyselyjen välillä vaihtelivat keskimäärin 12–14 prosenttiyksikköä eri tasojen välillä. Viimeistä kyselyä muodostaessaan tehtäväku-

vauksen ulkopuolisia hakuavaimia oli keskimäärin 18% käsitetason, 22% ilmaisutason ja jopa 25% esiintymätason kriteerien mukaan. Eri tehtäväkuvauksien kesken esiintyi myös vaihtelua. Enimmillään tehtäväkuvauksen ulkopuolisia hakuavaimia esiintyi tehtävän B1 kyselyissä, esiintymätason kriteerien mukaisesti tarkasteltuna. Tällöin jopa kolmannes hakuavaimista oli tehtäväkuvauksen ulkopuolelta.

8 Tutkimustulokset

Tässä luvussa esitellään tutkimuksen tulokset. Ensimmäiseksi tarkastellaan, kuinka suuri osa tehtäväkuvauksen ulkopuolisista hakuavaimista oli löydettävissä tuloslistasta katsotuista dokumenteista. Seuraavaksi tarkastellaan, millaisia relevanssiarvioita hakijat olivat antaneet niille dokumenteille, joista kyselyn hakuavaimia oli tunnistettavissa. Tämän jälkeen tarkastellaan kuinka pian hakija oli muodostanut uuden kyselyn katsottuaan tuloslistan dokumenttia, jossa hakuavain esiintyi. Lopuksi tutkitaan vielä miltä kohtaa dokumenttia hakuavaimia oli tunnistettavissa ja missä muodossa ne esiintyivät.

8.1 Tehtäväkuvauksen ulkopuolisten hakuavainten esiintyminen dokumenteissa

Nurmelan (2006) tutkimuksessa hakuavaimen laskettiin olevan tehtäväkuvauksesta, mikäli se täytti ilmaisutason määritelmän ehdot. Muussa tapauksessa hakuavain luokiteltiin tehtäväkuvauksen ulkopuoliseksi. Tässä tutkimuksessa lähdetään liikkeelle vastaavanlaisella luokittelulla. Nurmelan tutkimuksessa saatua tehtäväkuvauksen ulkopuolisten hakuavainten joukkoa muokattiin paremmin tämän tutkimuksen tarpeita vastaavaksi. Tehtäväkuvauksen ulkopuolisista hakuavaimista karsittiin pois sellaiset sanat, jotka tulkittiin liian yleisesti käytetyiksi. Tällaisia sanoja olivat *it, as, by* ja *for*. Myös hakuavaimet, jotka esiintyivät heti lähtökyselyissä, jätettiin tutkimuksen ulkopuolelle. Lista tehtäväkuvauksen ulkopuolisiksi lasketuista hakuavaimista löytyy liitteestä 1.

Tehtäväkuvauksen ulkopuolisten hakuavainten joukosta osa hakuavaimista esiintyi useampaan kertaan, sillä eri hakijat olivat käyttäneet kyselyissään samoja hakuavaimia. Hakuavain tulkittiin löytyvän dokumentista, mikäli se täytti ilmaisutason määritelmän kriteerit. Lisäksi käsitetason kriteereistä, johdokset laskettiin myös kuuluviksi tähän joukkoon. Ilmaisutason kriteerien mukaan yhdenmukaisiksi katsottiin hakuavaimet, jotka vastasivat täysin toisiaan (esiintymätaso) sekä hakuavaimet, jotka olivat niin sanotusti lähellä tehtäväkuvauksen sanoja (ilmaisutaso). Tällaiset sanat olivat joko yksikkö-monikko tai aikamuoto variaatioita tehtäväkuvauksen sanoista tai ne olivat sanoja, joissa esiintyi kirjoitusvirhe tehtäväkuvauksen sanaan nähden. Käsitetason kriteereistä huomioitiin ainoastaan johdokset. Nurmelan (2006) tutkielmassa käsitetason kriteerien mukaan yhdenmukaisiksi hakuavainten kanssa laskettiin myös synonyymit (advantage–benefit). Tässä tutkielmassa synonyymit jätettiin kuitenkin yhdenmukaisuus vertailun ulkopuolelle. Vaikka dokumentissa esiintyisi synonyymi hakuavaimeen nähden, ei ole järkevää olettaa sen päätyneen hakijan kyselyyn juuri tuloslistasta katsotusta dokumentista.

Tutkimustulokset osoittivat, että suurin osa hakuavaimista, jotka eivät olleet löydettävissä tehtäväkuvauksista, löytyivät sen sijaan hakijoiden katsomista tuloslistan dokumenteissa. Tehtävässä B1 osallistujat käyttivät tehtäväkuvauksen ulkopuolisia hakuavaimia yhteensä 93 kertaa. 86% näistä avaimista esiintyi osallistujien näkemissä dokumenteissa tai dokumenttien yksittäisissä elementeissä. Myös tehtäväkuvauksen C2 osalta osallistujien käyttämistä hakuavaimista suurin osa oli löydettävissä nähdyistä dokumenteista. C2 tehtäväkuvauksen 104 hakuavaimesta 78% oli löydettävissä nähdyistä dokumenteista. Kaiken kaikkiaan tehtäväkuvauksen C2 kyselyissä käytettiin hieman enemmän tehtäväkuvauksen ulkopuolisia hakuavaimia ja näiden hakuavainten joukossa oli myös useammin hakuavaimia, joita ei ollut tunnistettavissa hakijoiden katsomista dokumenteista (ks. taulukko 5).

Taulukko 5. Tehtäväkuvauksen ulkopuolisten hakuavainten esiintyminen dokumenteissa.

	Tehtävä-kuvaus B1		**Tehtävä-kuvaus C2**	
	kpl	**%**	**kpl**	**%**
Hakuavain esiintyi dokumentissa	80	86	81	77,9
Hakuavain ei esiintynyt dokumentissa	13	14	23	22,1
Yhteensä	93	100	104	100

8.2 Nähtyjen dokumenttien relevanttius ja relevanssiaste

Hakijat suorittivat katsomilleen dokumenteille relevanssiarvion. Taulukossa 6. on esitetty, kuinka suuri osa dokumenteista, joista oli tunnistettavissa kyselyjen hakuavaimia, arvioitiin relevanteiksi. Molempien tehtäväkuvausten osalta voidaan todeta, että suurin osa tällaisista dokumenteista arvioitiin relevanteiksi. Tehtäväkuvauksessa B1 käytetyt hakuavaimet olivat löydettävissä dokumenteista 80 kertaa. Dokumentit, joista hakuavaimia oli löydettävissä, olivat 90% tapauksia arvioitu relevanteiksi. Tehtäväkuvauksen C2 osalta dokumentit, joista hakuavaimia oli tunnistettavissa, arvioitiin hieman useammin epärelevanteiksi. Kuitenkin myös tehtäväkuvauksen C2 osalta dokumentit, joissa hakuavaimet esiintyivät, olivat 85% tapauksista arvioitu relevanteiksi.

