KYSELYNKÄSITTELYMENETELMIEN EVALUOINTITUTKIMUS SUOMALAISEN VERKKOARKISTON TAIVUTUSMUOTOINDEKSIÄ KÄYTTÄEN

Petteri Veikkolainen

Tampereen yliopisto
Informaatiotieteiden yksikkö
Informaatiotutkimus ja
interaktiivinen media
Pro gradu -tutkielma
Joulukuu 2015

TAMPEREEN YLIOPISTO, Informaatiotieteiden yksikkö
Informaatiotutkimus ja interaktiivinen media
VEIKKOLAINEN, PETTERI: Kyselynkäsittelymenetelmien evaluointitutkimus Suomalaisen verkkoarkiston taivutusmuotoindeksiä käyttäen
Pro gradu -tutkielma, 56 s., 3 liites.
Joulukuu 2015

Suomen kielen rikas morfologia aiheuttaa tiedonhaulle haasteita. Jotta tiedonhaku on tuloksellista, täytyy kyselyn sanamuoto saada täsmäämään dokumentissa esiintyvän sanamuodon kanssa. Tässä tutkimuksessa verrataan neljän eri kyselynkäsittelymenetelmän tuloksellisuutta dokumenteista rakennetussa taivutusmuotoindeksissä.

Aiempi suomenkielisellä aineistolla toteutettu tiedonhaun evaluointitutkimus on käyttänyt dokumenttikokoelmina pääasiassa lehtiartikkelikokoelmista rakennettuja testikokoelmia. Tässä tutkimuksessa käytetään artikkelikokoelman sijaan Suomalaisesta verkkoarkistosta rakennettua testikokoelmaa, joka sisältää verkkosivuja joiden sisältö ja laatu vaihtelevat paljon. Tutkielmassa verrattavat menetelmät ovat Frequent case generation 3 (FCG3), Simple word ending based rule generator (SWERG+), Snowball-stemmaus yhdistettynä villiin korttiin sekä käsittelemättömät kyselyt.

Tämän tutkimuksen tutkimusmenetelmä on tiedonhaun laboratoriomallin mukainen testaus. Sen suorittamiseksi Suomalaisesta verkkoarkistosta oli rakennettava testikokoelma. Testikokoelmaan valittiin lopulta 16 hakuaihetta, joista muodostetuilla lyhyillä kyselyillä suoritettiin kyselyajot. Ajojen tulokset mitattiin tarkkuudella kymmenen ensimmäisen tulosdokumentin kohdalla sekä kumuloituvan hyödyn mittarilla.

Tutkimuksessa havaittiin FCG3-menetelmän tuottavan perustasona toimineita käsittelemättömiä kyselyitä parempia tuloksia. Sen sijaan aiemmassa tutkimuksessa hyvin suoriutunut SWERG+-menetelmä ei tuottanut tässä tutkimuksessa perustasoa parempia tuloksia. Snowball-stemmaus yhdistettynä villiin korttiin taas tuotti perustasoa heikompia tuloksia.

Avainsanat: evaluointi, Kansalliskirjasto, kyselynkäsittelymenetelmä, morfologia, tiedonhaku, verkkoarkisto

Sisällysluettelo

1 JOHDANTO

Nykypäivän yhteiskunnassa tietoa on tarjolla paljon ja haluttua yksittäistä tietoa voi olla vaikeaa paikallistaa kaiken olemassa olevan tietomäärän joukosta. Tietoa tarvitaan ja etsitään jatkuvasti. Tiedon löytymistä helpottavia keinoja tutkitaan ja kehitetään tiedonhaun tutkimusalalla.

Tiedonhaun tutkimus on aikojen saatossa jakautunut tunnistettavasti kahteen eri koulukuntaan, luonnontieteelliseen niin kutsuttuun laboratoriotutkimukseen ja yhteiskuntatieteelliseen käyttäjän huomioivaan tutkimukseen. Laboratoriotutkimuksessa on tyypillisesti kyse siitä, kuinka hyvin tutkittava hakumenetelmä löytää relevanteiksi tiedetyt dokumentit tietystä dokumenttitietokannasta. Käyttäjäkeskeinen tutkimus taas selvittää kuinka hyvin käyttäjä kokee hakumenetelmän palvelevan häntä ja millaista vuorovaikutusta käyttäjän ja järjestelmän välillä tapahtuu.

Vaikka tämän tutkimuksen pääasiallisena mittarina toimii laboratoriotutkimuksen idean mukainen hakumenetelmien vertailu kvantitatiivisilla menetelmillä, pyritään siinä myös ottamaan huomioon käyttäjän näkökulma. Käyttäjän näkökulma on pyritty huomioimaan selvittämällä tutkimuksen kohteena olevan hakujärjestelmän käyttötapaukset aiemman tutkimuksen ja hakujärjestelmän oletettujen käyttötapausten avulla.

Tutkimuksen kohteena on Kansalliskirjaston ylläpitämän Suomalaisen verkkoarkiston tiedonhakutoiminto, jonka tarkoituksena on tuoda löydettäviksi vuodesta 2006 lähtien arkistoituja suomalaisia ja suomalaisille suunnattuja verkkosivuja. Kansallista kulttuuriperintöä säilyttävän toiminnan osana Suomalainen verkkoarkisto tarjoaa mahdollisuuden kulttuurintutkimukseen. Lisäksi verkkoarkistoja on hyödynnetty oikeudessa todistusmateriaalina. Arkistoitujen verkkosivujen löytämisen parantamiseksi tässä tutkimuksessa vertaillaan, miten eri tavat käsitellä kyselyitä vaikuttavat haun tuloksellisuuteen.

Kyselyn käsittelyn varioimisella pyritään tässä tutkimuksessa ratkaisemaan ongelmaa, jossa kyselyn sana ja etsityssä dokumentissa esiintyvä sana on esitetty eri muodoissa. Erityisesti vahvasti taipuvissa kielissä tällainen ongelma on yleinen. Aiempi tutkimus on pääasiassa suuntautunut dokumentin sanojen käsittelyyn perustuviin menetelmiin. Tässä tutkimuksessa tätä ongelmaa sen sijaan pyritään ratkaisemaan tuotantokäytössä oleviin

laajoihin tiedonhakujärjestelmiin paremmin sovellettavissa olevien kyselyn sanojen käsittelyyn perustuvilla menetelmillä.

Tutkimuksen tavoitteena on:

1. Selvittää, kuinka tuloksellisia suomen kielen morfologian aiheuttamien ongelmien hallintaan tarkoitetut sovellukset ovat tiedonhaussa Suomalaisessa verkkoarkistossa.

2. Selvittää, tuottaako verkkoarkiston käyttö testikokoelmana erilaisia evaluointituloksia kuin aiemmassa tutkimuksessa käytetyt artikkelitestikokoelmat.

3. Auttaa Suomalaisen verkkoarkiston kehittämisessä.

Tutkimuksen toisessa luvussa esitellään tiedonhakujärjestelmien rakennetta. Kolmannessa luvussa esitellään tiedonhaun evaluointitutkimusta ja perehdytään erityisesti tutkimuksessa käsiteltäviin luonnollisen kielen aiheuttamiin ongelmiin tiedonhaussa sekä tutkimuksessa suoritetun testikokoelman rakentamiseen. Lisäksi kolmas luku sisältää katsauksen aiempaan tutkimukseen, joka on jaettu luonnollisen kielen käsittelymenetelmiä vertaavaan tutkimukseen ja verkkoarkistojen käyttöä käsittelevään tutkimukseen. Neljännessä luvussa esitetään tutkimusasetelma ja tulokset. Viidennessä luvussa on keskustelu ja johtopäätökset.

2 TIEDONHAKUJÄRJESTELMÄ

Tiedonhakujärjestelmällä tarkoitetaan informaation tallentamiseen tarkoitettua järjestelmää, jonka avulla erilaiset käyttäjäryhmät voivat levittää, selata, etsiä ja prosessoida informaatiota (Salton & McGill 1987). Tiedonhakujärjestelmien tarkoituksena on auttaa käyttäjää löytämään relevanttia tietoa järjestetystä dokumenttikokoelmasta (Chowdhury 2010, 1). Tässä luvussa esitellään tekstitiedonhakuun tarkoitetun tiedonhakujärjestelmän rakennetta. Tämän jälkeen esitellään erityisesti dokumentti- ja kyselyesitysten muodostamista ja luonnollisen kielen käsittelyä osana tiedonhakujärjestelmän rakennetta. Lopuksi perehdytään dokumentti- ja kyselyesitysten vertaamiseen ja erityisesti esitellään tarkemmin tutkimuksessa käytetyn tiedonhakujärjestelmän perustana olevaa tiedonhaun vektorimallia ja tf*idf-painotusta.