Taulukko 6. Niiden dokumenttien relevanttius, joista oli tunnistettavissa kyselyjen hakuavaimia

	Tehtävä-kuvaus B1		**Tehtävä-kuvaus C2**	
	kpl	%	kpl	%
Relevantti	72	90	69	85,2
Epärelevantti	8	10	12	14,8
Yhteensä	80	100	81	100

Käyttäjät arvioivat elementtien relevanssia kymmenportaisen asteikon perusteella. Taulukossa 7. on esitetty elementtien saamat relevanssiarviot niiden elementtien osalta, joista kyselyjen hakuavaimia oli löydettävissä.

Taulukko 7. Niiden dokumenttien relevanssiaste, joista oli tunnistettavissa kyselyjen hakuavaimia

	Tehtäväkuvaus B1		Tehtäväkuvaus C2	
	kpl	**%**	**kpl**	**%**
A Erittäin käyttökelpoinen & erittäin spesifi	19	23,8	18	22,2
B Erittäin käyttökelpoinen & melko spesifi	9	11,3	4	4,9
C Erittäin käyttökelpoinen & jonkin verran spesifi	3	3,8	0	0
D Melko käyttökelpoinen & erittäin spesifi	10	12,5	6	7,4
E Melko käyttökelpoinen & melko spesifi	14	17,5	6	7,4
F Melko käyttökelpoinen & jonkin verran spesifi	5	6,3	9	11,1
G Jonkin verran käyttökelpoinen & erittäin spesifi	2	2,5	1	1,2
H Jonkin verran käyttökelpoinen & melko spesifi	1	1,3	4	4,9
I Jonkin verran käyttökelpoinen & jonkin verran spesifi	9	11,3	21	25,9
J Epärelevantti	8	10	12	14,8
Yhteensä	80	100	81	100

Molempien tehtäväkuvausten osalta yli 20% elementeistä arvioitiin relevanssiasteikon korkeimman asteen mukaan eli *”erittäin käyttökelpoiseksi ja erittäin spesifiksi”*. Muita relevanssiasteikon arvoja käytettiin hyvin vaihtelevasti. Tehtäväkuvauksen B1 osalta *”erittäin käyttökelpoiseksi ja erittäin spesifiksi”* arvioitujen elementtien lisäksi hakijoiden antamat relevanssiarviot painottuivat asteikon keskivaiheille. Relevanssiasteet *”melko käyttökelpoinen ja erittäin spesifi”* sekä *”melko käyttökelpoinen ja melko spesifi”* saivat yhteensä 30% hakijoiden suorittamista arvioista. Tehtäväkuvauksen C2 osalta hakijoiden arvioimat elementtien relevanssiasteet painottuivat molempiin ääripäihin. *”Erittäin käyttökelpoiseksi ja erittäin spesifiksi”* hakijat arvioivat 22% elementeistä. *”Jonkin verran käyttökelpoiseksi ja jonkin verran spesifiksi”* hakijat arvioivat puolestaan 26% elementeistä. Loput elementtien saamat relevanssiarviot painottuivat pääasiassa asteikon keskivaiheille.

Mielenkiintoista oli huomata, kuinka kahden eri tehtäväkuvauksen osalta, elementtien saamat relevanssiarviot vaihtelivat näinkin suuresti. Erityisesti *”jonkin verran käyttökelpoiseksi ja jonkin verran spesifiksi”* arvioitujen elementtien määrä vaihteli tehtäväkuvauksien välillä huomattavasti. Vertailevaa tietoa edustavan tehtävätyypin osalta hakijat olivat lähes 15 prosenttiyksikköä useammin arvioineet elementin *”jonkin verran käyttökelpoiseksi ja jonkin verran spesifiksi”*, verrattuna taustatietoa edustavan tehtävätyypin kohdalla. Ylipäänsä taustatietoa edustavan tehtävätyypin osalta hakijat arvioivat elementit keskimäärin relevantimmiksi kuin vertailevan tehtävätyypin kohdalla.

Kymmenportaisen relevanssiasteikon käyttö on voinut myös vaikuttaa käyttäjien suorittamiin relevanssiarvioihin. Tulokset voisivat olla erilaisia, jos käytössä ollut relevanssiasteikko olisi sisältänyt vähemmän arvoasteita. Tutkimusaineistossa käytetty kymmenportainen relevanssiasteikko on hyvin monimutkainen ja käyttäjien on voinut olla vaikeaa hahmottaa eroja eri relevanssiasteiden välillä. Tämän tutkimuksen osalta ei voida kuitenkaan vetää yleistettäviä johtopäätöksiä, sillä tutkimuksessa tarkasteltiin ainoastaan kahta erilaista tehtäväkuvausta.

Pharo ja Nordlie (2005) tutkivat moniasteisen relevanssiasteikon käyttöä INEX 2004 -hankkeen vuorovaikutteisen tiedonhaun tutkimuslinjan aineiston yhteydessä. Tutkimuksen tulosten mukaan hakijoiden antamissa relevanssiarvioissa ilmeni suurta ristiriitaisuutta. Käyttäjät olivat antaneet samoille dokumenteille useita eriasteisia relevanssiarvioita, saman kyselyn aikana. 10% tapauksia, käyttäjät olivat arvioineet saman dokumentin sekä *”Erittäin käyttökelpoiseksi ja erittäin spesifiksi”* että *”Epärelevantiksi”*. Lisäksi yli 30% dokumenteista oli arvioitu viiden tai useamman eri relevanssiasteikon arvoasteen mukaan, yhden kyselyn aikana, saman käyttäjän suorittamana. Ristiriitaisten relevanssiarvioiden arveltiin johtuvan muun muassa moniportaisen relevanssiasteikon käytöstä. Kymmenen luokkaa relevanssiasteikossa saattoi olla liian paljon käyttäjille, jotta he olisivat pystyneet tekemään yhdenmukaisia relevanssiarvioita. Myös kolmen eri arvon (erittäin, melko ja marginaalinen) käyttämistä käyttökelpoisuuden ja spesifisyyden määrittelemisessä pidettiin ongelmallisena. (Pharo & Nordlie 2005, 239, 246.)