2.1 Tiedonhakujärjestelmän rakenne

Yksinkertaistettuna tiedonhakujärjestelmän rakentumisen voidaan katsoa koostuvan *indeksointiprosessista* (kuva 1) ja *hakuprosessista* (kuva 2). Indeksointiprosessiin lasketaan mukaan *indeksin* rakentamiseen käytettävät toiminnot ja rakenneosat. Indeksillä tarkoitetaan listausta, joka sisältää kaiken dokumenttikokoelman sisältämän tiedon johon tiedonhaun voi kohdistaa. Indeksiä voisi verrata esimerkiksi kirjan lopussa olevaan sanahakemistoon. Tiedonhakujärjestelmän hakuprosessiin kuuluvat ne toiminnot ja rakenneosat, joiden avulla tiedonhakijan tiedontarve saadaan esitettyä tietojärjestelmälle sekä tuotettua indeksin tietojen avulla haun tuottamien tulosten esittäminen hakijalle.

Kuva 1, indeksointiprosessi havainnollistettuna Croftin ja kumppaneiden (Croft et al. 2010, 15) pohjalta

Indeksointiprosessin toiminnot ja rakenneosat:

1. Alkuperäiset dokumentit. Alkuperäiset dokumentit ovat tiedonhakujärjestelmän perusta. Niiden ominaisuudet asettavat vaatimuksensa indeksin rakentamiselle ja tiedonhaulle. Näitä ominaisuuksia ovat keskeisimmät dokumenttien tyypit (esimerkiksi verkkosivu, lehtiartikkeli, kuva), tallennusmuoto ja kokoelman päivittyminen.

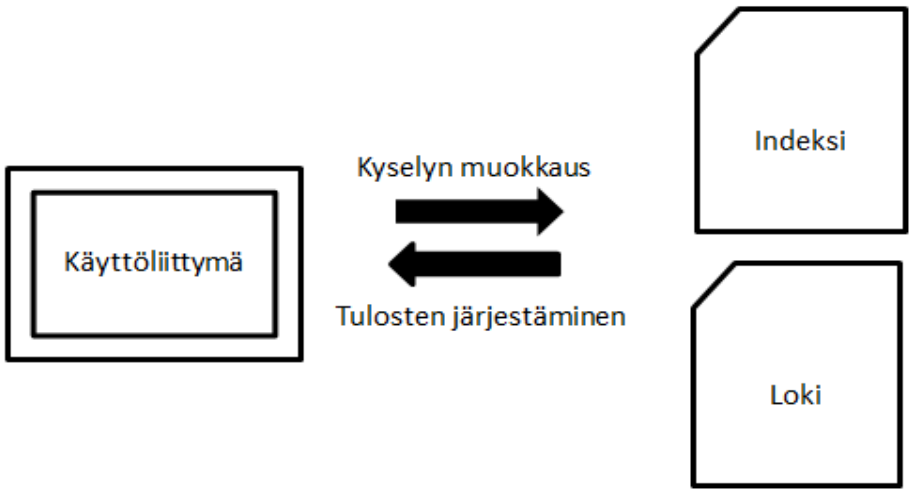
2. Dokumenttien kerääminen. Dokumenttien kerääminen tapahtuu joko automaattisesti hakurobotin avulla tai manuaalisesti lisäämällä uusi dokumentti kokoelmaan.

3. Dokumenttikokoelma. Dokumenttikokoelma on kerättyjen alkuperäisten dokumenttien varasto. Tallentamalla koko dokumentti varastoon, voidaan dokumentin saatavuus varmistaa vaikka se katoaisi alkuperäisestä esiintymispaikasta.

4. Dokumenttien muokkaaminen. Dokumentit on muokattava tarpeeksi yhtenäiseen muotoon, jotta niistä voi rakentaa indeksin. Tekstitiedonhaussa dokumentit yleensä muutetaan tiettyyn tekstitiedostomuotoon ja muokataan merkistökoodaus yhteneväksi.

5. Indeksin luominen. Indeksin luomisessa indeksiin kerätään talteen jokaisen dokumentin ominaisuudet. Yksinkertaisimmillaan indeksissä on tieto, missä dokumenteissa tietty sana esiintyy. Perinteisessä tiedonhakututkimuksessa yleistä indeksin luomisessa on sanojen käsittely siten, että saman sanan eri esiintymismuodot muokataan yhtenevään muotoon.

6. Indeksi. Indeksi sisältää kaiken tiedon mitä dokumenteista voidaan hakea. Tällaisia tietoja ovat esimerkiksi sanojen esiintymismäärät ja -kohdat dokumenteittain.



Kuva 2, hakuprosessi havainnollistettuna Croftin ja kumppaneiden (Croft et al. 2010, 16) pohjalta

Hakuprosessin toiminnot ja rakenneosat:

1. Käyttöliittymä. Käyttöliittymän avulla tiedonhakija voi tehdä kyselyitä tiedonhakujärjestelmään ja selata haun antamia tuloksia.

2. Kyselyn muokkaaminen. Kyselyssä esiintyvät sanat pyritään saamaan muotoon, jossa ne vastaavat indeksissä olevien sanojen muotoa.

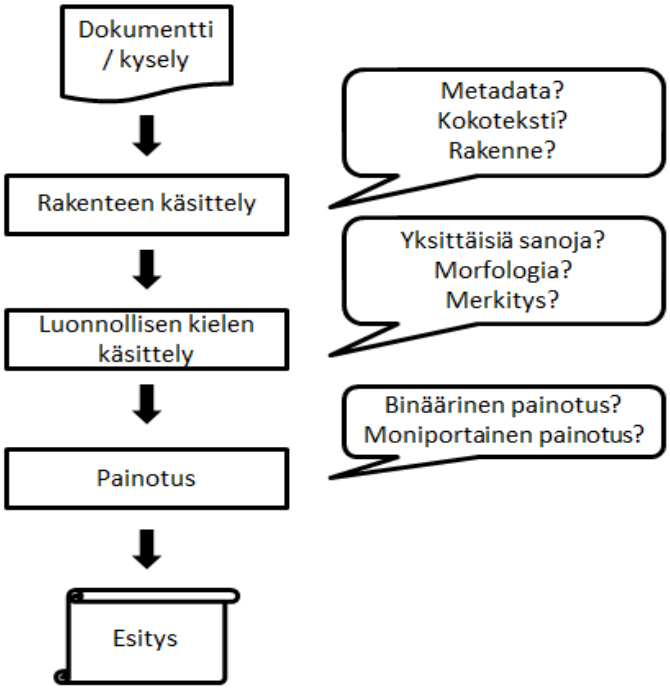
3. Tulosten järjestäminen. Tulokset järjestetään lokitietojen ja jonkin matemaattisen mallin mukaan siten, että mitä hyödyllisemmäksi arvioitu dokumentti on, sitä korkeammalla se on tuloslistalla.

4. Lokitiedot. Lokitietojen avulla voidaan tyypitellä käyttäjiä ja saada tietoa esitetyistä kyselyistä ja selatuista tuloksista. Lisäksi niiden avulla voidaan opettaa järjestelmälle suosituimmat tai viimeksi tehdyt kyselyt. Näiden tietojen avulla voidaan parantaa haun tehokkuutta ja tuloksellisuutta.

2.2 Dokumentti- ja kyselyesitysten muodostaminen

Tiedonhaun tulosdokumenttien määräytyminen perustuu yksinkertaisimmillaan *kyselyesityksen* ja *dokumenttiesityksen* vertailuun. Kyselyesityksellä tarkoitetaan käyttäjän antamista hakusanoista järjestelmän luomaa esitystä ja dokumenttiesityksellä järjestelmän luomaa esitystä dokumentista. Ingversen ja Järvelin (2005, 125–128) jakavat esityksenmuodostusprosessin kolmeen eri vaiheeseen, jotka on havainnollistettu

kuvassa 3. Rakennevaiheessa määritetään miten mahdollinen metadata ja dokumentin tai kyselyn sisältö järjestetään. Luonnollisen kielen käsittelyvaihe määrittää kuinka sisällön merkityksen tulkinta otetaan huomioon vai käsitelläänkö kyselyn hakusanoja yksittäisinä merkkijonoina. Painotusvaiheessa päätetään kuinka järjestelmä huomioi tuloslistaa rakentaessa sanojen suhteellisen esiintymismäärän. Esimerkkinä painotuksen merkityksestä olkoon, että tiedonhaun kannalta on luultavasti hyödyllisempää, mikäli haku ”graniitti kivi” tuottaisi tuloslistan alkupäähän erityisesti graniitista kertovia dokumentteja. Olettaen, että graniitista kertovia dokumentteja on vähemmän kuin yleisesti kivistä kertovia dokumentteja.



Kuva3, esityksen muodostaminen (Ingversen & Järvelin 2005, 127) sovellettuna

2.3 Luonnollisen kielen käsittely

Perinteinen vaihtoehto yhtenäisen merkkijonoesityksen muodostamiseksi on ollut karsia dokumentin ja kyselyn esityksestä mahdollisimman yksinkertainen. Toinen, uudempi vaihtoehto on säilyttää dokumentin esitys rikkaana ja yrittää rikastuttaa kyselyn esitysmuotoa. Voorheesin (1999) esimerkkiä perinteiselle englanninkieliselle esitysmuotoon saattamiselle mukaillen esitetään suomenkielisen dokumentin tai kyselyn esitys seuraavalla esimerkillä. Tässä perinteistä menetelmää kuvaavassa tapauksessa ei

ole merkitystä onko kyseessä dokumentin vai kyselyn esitys. Molempien esitysmuoto luodaan samalla periaatteella:

> ”Jo joutui armas aika, ja suvi suloinen.”