8.3 Monennestako elementistä hakuavain löytyi

Tutkittaessa hakuavainten esiintymistä dokumenteissa, mielenkiintoista oli myös tarkastella, kuinka monennesta hakijan katsomasta tuloslistan elementistä avain oli tunnistettavissa. Tutkittaessa kyselyn yksittäisiä hakuavaimia lokitiedoissa lähdettiin menemään dokumenttien elementtejä taaksepäin. Lukema kertoo siis sen, oliko hakija muodostanut uuden hakulausekkeen heti sellaisen elementin jälkeen, josta kyseisen hakuavain oli tunnistettavissa vai oliko käyttäjä selaillut useampia elementtejä ennen uuden kyselyn muodostamista. Taulukossa 8. on esitetty kuinka monennesta elementistä hakuavaimet olivat löydettävissä.

Taulukko 8. Kuinka monennesta elementistä käsite löytyi				
	Tehtäväkuvaus B1		Tehtäväkuvaus C2	
	kpl	%	kpl	%
1	43	53,8	27	33,3
2	14	17,5	16	19,8
3	8	10	11	13,6
4	6	7,5	8	9,9
5	5	6,3	8	9,9
6	1	1,3	5	6,2
7	0	0	2	2,5
8	0	0	2	2,5
9	1	0	0	0
10	0	1,3	2	2,5
11	2	2,5	0	0
Yhteensä	80	100	81	100

Molempien tehtäväkuvausten osalta voidaan todeta, että elementit, joista hakuavaimia oli löydettävissä, sijoittuivat hakijoiden tarkasteluissa lähelle uuden kyselyn muodostushetkeä. Tehtäväkuvauksen B1 osalta hakijat olivat yli puolessa tapauksista muodostaneet hakulausekkeen heti sellaisen elementin jälkeen, josta oli ollut löydettävissä kyseisen hakulausekkeen avaimia. Lisäksi yli 80% tapauksista hakuavain oli esiintynyt kolmen viimeksi katsotun elementin joukosta. Tehtäväkuvauksen C2 osalta hakijat muodostivat uusia kyselyitä harvemmin heti sellaisten elementtien jälkeen, joista kyselyn hakuavaimia oli tunnistettavissa. Kuitenkin myös tehtäväkuvauksen C2 hakuavaimista keskimäärin joka kolmas avain oli esiintynyt viimeksi katsotussa elementissä ja oli ollut löydettävissä kolmesta viimeisimmästä elementistä 67% tapauksia.

Tuloslistasta katsottujen elementtien sijoitusten tarkastelu kertoi myös jotakin siitä, kuinka paljon hakijat ylipäänsä silmäilivät tuloslistan dokumentteja. Aiemmat tutkimukset (ks.Esim. Spink ym. 2001; Jansen ym. 1998; 2000) osoittivat hakijoiden tarkastelevan vain vähän tuloslistan dokumentteja. Spinkin ym. (2001, 226–229) tutkimustulosten mukaan hieman alle kolmannes hakijoista tarkasteli ainoastaan yhtä ja noin 20% hakijoista tarkasteli vain kahta tuloslistan dokumenttia. Tämän tutkimuksen tulokset viittaavat toisenlaiseen hakukäyttäytymiseen. Hakijat tarkastelivat tämän tutkimuksen osalta huomattavasti enemmän tuloslistan dokumentteja. Tehtäväkuvauksen B1 osalta kyselyn hakuavain oli löydettävissä lähes joka neljäs kerta ja tehtäväkuvauksen C2 osalta peräti joka kolmas kerta vasta kolmannesta viidenteen, hakijan katsomasta dokumenttia. Tämä tarkoittaa hakijan tarkastelleen usein enemmän kuin kahta tuloslistan dokumenttia. Yleistettäviä johtopäätök-

siä ei voida tämän tutkimuksen osalta kuitenkaan tehdä, sillä elementtejä tarkasteltiin tuloslistassa vain niin pitkälle, kunnes kyselyn hakuavain tunnistettiin elementistä.

8.4 Hakuavainten esiintymiskohdat dokumenteissa

Tutkimuksessa haluttiin tarkastella, kuinka tarkasta kohdasta dokumenttia kyselyjen hakuavaimet löytyivät. Osa tarkastelluista elementeistä oli kokonaisia, hyvinkin laajoja dokumentteja ja osa puolestaan näiden dokumenttien yksittäisiä lukuja tai kappaleita. Lähtökohtana oli ajatus siitä, että mitä tarkemmasta kohdasta dokumenttia hakuavain oli löydettävissä, sitä todennäköisemmin hakija oli poiminut sen kyselyynsä juuri katsomastaan elementistä. Taulukossa 9. on esitetty jaottelu hakuavaimista, jotka löytyivät joko tarkasta kohtaa dokumenttia tai kokonaisesta dokumentista.

Taulukko 9. Tarkastellun elementin laajuus

	Tehtävä-kuvaus B1		Tehtävä-kuvaus C2	
	kpl	**%**	**kpl**	**%**
Dokumentin tarkka kohta	55	68,8	49	60,5
Koko dokumentti	25	31,3	32	39,5
Yhteensä	80	100	81	100

Molempien tehtäväkuvausten osalta hakuavain löytyi useammin dokumentin tarkasta kohdasta kuin kokonaisesta dokumentista. Tehtäväkuvauksen B1 osalta lähes 70% ja tehtäväkuvauksen C2 osalta hieman yli 60% hakuavaimista esiintyi tarkassa kohdassa dokumenttia.

Dokumentissa esiintyneen hakuavaimen paikallistaminen haluttiin tehdä yhä tarkemmin. Seuraavaksi tutkimuksessa selvitettiin, miltä kohtaa dokumenttia tai dokumentin elementtiä hakuavain oli löydettävissä. Esiintymiskohdat luokiteltiin neljään eri luokkaan, jotka olivat otsikko-, alku-, viite- ja tekstiluokka. Luokista kaikista tarkoin oli otsikkoluokka ja vastaavasti tekstiluokka oli laajin. Luokkien tarkemmat määrittelyt on esitelty aiemmin luvun 6.4.2 yhteydessä. Taulukossa 10. esitetään hakuavainten esiintymiskohdat elementeissä.

Taulukko 10. Hakuavainten esiintymiskohdat dokumenteissa.

	Tehtäväkuvaus B1		Tehtäväkuvaus C2	
	Esiintymät	**%**	**Esiintymät**	**%**
Otsikko	25	20,8	13	11,8
Alku	27	22,5	19	17,3
Viite	9	7,5	13	11,8
Teksti	59	49,2	65	59,1
Yhteensä	120	100	110	100

Yksittäinen hakuavain saattoi esiintyä useassa kohdassa elementtiä. Myös elementin samassa kohdassa saattoi esiintyä yksittäisen hakuavaimen eri variantteja. Tutkimuksessa on kuitenkin huomioitu vain hakuavaimen ensimmäinen ilmenemismuoto. Suurin osa hakuavaimista oli löydettävissä joltakin kohtaa dokumenttien varsinaisesta leipätekstistä. Tehtäväkuvauksen B1 osalta lähes puolet hakuavaimista esiintyi tekstiluokassa ja tehtäväkuvauksen C2 osalta lähes 60%. Mielenkiintoista oli kuitenkin havaita, että myös huomattavan suuri osa hakuavaimista oli tunnistettavissa alkuosasta dokumenttia, joko heti dokumentin otsikosta tai elementin alusta. Tehtävän B1 osalta otsikosta tai elementin alusta hakuavain löytyi 40 prosentissa tapauksia ja tehtävän C2 osalta lähes joka kolmas kerta.

8.5 Hakuavainten ilmenemismuoto dokumenteissa

Hakijoiden käyttämien tehtäväkuvauksen ulkopuolisten hakuavainten esiintymistä dokumenteissa tarkasteltiin seuraavaksi niiden ilmenemismuodon perusteella. Ilmenemismismuodolla tarkoitetaan tässä hakuavaimen ja dokumentin sanan suhdetta toisiinsa. Hakuavaimesta saattoi esiintyä yhtenevä ilmaus dokumentin sanaan nähden. Hakuavaimesta saattoi myös esiintyä yksikkö- tai monikkovariantti, aikamuotovariantti tai hakuavaimessa saattoi olla kirjoitusvirhe dokumentin sanaan nähden. Myös hakuavaimet, jotka olivat johdoksia dokumentin sanaan nähden tai dokumentissa esiintyneestä johdoksesta perusmuotoistettuja hakuavaimia, olivat oma ryhmänsä. Näin ollen hakuavaimet jaettiin ilmenemismismuotonsa perusteella kolmeen luokkaan; yhtenevä ilmaus, variantit ja johdos. Luokkien tarkemmat määrittelyt löytyvät luvun 6.4.2 yhteydestä. Yksittäinen hakuavain saattoi esiintyä dokumentissa useassa eri muodossa. Yksittäinen hakuavain saattoi myös ilmaantua tietyssä muodossa useassa kohdassa dokumentin elementtiä. Hakuavaimen tietty ilmenemismuoto otettiin kuitenkin huomioon vain kertaalleen. Taulukosta 11. käy ilmi missä muodoissa kyselyjen hakuavaimet esiintyivät dokumenteissa.

Taulukko 11. Hakuavainten ilmenemismuoto

	Tehtävä-kuvaus B1		**Tehtävä-kuvaus C2**	
	kpl	%	**kpl**	%
Yhtenevä ilmaus	72	67,9	46	43
Variantit	13	12,3	17	15,9
Johdos	21	19,8	44	41,1
Yhteensä	106	100	107	100

Hakuavainten ilmenemismuodoissa tehtäväkuvausten välillä esiintyi merkittävää vaihtelua. Tehtäväkuvauksen B1 hakuavaimista valtaosasta (68%) oli löydettävissä yhtenevä ilmaus dokumentista.

Lähes 20% oli johdoksia dokumenttien sanoista ja loput 12% olivat hakuavainten variantteja. Tehtäväkuvauksen C2 osalta hakuavainten ilmenemismuodot jakautuivat tasaisemmin eri luokkien kesken. Kyseisen tehtäväkuvauksen hakuavaimista lähes saman verran oli joko yhteneviä ilmauksia (43%) tai ne olivat johdoksia (41%). Loput 16% tehtäväkuvauksen C2 hakuavaimista olivat variantteja dokumenttien sanoihin nähden.

Taulukossa 12. on esitetty vielä tarkempi erittely hakuavainten ilmenemismuotojen jakautumisesta eri varianttien kesken. Tehtäväkuvauksen B1 osalta hakuavaimissa esiintyi yksikkö-monikkovariantteja kahdeksan kappaletta, aikamuotovariantteja kolme kappaletta ja kirjoitusvirheen sisältäviä hakuavaimia kaksi kappaletta dokumentin sanoihin nähden. Tehtäväkuvauksen C2 osalta sekä yksikkö-monikko- että aikamuotovariantteja hakuavaimista esiintyi molempia kahdeksan kappaletta. Yhdessä hakuavaimessa esiintyi kirjoitusvirhe dokumentin sanaan nähden.

Taulukko 12. Varianttien jakautuminen

	Tehtävä-kuvaus B1		**Tehtävä-kuvaus C2**	
	kpl	%	**kpl**	%
Yksikkö-monikkovariantti	8	7,5	8	7,5
Aikamuotovariantti	3	2,8	8	7,5
Kirjoitusvirhevariantti	2	1,9	1	0,9
Variantteja yhteensä	13	12,3	17	15,9

8.6 Hakuavainten ilmenemismuoto eri kohdissa dokumenttia

Hakuavainten esiintymistä dokumenteissa tutkittiin edellä erikseen sekä niiden ilmenemismuodon että sijaintikohdan perusteella. Mielenkiintoista tutkimukseni kannalta oli myös tarkastella näitä kahta muuttujaa yhdessä eli sitä, mistä kohdista dokumenttia tietyissä muodoissa esiintyneet hakuavaimet löytyivät. Taulukoissa 13. ja 14. on esitetty tulokset erikseen molempien tehtäväkuvausten osalta. Taulukoiden tulokset on laskettu siten, että yksittäisen hakuavaimen kaikki eri ilmenemismuodot eri kohdissa dokumenttia on huomioitu.