Perinteinen esitysmuotoistaminen voisi tuottaa tästä virkkeestä seuraavan esitysmuodon:

> ”joutu arma aik suv suloi”

Esitysmuoto muodostettiin:

1. Poistamalla välimerkit, ja jakamalla koko merkkijono useiksi pienemmiksi merkkijonoiksi välilyöntien perusteella.

2. Poistamalla yleiset, tiedonhaun kannalta merkityksettömät sulkusanat (jo, ja).

3. Katkaisemalla merkkijonot oletetun sanan vartalo-osan päätyttyä.

Vastaavasti rikastettuun kyselyn esitykseen perustuva dokumenttiesitys tapahtuisi samoin, paitsi sanoja ei katkaistaisi:

> ”joutui armas aika suvi suloinen”

Rikastettu kyselyesitys taas muodostuisi seuraavanlaiseksi, mikäli rikastusmenetelmä perustuisi vaikkapa kolmen yleisimmäksi oletetun taivutusmuodon päättelemiseen ja lisäämiseen:

> ”joutui joudutaan joutuu armas armaan armasta aika ajan aikaa suvi suvin suvia suloinen suloisen suloista”

Kyselyesityksen sanamuodot eivät välttämättä ole tosiasiassa yleisimmin suomen kielessä esiintyviä, vaan ovat tässä vain esimerkkinä. Dokumentti- ja kyselyesitysten muodostamiseen käytettäviä keinoja esitellään tarkemmin luvussa 3.2.

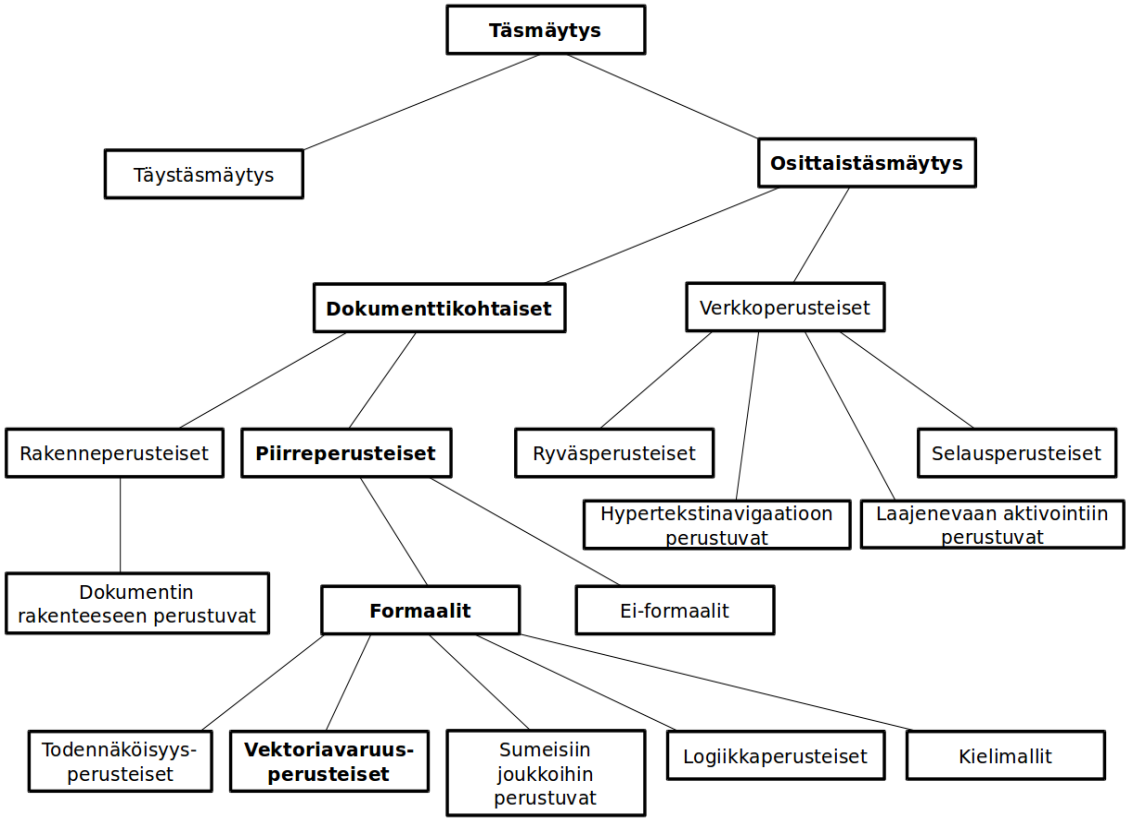
2.4 Täsmäytys

Täsmäytyksessä verrataan kyselyesityksen ja dokumenttiesitysten yhtenevyyttä. Käytännössä tämä tarkoittaa yleensä, että selvitetään missä kokoelman dokumenteista annetut hakusanat esiintyvät. Täsmäytys voidaan jakaa binääriseen *täystäsmäytykseen* ja tuloksia relevanssin asteen perusteella järjestävään *osittaistäsmäytykseen*.

Täsmäytysmenetelmien suhteet on esitetty kuvassa 4, jossa perehdytään osittaistäsmäytysmenetelmien suhteisiin. Täystäsmäytysmenetelmien osalta vastaava kuva ja lisätietoa löytyy Ingversenin ja Järvelinin (2005, 116) kirjasta.

Täystäsmäyttävässä täsmäytyksessä hakusanojen esiintymistä dokumenteissa tarkastellaan Boolen logiikan perusteella. Käyttäjän antaman tai järjestelmän oletusoperaattorin perusteella hakusanan joko täytyy (AND) tai se voi (OR) tai se ei saa (NOT) esiintyä dokumentissa. Lisäksi hakusanojen suhteita voidaan yleensä esittää käyttämällä sulkuja. Täystäsmäytys tuottaa tulokseksi vain ehdot täyttäneitä dokumentteja, joita ei ole lajiteltu järjestykseen relevanssin perusteella. Lisäksi täystäsmäyttävissä perinteisissä hakujärjestelmissä Boolen operaattorien käyttäminen vaatii loogista päättelyä sekä vaivannäköä käyttäjältä. Näiden syiden takia täystäsmäytyksen sijaan tiedonhakujärjestelmissä käytetään yleensä osittaistäsmäyttäviä menetelmiä.

Osittaistäsmäyttävät menetelmät ovat jaettavissa dokumenttikohtaisiin ja verkkoperusteisiin menetelmiin. Tiedonhaun evaluointitutkimus on vahvasti keskittynyt dokumenttikohtaisiin menetelmiin ja tarkemmin eriteltynä piirreperusteisiin formaaleihin menetelmiin. Formaalien menetelmien avulla yksittäisille dokumenteille voidaan laskea menetelmän mukainen arvo, joka kertoo kuinka samankaltainen dokumentti on suhteessa hakukyselyyn.



Kuva 4, täsmäytysmenetelmien suhteet sovellettuna aiempaan tutkimukseen Ingversenin ja Järvelinin (2005, 117) mukaan

Formaaleista menetelmistä erityisesti todennäköisyysperusteiset menetelmät ja vektoriavaruusperusteiset menetelmät ovat laajalti käytössä tiedonhakujärjestelmissä. Käytännön sovelluksissa hakujärjestelmistä on huomattava, että ne eivät yleensä perustu vain yhteen kuvassa 4 esitetyistä menetelmistä, vaan ne soveltavat useampaa eri menetelmää sekä huomioivat myös kuvan 4 listauksen ulkopuolisia tekijöitä. Tällaisia ovat esimerkiksi ajankohtaisuus tai käyttäjän huomioiminen, jotka ovat tärkeitä varsinkin web-tiedonhaun tuloksellisuuden kannalta (Alonso et al. 2007; Agichtein et al. 2006). Kuva 4 havainnollistaakin tieteellisten paradigmojen mukaisten menetelmien keskenäisiä suhteita.

Tässä tutkimuksessa keskitytään formaaleista menetelmistä erityisesti vektoriavaruusperusteisiin menetelmiin, joihin tutkimuksessa käytetyn järjestelmän tiedonhakutoiminto perustuu. Kyseisessä *tiedonhaun vektorimallissa* kyselyistä ja dokumentista muodostetaan vektorit, joiden yhtenevyyden perusteella dokumentin relevanssia suhteessa kyselyyn mitataan.

Vektorimalli on ensimmäinen osittaistäsmäytysparadigma ja huomattava osa nykyisinkin käytössä olevista tiedonhakusovelluksista perustuu siihen. Vektorimallissa dokumentti d_x voidaan esittää vektorina

$$v_{dx} = (d_{x1,} d_{x2,} \ldots, d_{xn}) \quad (1)$$

jossa d_{xn} on n-sijaluvun sanan paino dokumentissa d_x. Vastaavasti kysely q_x voidaan esittää vektorina

$$v_{qx} = (q_{x1,} q_{x2,} \ldots, q_{xn}) \quad (2)$$

jossa q_{xn} on n-sijaluvun hakusanan paino kyselyssä q_x. Vektoriesityksinä dokumentin ja kyselyn samankaltaisuus voidaan mitata lukujen 0 ja 1 välisenä arvona laskemalla vektoreiden välisen kulman kosini.