Taulukko 13. Hakuavainten ilmenemismuoto eri kohdissa dokumenttia. Tehtäväkuvaus B1

	Otsikko (%)	**Alku** (%)	**Lähdetiedot** (%)	**Teksti** (%)	**Yhteensä**
Yhtenevä ilmaus	12 (12,6)	23 (24,2)	9 (9,5)	51 (53,7)	95
Variantit	2 (11,1)	2 (11,1)	1 (5,6)	13 (72,2)	18
Johdos	12 (44,4)	5 (18,5)	0	10 (37)	27

Taulukko 14. Hakuavainten ilmenemismuoto eri kohdissa dokumenttia. Tehtäväkuvaus C2					
	Otsikko (%)	**Alku** (%)	**Lähdetiedot** (%)	**Teksti** (%)	**Yhteensä**
Yhtenevä ilmaus	7 (12,1)	10 (17,2)	5 (8,6)	36 (62,1)	58
Variantit	0	4 (20)	2 (10)	14 (70)	20
Johdos	6 (10,3)	7 (12,1)	6 (10,3)	39 (67,2)	58

Sekä tehtäväkuvauksen B1 että C2 yhtenevistä ilmauksista yli puolet esiintyi tekstissä. Tehtäväkuvauksen B1 osalta 54% ja tehtäväkuvauksessa C2 osalta 62% yhtenevistä ilmauksista oli löydettävissä tekstistä. Molempien tehtäväkuvausten hakuavaimista kuitenkin kohtalaisen suuri osa yhtenevistä ilmauksista löytyi joko heti dokumentin alusta tai sen otsikosta. Tehtäväkuvauksen B1 yhtenevistä ilmauksista 13% löytyi otsikosta ja 24% dokumentin alusta. Tehtäväkuvauksen C2 yhtenevistä ilmauksista puolestaan 12% esiintyi otsikossa ja 17% dokumentin alussa. Molempien tehtäväkuvauksien osalta hakuavainten variantit esiintyivät pääasiassa elementtien leipätekstissä. Tehtäväkuvauksen B1 osalta 72% ja tehtäväkuvauksen C2 osalta 70% hakuavainten varianteista oli löydettävissä elementtien leipätekstistä. Eroa tehtäväkuvausten B1 ja C2 kesken löytyi kuitenkin siitä, miltä kohtaa dokumenttien sanoista johdetut tai niistä lyhennetyt hakuavaimet esiintyivät. Tehtäväkuvauksen B1 osalta johdosten esiintyminen joko dokumentin otsikossa tai tekstissä jakaantui hyvin tasaisesti. Johdoksista 44% löytyi otsikosta ja 37% tekstistä. Kun huomioidaan dokumenttien alusta löytyneet johdokset (19%), voimme huomata, että tämän tehtäväkuvauksen osalta suurempi osa johdoksista löytyi useammin joko dokumentin otsikosta tai alusta kuin myöhemmästä vaiheesta tekstiä. Johdosten ilmenemiskohdissa tehtäväkuvauksen C2 osalta esiintyi kuitenkin huomattavasti enemmän eroja. Dokumentin sanoista johtamalla muodostetuista hakuavaimista vain 10% löytyi otsikosta ja 12% dokumentin alusta. Sen sijaan peräti 67% johdoksista oli löydettävissä dokumenttien varsinaisesta tekstiosuudesta.

9 Yhteenveto ja johtopäätökset

Tutkimuksen tarkoituksena oli tarkastella nähtyjen dokumenttien vaikutusta kyselyjen uudelleen muotoiluun. Lähtökohtana tutkimukselle oli oletus siitä, että uudelleen muotoilluissa kyselyissä esiintyvät muut kuin tehtäväkuvauksen hakuavaimet olisivat tunnistettavissa tuloslistasta esiintyvistä dokumenteista.

Nurmelan (2006) pro gradu -tutkimuksessa osoitettiin, että hakijoiden käyttämistä kyselyjen hakuavaimista keskimäärin 10–20% ei ollut tunnistettavissa tehtäväkuvauksesta. Kun kyselyitä tutkittiin erikseen ensimmäisten ja viimeisten kyselyjen osalta, tehtäväkuvauksen ulkopuolisten hakuavainten osuus kasvoi selvästi viimeisten kyselyjen kohdalla. Nurmelan tutkimuksessa osoitettiin, että viimeisten kyselyjen osalta tehtäväkuvauksen ulkopuolisten hakuavainten määrä oli keskimäärin 18–25%. Tässä tutkimuksessa haluttiin selvittää, esiintyivätkö nämä tehtäväkuvauksen ulkopuoliset hakuavaimet, tuloslistassa esiintyvissä dokumenteissa.

Tutkimustulokset osoittivat, että suurin osa jatkokyselyjen hakuavaimista, jotka eivät olleet tunnistettavissa tehtäväkuvauksesta, olivat tunnistettavissa tuloslistan dokumenteista. Tehtäväkuvauksen B1 yhteydessä käytetyistä jatkokyselyjen hakuavaimista 86% oli löydettävissä tuloslistan dokumenteista ja tehtäväkuvauksen C2 yhteydessä käytetyistä hakuavaimista 78%.

Tuloksia hakijoiden turvautumisesta tuloslistan dokumenttien sanoihin vahvistaa myös se, millä kohtaa dokumentteja, kyselyjen hakuavaimet esiintyivät. Molempien tehtäväkuvausten osalta hakuavain löytyi useammin dokumentin tarkasta kohdasta kuin kokonaisesta dokumentista. Tehtäväkuvauksen B1 osalta lähes 70% ja tehtäväkuvauksen C2 osalta hieman yli 60% hakuavaimista esiintyi tarkassa kohdassa dokumenttia. Vaikka tässä tutkimuksessa ei pystytä todistamaan, kuinka hakijat käyttäytyvät tarkastellessaan tuloslistan dokumentteja, tarkastellun elementin laajuudesta voidaan kuitenkin päätellä paljon. Mitä tarkemmasta elementistä hakuavain on löydettävissä, sitä todennäköisemmin se on voinut päätyä juuri katsotusta elementistä kyselyyn.

Suurin osa hakuavaimista oli löydettävissä dokumenttien leipätekstistä. Kuitenkin huomattavan suuri osa hakuavaimista oli löydettävissä joko elementin otsikosta tai elementin alusta. Tehtäväkuvauksen C2 yhteydessä käytetyistä jatkokyselyjen hakuavaimista lähes joka kolmas ja tehtäväkuvauksen B2 yhteydessä käytetyistä hakuavaimista 40% löytyivät joko elementin otsikosta tai elementin alusta. Hakuavaimen esiintyminen niinkin keskeisessä paikassa kuin dokumentin otsikko tai

elementin alku, tukee tämän tutkimuksen oletusta, hakuavaimen päätymisestä tuloslistan dokumentista jatkokyselyyn.

Hakuavainten ilmenemismuodoissa tehtäväkuvausten välillä esiintyi huomattavaa vaihtelua. Tehtäväkuvauksen B1 hakuavaimista valtaosasta (68%) oli löydettävissä yhtenevä ilmaus dokumentista. Lähes 20% oli johdoksia dokumenttien sanoista ja loput 12% olivat hakuavainten variantteja. Tehtäväkuvauksen C2 osalta hakuavainten ilmenemismuodot jakautuivat tasaisemmin eri luokkien kesken. Kyseisen tehtäväkuvauksen hakuavaimista lähes saman verran oli joko yhteneviä ilmauksia (43%) tai ne olivat johdoksia (41%). Loput 16% tehtäväkuvauksen C2 hakuavaimista olivat variantteja dokumenttien sanoihin nähden.

Hakuavaimet, jotka olivat yhteneviä ilmauksia dokumenttien sanoihin nähden, esiintyivät yli puolessa tapauksista dokumenttien leipätekstissä. Kohtalaisen suuri osa yhtenevistä ilmauksista löytyi kuitenkin joko heti dokumentin alusta tai sen otsikosta. Tehtäväkuvauksen B1 yhtenevistä ilmauksista 13% löytyi otsikosta ja 24% dokumentin alusta. Tehtäväkuvauksen C2 yhtenevistä ilmauksista puolestaan 12% esiintyi otsikossa ja 17% dokumentin alussa. Sen lisäksi, että kyselyjen hakuavaimet esiintyivät kohtalaisen usein keskeisessä kohdassa dokumenttia, kuten dokumentin otsikossa, ilmenivät hakuavaimet tällaisessa kohdassa täsmälleen samassa muodossa kuin ne esiintyivät kyselyssä. Tämänkaltaiset tulokset vahvistavat vielä entisestään oletusta, hakuavaimen päätymisestä tuloslistan dokumentista hakijan kyselyyn.

Dokumenttien elementit, joista jatkokyselyjen hakuavaimia oli löydettävissä, arvioitiin pääsääntöisesti relevanteiksi. 90% tehtäväkuvauksen B1 ja 85% tehtäväkuvauksen C2 elementeistä arvioitiin relevanteiksi. Myös Efthimiadis (1996, 134–135) toteaa tutkimuksessaan, että hakuavaimia poimitaan kyselyihin pääsääntöisesti relevanteiksi todetuista dokumenteista.

Hakijoiden antamat kymmenportaiseen relevanssiasteikkoon pohjautuvat relevanssiarviot elementeille vaihtelivat kuitenkin suuresti sekä tehtäväkuvausten sisällä että niiden välillä. Tehtäväkuvauksen B1 osalta ”*erittäin käyttökelpoiseksi ja erittäin spesifiksi*” arvioitujen elementtien lisäksi hakijoiden antamat relevanssiarviot painottuivat asteikon keskivaiheille. Tehtäväkuvauksen C2 osalta hakijoiden arvioimat elementtien relevanssiasteet painottuivat molempiin ääripäihin sekä myös asteikon keskivaiheille. Erityisesti ”*jonkin verran käyttökelpoiseksi ja jonkin verran spesifiksi”* arvioitujen elementtien määrä vaihteli tehtäväkuvauksien välillä huomattavasti. Vertailevaa tietoa edustavan tehtävätyypin osalta hakijat olivat lähes 15 prosenttiyksikköä useammin arvioineet elementin

”*jonkin verran käyttökelpoiseksi ja jonkin verran spesifiksi*”, verrattuna taustatietoa edustavan tehtävätyypin kohdalla. Ylipäänsä taustatietoa edustavan tehtävätyypin osalta hakijat arvioivat elementit keskimäärin relevantimmiksi kuin vertailevan tehtävätyypin kohdalla. Tämän suuntaiset tulokset ovat yhdenmukaisia Tombrosin ym. (2004, 24) tutkimusten kanssa siitä, että erilaisten tehtävätyyppien käyttö vaikuttaa siihen, millaisin kriteerein hakijat arvioivat hakutulostensa relevanssia.

Kymmenportaisen relevanssiasteikon käyttö on voinut myös vaikuttaa käyttäjien suorittamiin relevanssiarvioihin. Tulokset voisivat olla erilaisia, jos käytössä ollut relevanssiasteikko olisi sisältänyt vähemmän arvoasteita. Relevanssiarvioiden painottuminen asteikon keskivaiheille ja ääripäihin voi olla seurausta siitä, ettei hakija ole osannut tehdä selvää eroa eri relevanssiasteikon arvoille. Myös Pharo & Nordlie (2005) totesivat moniarvoisen relevanssiasteikon käytön olevan ongelmallista. Kymmenen luokkaa relevanssiasteikossa, saattoi heidän mukaan olla liian paljon käyttäjille, jotta käyttäjät olisivat pystyneet tekemään yhdenmukaisia relevanssiarvioita. Myös kolmen eri arvon (erittäin, melko ja marginaalinen) käyttämistä käyttökelpoisuuden ja spesifisyyden määrittelemisessä pidettiin tutkimuksessa ongelmallisena.

Tämän tutkimuksen osalta ei voida kuitenkaan vetää yleistettäviä johtopäätöksiä erilaisten tehtäväkuvausten käytön tai moniportaisen relevanssiasteikon käytön vaikutuksesta hakijoiden antamiin relevanssiarvoihin. Tutkimuksessa tarkasteltiin ainoastaan kahta erilaista tehtäväkuvausta. Hakijoiden antamat relevanssiarviot vaihtelivat suuresti myös tehtäväkuvausten sisällä. Tutkimusta olisi mielenkiintoista laajentaa, ottamalla tarkasteluun laajempi otos erilaisia tehtäväkuvauksia.