Jo Salton (Salton et al. 1975) hyödynsi kehittämässään vektorimallissa olemassa olevaa havaintoa siitä, että mitä harvemmassa kokoelman dokumentissa sana esiintyy, sitä arvokkaampi se on hakumenetelmän tuloksellisuuden parantamisessa (Spärck-Jones 1972). Tätä kutsutaan *käänteiseksi dokumenttifrekvenssiksi* (inverted document frequency, idf). Idf kerrottuna termin esiintymismäärällä dokumentissa (term frequence, tf) on klassikoksi muodostunut tapa (tf*idf) laskea sanan paino dokumenttiesityksessä niin vektorimallissa kuin todennäköisyysmallissakin. Siihen perustuvat painolaskukaavat ovat nykyisinkin käytössä laajasti tekstitiedonhakujärjestelmissä. Yksi vektorimallin mukainen tf*idf-painotukseen perustuva menetelmä esitellään tarkemmin tutkimusasetelmassa luvussa 4.3.

3 TIEDONHAKUJÄRJESTELMIEN TUTKIMUS

Tässä luvussa esitellään tiedonhakututkimuksen kehitystä ja edetään tämän tutkimuksen varsinaisen aiheen eli luonnollisesta kielestä johtuvien tiedonhaun ongelmien käsittelyyn. Luonnollisesta kielestä johtuvissa ongelmissa syvennytään tarkemmin morfologiaan ja siitä aiheutuvien ongelmien ratkaisemiseksi kehitettyyn tiedonhaun teoriaan ja sovelluksiin. Tutkimusongelmien jälkeen esitellään tiedonhakumenetelmien evaluoimisen pohjana olevaa teoriaa ja ohjeistusta, joiden perusteella tutkimuksessa käytetty testikokoelma rakennettiin ja saadut tulokset käsiteltiin. Lopuksi esitellään aiempaa suomen kielen käsittelymenetelmiä vertailevaa tutkimusta ja verkkoarkistojen tiedonhakua käsittelevää tutkimusta.

3.1 Merkkijonoihin keskittyvä tiedonhaku

Tiedonhaun tutkimuksen voidaan katsoa saaneen alkunsa 1950–1960-luvuilla, jolloin syntyi tarve kehittää tuloksellisempia ratkaisuja dokumenttien löytymiseen tietojärjestelmistä. Tuolloin toteutetussa Cranfield-projektissa rakennettiin ensimmäinen merkittävä testikokoelma. Cranfield-projektia seurasivat muun muassa MEDLARS- ja STAIRS-tutkimukset, joissa Cranfield-projektin tapaan selvitettiin dokumenttien indeksointimenetelmän vaikutusta tiedonhaun tuloksellisuuteen. (Chowdhury 2010, 296–308.)

Tiedonhaun kehittämistutkimus sai merkittävän sysäyksen täysin tietokonepohjaisiin hakujärjestelmiin, kun Salton (1986) esitti, ettei ihmisen suorittamaa intellektuaalia indeksointia voida pitää tiedonhaun tuloksellisuuden kannalta parempana kuin tietokoneen automaattisesti muodostamaa indeksointia. Automaattisen indeksin rakentamisen myötä tietokoneen täytyi itsenäisesti päätellä, minkälaisia dokumentteja käyttäjä haluaa löytää tietyllä hakusanalla. Tähän päättelyyn tietokone käyttää perinteisesti kolmea komponenttia: annettua kyselyä, tietojärjestelmän dokumenteista rakennettua indeksiä sekä kaavaa, jolla kyselyn hakusanoja verrataan indeksin sanoihin (Ingversen & Järvelin 2005, 172).

Aiemmin datan säilyttäminen ja hakeminen vaati suhteellisesti paljon enemmän resursseja kuin nykyään, ja tiedonhaun tutkimuksessa pyrittiin rakentamaan taloudellisia indeksejä (kts. Alkula 2000, 48). Tyypillisesti dokumentissa esiintyvät sanat katkaistiin,

jolloin indeksiin tuli vähemmän yksittäistä sanaa esittäviä merkkijonoja. Vastaavasti annetussa kyselyssä esiintyvät hakusanat katkaistiin, jotta ne täsmäisivät indeksissä oleviin sanoihin. Tällaisia sanan esiintymismuotojen vähentämiseen pyrkiviä menetelmiä kutsutaan *reduktiivisiksi menetelmiksi*.

Perinteisillä reduktiivisilla menetelmillä tuotetut indeksit eivät kuitenkaan ole ainakaan täysin riittäviä täyttämään isojen tiedonhakujärjestelmien tarpeita. Tällaiset järjestelmät käyttävät usein taivutusmuotoindeksiä. Esimerkiksi Googlen indeksi vaikuttaa olevan taivutusmuotoindeksi, mutta sen tarkka toimintalogiikka on useiden kaupallisten sovellusten tapaan salainen.

Taivutusmuotoindeksin haasteena tiedonhaussa on, että yksittäinen sana esiintyy indeksissä yleensä useassa eri merkkijonomuodossa. Perinteisiä reduktiivisia menetelmiä on hankala soveltaa tällaiseen indeksiin. Taivutusmuotoindeksiä varten on kehitetty *generatiivisia menetelmiä*, joissa yksittäisestä hakusanan merkkijonoesityksestä muodostetaan eri merkkijonoja (yleensä sanojen taivutusmuotoja) joiden oletetaan täsmäävän indeksissä esiintyviin useassa eri muodossa esiintyviin kirjoitusmuotoihin.

3.2 Luonnollisesta kielestä johtuvat ongelmat

Valitettavan usein tiedonhakija törmää ongelmaan, jossa tietyllä hakusanalla ei löydy toivottua tietoa. Tällainen ongelma johtuu usein siitä, että hakijan käyttämä hakusana esiintyy eri taivutusmuodossa tai tyystin eri sanana etsityissä dokumenteissa tai kyseisellä hakusanalla on useita eri merkityksiä. Esimerkiksi sana "hiiri" voi viitata eläimeen tai näyttöpäätteessä näkyvän kursorin ohjaamiseen tarkoitettuun laitteeseen. Tällaista ilmiötä, jossa kahdella eri asiaa tarkoittavalla sanalla on sama kirjoitusasu, kutsutaan homografiaksi. Ingversen ja Järvelin (2005, 151) tunnistavat homografian lisäksi yhteensä yksitoista tiedonhakua vaikeuttavaa luonnollisen kielen ominaisuutta. Näistä ominaisuusluokista kolmannes, eli affiksit, sanojen taipuminen, sanojen johtaminen sekä kirjoitusvirheet ovat peräisin kielen *morfologisista* ominaisuuksista. Morfologialla tarkoitetaan kielitieteen tutkimusalaa, jossa tutkitaan sanoja ja niiden rakennetta (Alkula 2000, 30).

Etenkin suomen kielen kaltaisissa morfologisesti rikkaissa kielissä morfologiasta johtuvat ongelmat ovat yksi suurimmista tiedonhaun ongelmatekijöistä (Alkula 2000,

48). Yhdellä substantiivilla voi suomen kielessä olla varovaisen arvion mukaan 2000 eri kirjoitusasua, kun taas adjektiivilla niitä voi olla vielä useita tuhansia enemmän ja verbeillä peräti 12000 (Karlsson 2008).

Tässä alaluvussa perehdytään ensin tiedonhaun tutkimuksessa kehitettyihin keinoihin ja sitten niistä johdettuihin sovellutuksiin morfologiasta johtuvien ongelmien ratkaisemiseksi.

3.2.1 Keinoja morfologian aiheuttamien ongelmien ratkaisemiseksi

Yksi keino morfologian aiheuttamien ongelmien voittamiseen on *jokerimerkit* (wild cards). Jokerimerkki on apukeino, jolla tiedonhakija voi oman harkintansa mukaan katkaista tai korvata tietyn osan kyselyn merkkijonosta jokerimerkillä. Yleensä yhden merkin korvaavaa jokerimerkkiä merkitään kysymysmerkillä (?) ja useita merkkejä korvaavaa jokerimerkkiä asteriskillä (*). Tällöin esimerkiksi hakumerkkijonoilla "kon*" ja "konn?" löytyisi indeksin merkkijono "konna", mutta hakumerkkijono "kon?" ei löytäisi kyseistä merkkijonoa. Jokerimerkkien hyödyntäminen vaatii kuitenkin käyttäjältä merkittävän määrän vaivannäköä ja päättelykykyä, eikä niiden käyttö ole enää kovin suosittua.

Toinen keino on merkkijonojen katkaiseminen eli *stemmaus* (stemming). Stemmauksessa yksittäiset merkkijonot katkaistaan katkaisualgoritmin perusteella. Esimerkkialgoritmi voisi tuottaa merkkijonoista "konna" ja "konnittain" stemmatun muodon "konn". Stemmauksen onkin todettu toimivan erityisen tehokkaasti erityisesti suomen kielen kaltaisissa runsaasti taipuvissa kielissä (Tomlinson 2004; Hollink et al. 2004). Stemmauksen varjopuolena on se, että tarkkoja merkkijonotason hakuja ei voida suorittaa eli sanaa ei voi etsitä sen alkuperäisen taivutusmuodon perusteella.

Kolmas keino morfologian hallitsemiseen on sanan sanakirjamuotoon saattaminen eli *lemmaaminen* (lemmatization). Lemmauksessa merkkijonot muokataan sanakirjamuotoon perusmuotoistamisalgoritmin ja kielikohtaisen sanakirjan avulla. Tarkoituksena on saada esimerkiksi merkkijonot "konnani" ja "konnasta" muotoon "konna". Lemmauksen ongelmana on erityisesti homografia, sanakirjassa esiintymättömät erisnimet sekä yhdyssanat (Ingversen & Järvelin 2005, 154).

Neljäs keino on *sumeat merkkijonomenetelmät*. Sumeissa merkkijonomenetelmissä merkkijonot jaetaan osiin ja verrataan kuinka monta yhteistä osaa verrattavilla

merkkijonoilla on. Sumeiden merkkijonomenetelmien oletuksena on, että merkkijonomuotojen yhtenevyys tarkoittaa merkityksen yhtenevyyttä (Robertson & Willett 1988). Esimerkiksi sanat ”konna” ja ”konnan” olisivat varsin yhteneviä, mikäli niitä vertailtaisiin kolmen merkin merkkijonopätkissä. Konna-sanan merkkijonot olisivat ”kon”, ”onn” ja ”nna” sekä konnan-sanan ”kon”, ”onn”, ”nna” ja ”nan”. Yhtenevyyttä voidaan mitata esimerkiksi jakamalla parin molemmissa sanoissa esiintyvien merkkijonojen määrä parissa esiintyvien kaikkien uniikkien merkkijonojen määrällä, mikä tässä tapauksessa tuottaisi yhtenevyydeksi 3 / 4 eli 0,75.

Viides ja uusin menetelmä morfologian hallintaan on *sanamuotojen generointi* (kts. esim. Kettunen & Arvola 2012). Sanamuotojen generoinnissa pyritään dokumentissa esiintyvä merkkijono löytämään generoimalla hakusanasta useita eri merkkijonoja, joiden oletetaan olevan sanan esiintymismuotoja dokumenteissa. Esimerkiksi hakusanasta ”konna” voitaisiin generoida muodot ”konna”, ”konnan”, ”konnaa”, jolloin hakusanalla ”konna” löytyisi oletettu dokumentissa esiintyvä muoto ”konnan”.

3.2.2 Käytännön sovelluksia suomen kielen morfologian hallintaan

Suomalaisen luonnollisen kielen morfologian hallintaan suosituimmat menetelmät ovat erityisesti tutkimuksessa olleet stemmaukseen perustuva Snowball-stemmeri (Snowball 2015) ja lemmaukseen perustuva FINTWOL (Lingsoft 2015). Viime aikoina on kehitetty hakusanojen esiintymismuotoja generoivat Frequent case generation (FCG, kts. Kettunen 2008) sekä siitä johdetut Simple word ending based rule generator (SWERG, SWERG+, kts. Kettunen & Arvola 2012) -menetelmät. Seuraavaksi esitellään tarkemmin taivutusmuotoindeksiin hyödynnettävissä olevat ratkaisut Snowball-stemmeri sekä FCG- ja SWERG-menetelmät että hieman suppeammin perusmuotoistettuun indeksiin hyödynnettävissä oleva FINTWOL-lemmaus. Aiempia tutkimustuloksia kyseisten menetelmien vertailuista esitellään luvun lopussa.

FCG-menetelmän ideana on laajentaa hakukyselyssä olevia substantiiveja, joiden voidaan katsoa olevan tiedonhaun kannalta selkeästi merkityksellisen sanaluokka (Kettunen & Airio 2006). FCG-menetelmissä hakusanasta tuotetaan sen yleisimmin kielessä esiintyviä sijamuotoja. Tutkimuksessa FCG-menetelmällä on käsitelty myös adjektiiveja, jotka ainakin suomen kielessä ovat substantiivien tapaan sijamuodoissa taipuvia. FCG-menetelmien tuloksellisuutta on testattu myös ainakin ruotsin, saksan sekä venäjän kielellä tehtävässä tiedonhaussa.

Kettusen ja Airion (2006) tutkimuksen mukaan suomalaisessa sanomalehtikirjoituksessa esiintyvistä substantiiveista 84–88 % esiintyvät joko nominatiivissa, genetiivissä, partitiivissa, inessiivissä, elatiivissa tai illatiivissa. Tarvittaessa FCG-menetelmän voi sovittaa dokumenttikokoelman ominaisuuksien perusteella tuottamaan sellaisia taivutusmuotoja, joiden on todettu tai voidaan arvella olevan merkittäviä kokoelmassa.

FCG-menetelmästä on tutkittu neljää eri versiota, jotka nimetään lisäämällä lyhenteen perään generoitavien muotojen määrä. FCG3 tuottaa yksikön nominatiivin, genetiivin ja partitiivin, kun taas FCG6 tuottaa samat sijamuodot myös monikossa. FCG9 tuottaa noiden muotojen lisäksi yksikön inessiivin, elatiivin ja illatiivin, sekä FCG12 kyseiset sijamuodot myös monikossa. (Kettunen & Airio 2006.)

FCG-menetelmä vaatii toimiakseen hakusanan lemman. FCG-menetelmän operationaalista toimintaa on mallintanut Kettunen (2008) toteuttamalla luonnollisen kielen hakulauseista laajennettuja hakulauseita, joissa substantiivit ja adjektiivit lemmataan automaattisesti ja niihin generoidaan FCG-menetelmän avulla taivutusmuotoja. Kyseisessä toteutuksessa generointi tapahtui myös niissä tapauksissa, joissa lemmaus-menetelmä ei tunnistanut sanaa.

SWERG-menetelmät ovat Kettusen ja Arvolan (2012) kehittämiä generatiivisia menetelmiä suomen kielen morfologian hallintaan. Näiden menetelmien on havaittu tutkimuksessa yltävän hyvin lähelle sanakirjapohjaisen FINTWOL-lemmauksen tulostasoa ja jopa toimivan tuloksellisemmin kuin Snowball-stemmaus (Kettunen & Arvola 2012). Vaikka SWERG-menetelmät eivät vaadi indeksin lemmausta, vaativat ne hakusanojen lemmauksen. Tehokkuuden kannalta indeksin lemmaaminen on kuitenkin moninkertainen vaatimus hakusanojen lemmaukseen verrattuna. Lisäksi eri menetelmien laskennallista tehokkuutta vertailtaessa on huomioitava, että lemmaukseen perustuvassa tiedonhaussakin hakusanat on lemmattava.

SWERG-menetelmä muodostaa yksittäiselle hakusanalle useita muotoja, joista osa on epäaitoja suomen kielen sanamuotoja. Muotojen generointi perustuu 261087 sanan nominatiivi-, genetiivi-, partitiivi-, inessiivi-, elatiivi- ja illatiivimuotoihin yksikössä sekä monikossa sisältävään tietokantaan. SWERG+-menetelmässä tietokantaa on laajennettu tuottamalla sanan translatiivi-, adessiivi-, ablatiivi-, allatiivi-, essiivi- ja abessiivimuodot yksikössä ja monikossa korvaamalla elatiivin ssa-pääte kyseisen taivutusmuodon päätteellä. Menetelmä generoi yksittäiselle perusmuodolle lukuisia

(oikeita ja vääriä) taivutusmuotoja yksikkö- ja monikkomuotoihin perustuvan vertailun pohjalta. Vertailun tuottamia sääntöjä sovelletaan suomen kielen perusmuotoisten sanojen 3052 erilaiseen loppuosaan. SWERG-menetelmässä yksittäisiä sääntöjä syntyy 21395 ja SWERG+-menetelmässä 47966. (Kettunen & Arvola 2012.)

Esimerkiksi sana ”sana” tuottaa SWERG+-menetelmälla seuraavat merkkijonot: sana, sanoiden, sanoita, sanoissa, sanoina, sanoiksi, sanoilla, sanoilta, sanoille, sanoitta, sanoista, sanoihin, sanojen, sanoja, sanat, sanan, sanaa, sanassa, sanana, sanaksi, sanalla, sanalta, sanalle, sanatta, sanasta, sanaan, sant, sann, sanl. Näistä 29 muodosta 4 eivät siis oikeasti ole suomen kielen sanamuotoja, mutta ne eivät heikennä tiedonhaun tuloksellisuutta.

Snowball-stemmaus perustuu Porterin (1980) kehittämään stemmausalgoritmiin. Snowball-menetelmä on varsin tehokas ja tuloksellinen apuväline morfologian hallintaan (Hollink et al. 2004; Kettunen & Airio 2006). Snowball-menetelmästä menestyneen tekee sen yksinkertaisuus ja sanakirjariippumattomuus, vaikkakin sen kanssa voidaan hyödyntää myös sanakirjaa.