Elementit, joissa kyselyjen hakuavaimia esiintyi, sijoittuivat hakijoiden tarkasteluissa lähelle uuden kyselyn muodostushetkeä. Tämä vahvistaa ajatusta siitä, että hakijat hyödyntävät tuloslistan dokumenttien sanoja valitessaan uusia hakuavaimia jatkokyselyihinsä. Tuloslistasta katsottujen elementtien sijoitusten tarkastelu kertoi myös hakijoiden tarkastelevan tuloslistan dokumentteja enemmän kuin aikaisemmat tutkimustulokset ovat osoittaneet. (ks. Esim. Spink ym. 2001; Jansen ym. 1998; 2000). Spinkin ym. (2001, 226–229) tutkimustulosten mukaan hieman alle kolmannes hakijoista tarkasteli ainoastaan yhtä ja noin 20% hakijoista tarkasteli vain kahta tuloslistan dokumenttia. Tässä tutkimuksessa osoitettiin hakijoiden tarkastelevan kohtalaisen usein jopa viittä tuloslistan dokumenttia. Yleistettäviä johtopäätöksiä ei voida tämän tutkimuksen osalta kuitenkaan tehdä, sillä elementtejä tarkasteltiin tuloslistassa vain niin pitkälle, kunnes kyselyn hakuavain tunnistettiin elementistä. Jos hakuavain oli löydettävissä heti ensimmäisestä hakijan katsomasta tuloslistan dokumentis-

ta, tuloslistan tarkastelu lopetettiin tähän. Tuloslistasta tarkasteltujen dokumenttien määrien tutkimista olisi mielenkiintoista jatkaa vielä sen jälkeen, kun kyselyn hakuavain on tunnistettu.

Lähteet

Aitchison, J. & Gilchrist, A. & Bawden, D. (1997). Thesaurus construction and use: a practical manual. London: ASLIB.

Alaterä, A. & Halttunen, K. (2002). Tiedonhaun perusteet – osa lukutaitoa. Helsinki :BTJ Kirjastopalvelu.

Bates M. J. (1979). Information search tactics. Journal of the American Society for Information Science 30 (4), 205–214. Saatavilla www-muodossa: <http://www.gseis.ucla.edu/faculty/bates/articles/Information%20Search%20Tactics.html> (Käytetty 8.4.2008).

Bates, M. J. (1987). How to use information search tactics online. Online 11 (3), 45–54.

Borlund, P. (2000). Experimental components for the evaluation of interactive retrieval systems. Journal of Documentation 56 (1), 71–90.

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

Borlund, P. (2003b). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Information Research 8 (3). Saatavilla www-muodossa: < http://informationr.net/ir/8-3/paper152.html>. (Käytetty 11.3.2008).

Borlund, P. & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation 53 (3), 225–250.

Bruza, P. D. & Dennis, S. (1997). Query ReFormulation on the Internet: Empirical Data and the Hyperindex Search Engine. 5th RIAO Conference – Computer Assisted Information Searching on the Internet, 488–499.

Cosijn, E. & Ingwersen, P. (2000). Dimensions of relevance. Information Processing & Management 36 (4), 533–550.

Desmarais, N. (2000). The ABC of XML. The Librarian's guide to the eXtensible Markup Language. Houston: New Technology Press.

Ding, Y. & Foo, S. (2001). Ontology research and development. Part 1 – a review of ontology generation. Journal of Information Science 28 (2), 123–136.

Dowling, T. (2001). Lies, damned lies, and Web logs. Library Journal. Saatavilla www-muodossa: <http://www.libraryjournal.com/article/CA106218> (Käytetty 15.4.2008).

Efthimiadis, E. N. (1992). Interactive query Expansion and Relevance Feedback for Document Retrieval Systems. London. UK: City University.

Efthimiadis, E. N. (1996). Query expansion. Saatavilla www-muodossa: <http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html> (Käytetty 13.3.2008).

Fidel, R. (1984). Online searching styles: A case-study-based model of searching behaviour. Journal of the American Society for Information Science 34 (4), 211–221.

Fidel, R. (1985). Moves on online searching. Online Review 9 (1), 61–74.

Fuhr, N., Lalmas, M., Malik, S. & Szlavik, Z. (2004). INEX 2004 Workshop Pre-Proceedings. Saatavilla www-muodossa: <http://INEX.is.informatik.uni-duisburg.de:2004/pdf/INEX2004PreProceedings.pdf> (Käytetty 5.3.2008).

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5 (2), 199–220. Saatavilla www-muodossa: <http://tomgruber.org/writing/ontolingua-kaj-1993.pdf> (Käytetty 22.4.2008).

Gövert, N., Kazai, G., Fuhr, N. & Lalmas, M. (2003). Evaluating the effectiveness of content-oriented XML-retrieval. Saatavilla www-muodossa: <http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Goevert_etal:03a.pdf>. (Käytetty 5.3.2008).

Harman, D. K. (1992). Relevance feedback revisited. Teoksessa Belkin, N., Ingwersen, P. & Pejtersen, A., M. (toim.) Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. Copenhagen, Denmark, 1–10.

Iivonen, M. (1995). Hakulausekkeiden muotoilun yhdenmukaisuus onlineviitehaussa. Tampereen yliopisto. Acta Universitatis Tamperensis 443.

INEX 2004. Initiative for the evaluation of XML Retrieval. Saatavilla www-muodossa: <http://INEX.is.informatik.uni-duisburg.de:2004/> (Käytetty 9.11.2007).

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

Jansen, B.J., Spink, A., Bateman, J. & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. ACM SIGIR Forum 33 (1), 5–17.

Jansen, B.J., Spink, A. & Saracevic, T. (2000). A study of user queries on the Web. Information Processing and Management 36 (2), 207–227.

Järvelin, K. (1993). Merkkijonot, sanat, termit ja käsitteet informaation haussa. Kirjastotiede ja informatiikka 12 (4), 119–128.

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

Järvelin, K. & Sormunen, E. (1999). Dokumentit kateissa? Tiedontallennus ja haku avuksi. Teoksessa Ilkka Mäkinen (toim.) Tiedontie: johdatus informaatiotutkimukseen. Helsinki: BTJ. Kirjastopalvelu, 110–143.

Karlsson, F. (1998). Yleinen kielitiede. Helsingin yliopisto. Yliopistopaino.