Snowball-stemmaus perustuu joukkoon sääntöjä, jotka sanan loppuosan perusteella katkaisevat sanan vartalomuotoon. Porterin kehittämä menetelmä tunnistaa ja käsittelee suomen kielen nomineista seuraavat päätteet:

1. Liitepartikkelit (esim. -kö, -kaan, -kin)
2. Possessiivit (esim. -ni, -si, -nsä)
3. Suffiksit (esim. -siin, -ssa, -llä)
4. Muita päätteitä (esim. -mpi, -mmi, -ejä)

Merkkijonot käsitellään yllä mainitussa järjestyksessä. Jokaisessa kohdassa käsiteltäväksi valitaan eniten merkkejä sisältävä pääte. Tällöin esimerkiksi mpi- ja impi-päätteistä käsiteltäväksi valikoituisi impi-pääte.

FINTWOL-menetelmä on osa kaupallista TWOL-tuoteperhettä, joka sisältää sanakirjamuotoistamismenetelmät yhdeksälle eri kielelle. TWOL-menetelmien on tutkimuksissa todettu olevan tehokkain menetelmä suomen kielen kaltaisten vahvasti taipuvien kielten morfologian hallintaan (Kettunen & Airio 2006; Kettunen & Arvola

2012). TWOL-menetelmien ongelmana ovat lemmaukselle tyypilliset suuret suorituskyvylliset vaatimukset, jotka heikentävät indeksoinnin tehokkuutta ja vaikeuttavat sen päivittämistä. Lemmaukseen perustuvat menetelmät ovat lisäksi riippuvaisia sanakirjasta (Kettunen 2007).

3.3 Tiedonhakujärjestelmien evaluointi

Lancaster (1971) jakaa tiedonhakujärjestelmien evaluoinnin kolmeen eri tarkastelunäkökulmaan:

1. Kuinka hyvin järjestelmä täyttää sille asetut tavoitteet?
2. Kuinka tehokkaasti järjestelmä täyttää sille asetetut tavoitteet?
3. Oikeuttaako järjestelmä olemassaolonsa?

Ensimmäistä näkökulmaa kutsutaan tiedonhaun *tuloksellisuuden* (effectiveness) mittaamiseksi. Perinteisesti tuloksellisuutta on mitattu selvittämällä kuinka hyvin järjestelmä tarjoaa relevanttia tietoa (suhteessa epärelevanttiin tietoon). Toinen näkökulma tarkoittaa järjestelmän *tehokkuutta* (efficiency), jolla viitataan järjestelmän tekniseen suorituskykyyn ja sen taloudellisuuteen. Kolmas näkökulma selvittää tuloksellisuuden ja tehokkuuden suhdetta.

Tiedonhaun evaluointitutkimuksen suosituin mittari on tuloksellisuus, ja merkittävä osa tutkimuksesta keskittyy evaluoimaan järjestelmiä mittaamalla tulosdokumenttien relevanssia. Kuitenkin myös tehokkuus on ollut tutkimuksessa esillä erityisesti aiemmin kun datan säilyttäminen ja käsittely ovat olleet kallista nykyään jopa varsin vaatimattoman kokoisiksi koetuissa testikokoelmissa. Lisäksi web-tiedonhaun yleistymisen ja informaatiotulvan myötä on myös tutkimuksessa kiinnitetty enemmän huomiota järjestelmien tehokkuuteen.

Tässä alaluvussa esitellään ensin tiedonhaun tuloksellisuuden evaluointia ja sen perustana toimivaa tiedonhaun laboratoriomallia, jota soveltaen tämäkin tutkimus on tehty. Toiseksi perehdytään tekstikokoelman rakentamisen teoriaan ja ohjeistuksiin, joita hyödyntäen tutkimuksessa rakennettu testikokoelma on suunniteltu ja toteutettu. Lopuksi syvennytään evaluoinnissa käytettävien tulosten mittaamiseen ja esitellään mittausmenetelmien toimintaperiaatteita.

3.3.1 Tiedonhaun tuloksellisuuteen perustuva evaluointi ja laboratoriomalli

Tiedonhaun tuloksellisuuden mittaamiseen yleisesti käytetty keino on tiedonhaun tulosdokumenttien relevanssin selvittäminen. Relevanssia lienee yksinkertaisinta tarkastella joko käyttäjärelevanssina tai aiherelevanssina. Käyttäjärelevanssi ilmaisee, kuinka relevantteja tulokset ovat tiedonhakijalle yksilönä. Aiherelevanssi sen sijaan määrittää tulosten relevanssin yleisellä tasolla tiettyyn hakuaiheeseen liittyen.

Tutkimuksessa käytettävän menetelmän avaamiseksi tarkastellaan Saracevicin (1996) viiteen eri näkökulmaan jaettua relevanssia:

1. Järjestelmä eli algoritminen relevanssi, joka kuvaa relevanssia tietyn matemaattisen kaavan perusteella. Tällainen on esimerkiksi tutkimusasetelmassa luvussa 4.3 esitetty vektorimallin mukainen täsmäytys, jossa relevanssia kuvaa kaavalla saatu numeerinen arvo.

2. Aiherelevanssi eli objektiivinen aiheeseen liittyvä relevanssi.

3. Kognitiivinen relevanssi kuvaa tiedon jäsentymistä käyttäjän sen hetkisessä kognitiivisessa tilassa.

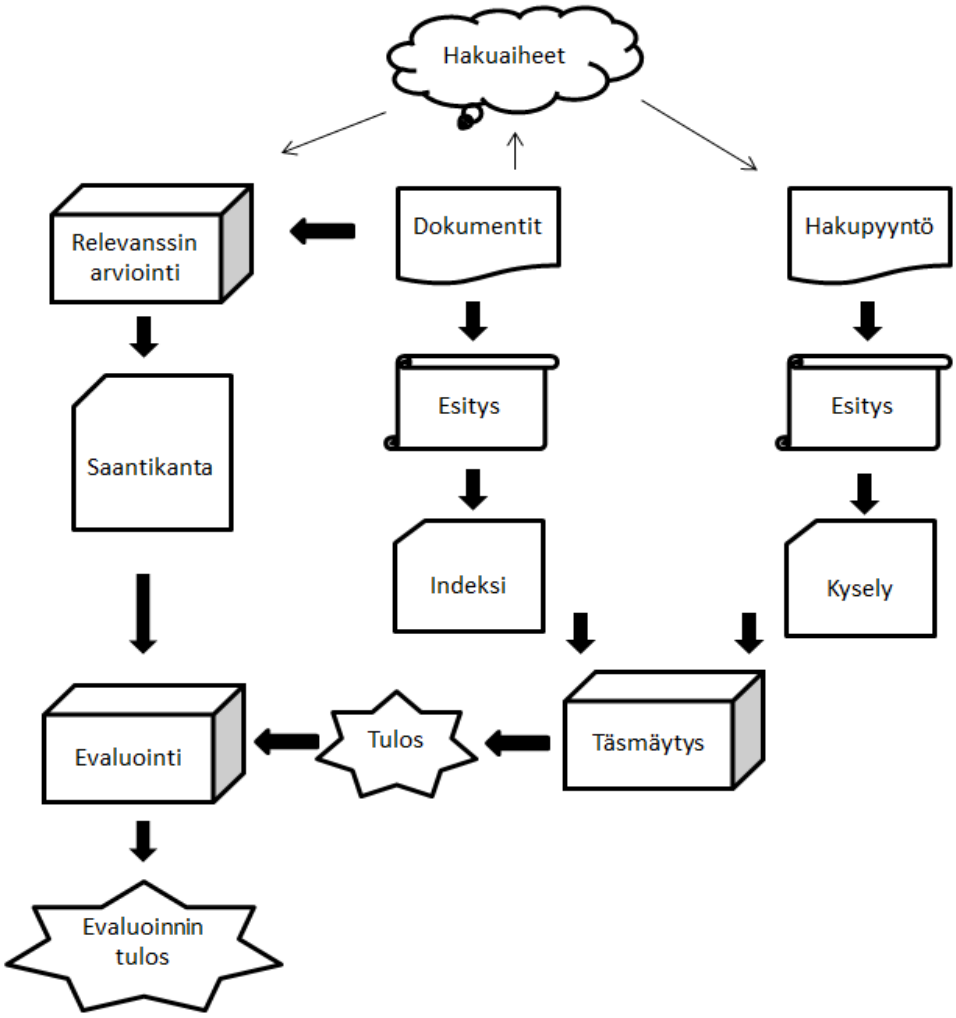
4. Tilannerelevanssi tarkoittaa dokumentin relevanssia tai hyödyllisyyttä tietyssä tilanteessa.

5. Motivaatio- eli affektiivinen relevanssi tarkoittaa dokumentin relevanssia suhteessa käyttäjän motivaatioon tai tarkoitusperiin.

Kolme viimeistä näkökulmaa ovat vahvasti sidoksissa käyttäjään. Erityisesti käyttäjään liittyvän relevanssiin liittyen erilaisia teoreettisia näkökulmia on esitetty paljon. Mizzaro (1997) on koonnut artikkeliinsa kattavasti erilaisia relevanssin näkökulmiin liittyviä teorioita. Tässä tutkimuksessa kuitenkin keskitytään erityisesti kahteen ensimmäiseen relevanssin näkökulmaan, jotka sisältyvät vahvasti laboratoriomallin mukaiseen tiedonhakumenetelmien evaluointiin.

Laboratoriomalli on yleisesti tiedonhakujärjestelmien evaluoinnissa käytetty tutkimusasetelma ja sen avulla voidaan kuvata tiedonhakujärjestelmän kehittämiseen vaikuttavat tekijät ja niiden väliset suhteet. Laboratoriomallin tavoitteena on selvittää kyselyesityksen, dokumenttiesityksen ja täsmäytyksen aikaansaama tuloksellisuus

aihekohtaisesti (Ingversen & Järvelin 2005, 172). Yleensä varsinaisena tutkimuksen kohteena eli muuttujana on yksi näistä kolmesta tekijästä, kun taas kaksi muuta tekijää ovat vakioita. Tällöin voidaan verrata luotettavasti laboratorio-olosuhteissa esimerkiksi kolmen eri indeksointimenetelmän (ts. dokumenttiesityksen) tuloksellista paremmuutta tietyn täsmäytysmenetelmän ja kyselyesityksen kanssa.



Kuva 5,tiedonhaun laboratoriomalli (Ingversen & Järvelin 2005, 5) sovellettuna

Jotta tiedonhakujärjestelmän jotain osaa voidaan evaluoida, tarvitaan jokin aineisto jolla tutkimus suoritetaan. Perinteisesti tällainen aineisto on ollut artikkelikokoelma, joka pysyy vakiona. Varsinaisten kokoelman dokumenttien lisäksi tarvitaan hakutehtävät, jotka suoritetaan ja jotka tyypillisesti pysyvät vakiona jokaiselle tutkimuksessa vertailtavalle menetelmälle. Lopuksi hakutehtävien tulokseksi saatujen dokumenttien relevanssi pitää arvioida ja arviointia varten päättää tapa jolla relevanssia mitataan. Näitä osa-alueita selitetään tarkemmin myöhemmin tässä luvussa.

Laboratoriomallin vahvuutena on, että sen perusteella tehtävä merkkijonojen täsmäytyksen voidaan todeta korreloivan aiherelevanssin kanssa (Kekäläinen & Järvelin 2002). Relevanssin mittaamisen, skaalautuvuuden ja vakioiden sekä muuttujan asettamisen ansiosta laboratoriomalli muodostaa erinomaisen työkalun tiedonhakumenetelmien evaluointiin. Tämän takia se on saanut aseman standardina tapana evaluoida tiedonhakumenetelmiä.

Vaikka laboratoriomallin mukainen evaluointitutkimus on ollut johtava tapa evaluoida tiedonhakumenetelmiä, on se saanut osakseen myös kritiikkiä. Keskustalo (2010, 18) on koonnut ja jäsentänyt perinteiseen laboratoriomallin mukaiseen tutkimuksen kohdistuvan kritiikin kuuteen eri ryhmään, joissa huomioidaan kattavasti laboratoriomallin tutkimuksellisia rajoitteita, jotka ovat

1. käyttäjämallinnuksen puutteellisuus,
2. vain yksi kysely per hakuaihe,
3. tarkasti määritetyt kyselyt ja hakuaiheet,
4. aiheperusteinen relevanssi,
5. tulosdokumenttien käsittely itsenäisinä yksikköinä (eikä tulosjoukkona),
6. arvioinnin mittaaminen (perustuu yleensä vain saantiin ja tarkkuuteen).

Tähän listaan voisi lisätä myös sen, että evaluointi on sidoksissa testikokoelmaan. Kun evaluoinnin tuloksia on tarkoitus hyödyntää johonkin käytössä olevaan järjestelmään, täytyy testikokoelman dokumenttien yhdessä hakuaiheiden ja kyselyjen sekä relevanssien kanssa olla vertailukelpoisia käytännön järjestelmään (Sanderson & Braschler 2009).

Laboratoriomalliin kohdistuvalle kritiikille yhteistä on, että sen mukainen evaluointi ei välttämättä vastaa tosi elämässä tapahtuvaa tiedonhakua. Tämän puutteen paikkaamiseksi Ingversen ja Järvelin (2005, 2) ovat esittäneet, että tiedonhaun ja tiedonhankinnan tutkimusten pitäisi tehdä enemmän yhteistyötä. Uudempia tutkimussuuntauksia kuten vuorovaikutteinen tiedonhaku (kts. Borlund 2000) ja tehtäväperusteinen tiedonhaku (kts. Vakkari 2001) onkin kehittynyt perinteisen tiedonhaku- ja tiedonhankintatutkimuksen rinnalle. Näillä tutkimuksen aloilla

Saracevicin (1996) yllä listatuista relevanssin näkökulmista kolme viimeistä pystytään huomioimaan.

Vaikka laboratoriomalli on saanut osakseen kritiikkiä, on se hyvä menetelmä osana melkein kaikkea tiedonhaun tutkimusta yhtenä relevanssin mittaamisen osana. Vaikka se ei vastaakaan kaikkiin relevanssin näkökulmien vaatimuksiin, vastaa se kuitenkin osaan niistä ja joita ei oikein muuten voida selvittää. Keskustalo (2010, 24–) esittelee miten muita relevanssin olemuksia voidaan ottaa huomioon laboratoriomallia hyödyntävissä tutkimuksissa.

3.3.2 Testikokoelmat ja niiden rakentaminen

Evaluointitutkimuksessa on perinteisesti käytetty artikkelitietokantoja, jotka koostuvat jonkin tieteenalan tai ihan vain sanomalehtien artikkeleista tai esimerkiksi oikeudellisista dokumenteista (Chowdhury 2010, 295-). Erityisesti 1990-luvulla tutkimuksen kasvaessa suosituimpana aineistona olivat sanomalehtien artikkelit (Ingversen & Järvelin 2005, 123). Näiden lisäksi on jonkin verran käytetty esimerkiksi historiallista sanomalehtiaineistoa, jonka OCR-tekniikalla suoritettu digitalisointi ja vanhat kirjoituskäytännöt tekevät sanoista kirjoitusasultaan erilaisia kuin ne ovat nykypäivinä. Lisäksi nykyään verkkosivut ovat yleistyneet tiedonhaun tutkimusaineistona (Voorhees & Harman 2005, 8).

Testikokoelmaa rakentaessa kannattaa valita dokumenttikokoelma, joka edustaa mahdollisimman hyvin kokoelmaa jossa testattavien menetelmien halutaan toimivan tuloksellisesti hyvin (Sanderson & Braschler 2009). Myös esimerkiksi valittavien dokumenttien määrällä ja ominaisuuksilla on vaikutuksensa evaluointiin. Clough ja Sanderson (2013) ovat listanneet testikokoelman dokumenttien valinnan keskeisiksi päätöksiksi:

- Kuinka monta dokumenttia pitäisi kerätä?
- Miten dokumentteja pitäisi valita edustavan kokoelman saamiseksi?
- Onko kokoelmalla tekijänoikeudellisia rajoituksia?

Kerättävien dokumenttien määrää päätettäessä Clough ja Sanderson (2013) esittävät, että pienemmissä kokoelmissa kaikki dokumentit kannattaa ottaa mukaan

testikokoelmaan. Kuitenkin käytössä olevissa kokoelmissa kokoelma on harvoin stabiili. Erityisesti internetissä dokumentit voivat muuttua tai ne voivat siirtyä (Chowdhury 2010, 382) (tai kadota kokonaan). Näiden syiden vuoksi dokumenteista täytyy yleensä ottaa tietyn hetken otos. Isoissa kokoelmissa dokumenttien määrää täytyy usein rajata testiasetelman suorituskyvyn pitämiseksi hyvänä.

Dokumenttien edustavuudesta Clough ja Sanderson (2013) esittävät, että dokumentit pitäisi valita joko satunnaisotannalla tai valitsemalla tietty alajoukko. Tämä kannattanee tulkita siten, että dokumentteja valitessa ei sovi valita esimerkiksi laadukkaimmaksi oletettuja dokumentteja, ellei siihen ole erityinen tarve. Esimerkiksi oletettuun dokumenttien laatuun perustuva valinta ei luultavasti tuottaisi tuloksia, joiden perusteella voisi valita menetelmiä jotka tuottaisivat mahdollisimman hyviä tuloksia sovellettuna järjestelmään käytännössä.

Varsinaisen dokumenttikokoelman määrittämisen jälkeen on hakumenetelmien vertailemiseksi kehitettävä hakuaiheet. Clough ja Sanderson (2013) ovat listanneet hakuaiheiden kehittämiseen työkaluiksi seuraavat kohdat:

- Kuinka hakuaiheet pitäisi määrittää?
- Kuinka monta hakuaihetta pitäisi olla?
- Edustavatko hakuaiheet tarpeeksi kattavasti mahdollisia tiedontarpeita?
- Kuinka hakuaiheet pitäisi ilmaista?