Liu, S., Zou, Q. & Chu, W. (2004). Configurable Indexing and Ranking for XML Information Retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, UK, 88–95.

Lykke Nielsen, M. (2002). The word association method: a gateway to work-task based retrieval. Åbo: Åbo Academi University Press

Magennis, M. & van Rijsbergen, C.J. (1997). The potential and actual effectiveness of interactive query expansion. Teoksessa: Belkin, N. J., Narasimhalu, A. D. & Willett, P. (toim.) ACM SIGIR Forum 31 (20), 324–332.

Malik, S., Tombros, A. & Larsen, B. (2004). Hyrex for INEX iTrack. Teoksessa: Teoksessa: Fuhr, N., Lalmas, M., Malik, S. & Szlavik, Z. (2004). INEX 2004 Workshop Pre-Proceedings. International Conference and Research Center for Computer Science, 264–269 Saatavilla www-muodossa: <http://inex.is.informatik.uni-duisburg.de:2004/pdf/INEX2004PreProceedings.pdf> (Käytetty 9.4.2008).

Marchionini, G. (1995). Information seeking in electronic environments. Cambridge: Cambridge University Press.

Nurmela, M. (2006). Simuloidun tehtävän vaikutus kyselyn muodostukseen: INEX 2004-hankkeen vuorovaikutteisen tiedonhaun tutkimuslinja. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu -tutkielma.

Pharo, N. & Nordlie, R. (2005). Context Matters: An Analysis of Assessments of XML Documents. Oslo.Oslo Univercity College. Faculty of Journalism, Library and Information Science, 238–248.

Rieh, S. Y., & Xie, H. (2005). Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing and Management 42 (2006), 751–768.

Robertson, S. E. & Hancoc-Beaulieu, M. M. (1992). On the evaluation of IR systems. Information Processing and Management 28 (4), 457–466.

Rubin, J. H. (2001). Introduction to Log Analysis Techniques: Methods for Evaluating Network Services. Teoksessa Mc Clure (toim.) Evaluating Networked Information Services: Techniques, Policies and Issues. Medford, NJ. American Society for Information and Technology, 197–212.

Ruthven, I., Lalmas, M. & van Rijsbergen C. J. (2003). Incorporating user search behavior into relevance feedback. Journal of the American Society for Information Science and Technology 54 (6), 528–548.

Saracevic, T. (1996). Relevance reconsidered. Information science: Integration in perspectives. Proceeding of the Second Conference on Conceptions of Library and Information Science. Copenhagen, Denmark, 201–218.

Schamber, L., Eisenberg, M. B. & Nilan, S. (1990) A re-examination of relevance: toward a dynamic, situational definition. Information Processing & Management 26 (6), 755–776.

Sihvonen, A. & Vakkari, P. (2004). Subject knowledge, thesaurus-assisted query expansion and search success. Proceedings of RIAO 2004 Conference. CID, Paris, 393–404.

Spink, A., Wolfram, D., Jansen, M.B.J. & Saracevic, T. (2001). Searching the Web: Public and Their Queries. Journal of the American Society for Information Science and Technology 52 (3), 226–234.

Suomela, S. & Kekäläinen, J. 2005. Ontology as a search-tool: A study of real users' query. Formulation with and without conceptual support. Proceedings of ECIR 2005 Conference. Santiago de Compostela, Spain, 315–329.

Tietohuollon sanasto: suomi, ruotsi, englanti, saksa, ranska. (1993). Tekniikan sanastokeskus ja Tietopalveluseura. Helsinki: Kirjastopalvelu.

Tietotekniikan sanasto. (1990). Helsinki: Tietosanoma Oy.

Tombros, A., Larsen, B. & Malik, S. (2004). The Interactive Track at INEX 2004. Teoksessa: Fuhr, N., Lalmas, M., Malik, S. & Szlavik, Z. (2004). INEX 2004 Workshop Pre-Proceedings. International Conference and Research Center for Computer Science, 24–32. Saatavilla

www.muodossa: <http://inex.is.informatik.uni-duisburg.de:2004/pdf/INEX2004PreProceedings.pdf> (Käytetty 8.4.2008).

Text REtrieval Conference. (2000). Saatavilla www-muodossa: <http://trec.nist.gov/> (Käytetty 18.4.2008).

Vakkari, P. (1999). Tiedonhankinnan tukeminen ja informaatiotutkimus. Teoksessa: Ilkka Mäkinen (toim.) Tiedontie: johdatus informaatiotutkimukseen. Helsinki: BTJ. Kirjastopalvelu, 9–31.

W3C. (2003). XML in 10 points. Saatavilla www-muodossa: <http://www.w3c.tut.fi/translations/xml/xmlin10pts/.> (Käytetty 5.3.2008).

Liitteet

Liite 1. Tehtäväkuvauksen ulkopuoliset hakuavaimet kyselyissä

Tehtäväkuvaus B1	
"eye strain"	kennedy
"motion sick"	kolasinski
"motion sickness"	ltd.
"simulator sickness"	maladies
addiction	motion
affecting	nausea
analysis	percent
blur	percentage
cause	prevalence
causes	problem
computer	problems
cyber	profile
cybersex	quasy
cybersick	queasiness
cyberspace	related
disorder	residual
disorders	sick
disorientation	sickness
dizzy	simulation
environment	simulator
eugenia	statistics
experiencing	strain
games	systems
harm	therapy
health	treating
human	virtuality
illness	vision
immersion	witness

Tehtäväkuvaus C2	
"alan brown"	distibuted
"development cost"	distributed
"development using java"	drawback
"execution speed"	drawbacks
"exploiting python resources"	efficient
"java advantages"	enterprise
"java versus python"	experience
"java vs. python"	faqs
"lage scale development"	features
"man hours"	flexible
"man months"	fortran
"object oriented progaming languages"	framework
"object oriented progamming languages"	future
"rapid application development"	interactive
"advantage"	library
"advantages"	memory
alan	net
appropriate	new
bad	object
benchmark	open
best	performance
big	pitfall
brown	pros
combining	reengineering
compare	scalability
compared	scale
comparing	scaleability
compiler	script
componenets	scripting
computing	source
cons	suitable
conversion	suited
costs	tutorial
cots	using
design	Watters
difference	weakness
disadvantage	vs
disadvantages	vs.