Hakuaiheiden määrittämisessä olisi Cloughin ja Sandersonin (2013) mukaan hyvä käyttää mahdollisimman paljon realismia. Realismin huomioimista voidaan parantaa esimerkiksi lokitiedostojen, loppukäyttäjien haastattelun tai kokoelman asiantuntijoiden avulla. Erityisesti lokianalyysistä on hyötyä myös kattavuuden toteamisessa.

Sandersonin ja Braschlerin (2009) mukaan tieteenalalla vallitsee konsensus, että evaluointitutkimuksessa noin 50 hakuaihetta olisi riittävä otos. Tutkimusten mukaan hakuaiheiden määrällä on positiivinen vaikutus evaluoinnin luotettavuuteen (Carterette et al. 2008; Sanderson & Zobel 2005).

Hakuaiheista muotoillaan yleensä otsikko, jota usein käytetään baseline-kyselynä testiajoissa. Kyselyn lisäksi on tarpeellista erikseen kuvailla tarkemmin millainen

hakuaiheen tiedontarve on ja millaisia tuloksia mahdollisesti halutaan. Kuvailuja on käytetty tutkimuksessa myös hakukoneelle annettavina hakusanoina, mutta käyttäjätutkimuksen mukaan käyttäjät suosivat lyhyitä kyselyjä (Jansen et al. 2000). Lisäksi varsinaisten käyttäjien kyselyt voivat olla monitulkintaisia, eivätkä ne ole välttämättä optimaalisia hyvien tulosten löytämiseksi. Kokeilemalla kannattaa varmistaa, että kyselyillä löytyy relevantteja tuloksia hakuaiheille. Hakuaihe ilman relevantteja kokoelman dokumentteja ei auta evaluointia.

Yleensä tekstitietokannat ovat niin suuria, että niiden jokaisen dokumentin relevanssia jokaiselle tutkimuksessa käytetylle hakuaiheelle ei voida arvioida. Tämä olisi työmäärän osalta kohtuuttoman tai esimerkiksi internetin ollessa kyseessä jopa mahdottoman suuri työ. Työmäärän kohtuullistamiseksi jokaiselle hakuaiheelle on määritetty oma relevanttien dokumenttien joukko hyväksi todettujen hakumenetelmien tulosten perusteella. Tätä tapaa määrittää kullekin hakuaiheelle omat relevanttien dokumenttien joukot eli relevanssikorpus kutsutaan *pooling-menetelmäksi*. Clough ja Sanderson (2013) ovat listanneet relevanssiarvioiden tekemiseen seuraavat avainkysymykset:

- Kenen pitäisi tehdä relevanssiarviot?
- Kuinka monta tulosdokumenttia pitäisi arvioida?
- Mitä arvioijien kuuluu tehdä?
- Miten paikallistaa löytämättömät relevantit dokumentit?

Bailey ja kumppanit (Bailey et al. 2008) selvittivät, kuinka paljon arvioijan tietämyksen tasolla oli vaikutusta relevanssiarvioon. Heidän mukaansa aihetta laatimassa olleet asiantuntijat ja laatimiseen osallistumatta olleet asiantuntijat tekivät kohtalaisen samanlaisia relevanssiarvioita. Sen sijaan aihetta tuntemattomien arvioijien arviot erosivat selkeästi alkuperäisten aiheita laatineiden asiantuntijoiden arvioista ja he arvioivat aiheita laatineita asiantuntijoita useammin dokumentit relevanteiksi. Myös Kinney ja kumppanit (Kinney et al. 2008) totesivat asiantuntijoiden ja ei-asiantuntijoiden relevanssiarvioiden eroavan selvästi toisistaan, joskin ei-asiantuntijoiden arviot olivat tarkempia heidän saadessa laadukkaammat ohjeet arviointia varten.

Esimerkiksi TREC-kokoelmissa on perinteisesti otettu pooling-menetelmään 100 ensimmäistä tulosdokumenttia per hakumenetelmä (Voorhees 1998). Sittemmin Sanderson ja Zobel (2005) sekä Carterette ja kumppanit (Carterette et al. 2008) ovat havainneet, että arvioitavien tulosdokumenttien määrän sijaan olisi hyödyllisempää panostaa hakuaiheiden määrään. Tämä huomio tukee myös vuorovaikutteisen tiedonhakututkimuksen huomiota, jonka mukaan käyttäjän huomio kiinnittyy yleensä vain tuloslistan kärjessä oleviin dokumentteihin (Costa & Silva 2011; Jansen et al. 2000). Lisäksi Zobel (1998) vertasi tutkimuksessaan, miten pooling-menetelmällä kerättävien tulosdokumenttien rajaaminen 10 ensimmäiseen dokumenttiin vaikuttaa evaluoinnin tuloksiin. Hänen tutkimuksensa mukaan rajaaminen ei juurikaan vaikuttanut vertailtavien menetelmien suhteelliseen paremmuuteen, vaikka tarkkuuslukemat hieman muuttuivatkin.

Relevanssin arvioijien työ on yksinkertainen. Heidän tehtävänään on arvioida relevanssi annettujen ohjeiden mukaan. Tämän takia on syytä kiinnittää huomiota opastuksen laatuun. Opastuksen laadulla onkin selvä vaikutus siihen, kuinka laadukkaasti ei-asiantuntija-arvioijat pystyvät arvioimaan relevanssia (Kinney et al. 2008).

Testikokoelman relevanssiarvioita pooling-menetelmällä rakentaessa olisi hyvä ottaa huomioon mahdollisimman kattavasti erityyppisiä menetelmiä joita testiaineistoon mahdollisesti käytettäisiin (Clough ja Sanderson 2013). Kuitenkin mikäli jokin uusi menetelmä tuottaa uusia tulosdokumentteja, olisi kannattavinta arvioida niille relevanssi. Todennäköisyyttä uusien arvioiden tarvitsemiselle voi laskea valitsemalla evaluointiin mittareita, jotka ottavat huomioon vain tuloslistan kärkipään dokumentit.

Aiemmin tässä luvussa esitettyjä kysymyslistoja sopii täydentää sisäisen reliabiliteetin tarkastelulla (kts. Heikkilä 2008, 187). Arviointia voidaan pitää arvioijan tietotasosta riippumatta luotettavampana, mikäli arvioija arvioidessaan uudelleen tietyn dokumentin relevanssia arvioi sen samaksi kuin aiemmin. Lisäksi on olemassa menetelmiä, joilla voidaan arvioida relevanssia ilman arvioijana toimivaa ihmistä. Tällä tavalla arvioitua relevanssia kutsutaan *pseudorelevanssiksi*. Pseudorelevanssia ensimmäisenä tutkineet Soboroff ja kumppanit (Soboroff et al. 2001) havaitsivat pseudorelevanssiin perustuvien tulosten olevan hämmästyttävän lähellä perinteisten relevanssiarvioiden perusteella saatuja tuloksia.

Efron ja Winget (2010) kehittivät menetelmän, joka perustui useaan kyselynäkökulmaan eli käytännössä useampaan kyselyyn per hakuaihe. Näkökulmien tekeminen vaati ihmistyötä, mutta huomattavan paljon vähemmän kuin perinteisten relevanssiarvioiden tekeminen. He havaitsivat, että vain tuloslistan kärkipään dokumentit huomioivat evaluointimenetelmät saivat hyvin samankaltaisia tuloksia pseudorelevanssin ja perinteisiin relevanssiarvioihin perustuvilla mittauksilla.

3.3.3 Tuloksellisuuden mittaaminen

Tiedonhaun tuloksellisuutta on heti tutkimuksen alkuhetkistä saakka 1960-luvulla mitattu *saannin* ja *tarkkuuden* mittareilla. Saanti ilmaisee kuinka suuri osa tulokseksi halutuista dokumenteista kuuluu saatuun tulosjoukkoon. Tarkkuus taas ilmaisee kuinka suuri osa tulosjoukon dokumenteista on haluttuja tuloksia. (Chowdhury 2010, 286.) Saannin ja tarkkuuden on todettu korreloivan keskenään käänteisesti (Cleverdon & Keen 1966). Saannista ja tarkkuudesta saatava saanti-tarkkuus-käyrä on edelleenkin suosittu ja yksinkertainen keino tuloksellisuuden vertaamiseksi.

Informaatiomäärän kasvaessa binäärinen tarkkuuden määrittäminen voidaan kyseenalaistaa riittämättömänä, koska enemmän relevantit tulosdokumentit hukkuvat binäärisessä relevanssimäärittelyssä marginaalisesti relevanttien dokumenttien joukkoon (Sormunen et al. 2001). Tuloksellisuuden mittaamisessa voidaankin useissa tapauksissa pitää mielenkiintoisena painottaa vahvemmin relevantteja tulosdokumentteja. Tällaisen painotuksen huomioiva kumuloituvan hyödyn menetelmä (kts. Järvelin & Kekäläinen 2002) on vakiintunut suosittuna tuloksellisuuden mittarina. Kumuloituvaa hyötyä mittaava menetelmä ottaa myös huomioon käyttäjätutkimuksissa havaitun seikan, jonka mukaan enemmistö käyttäjistä tarkastelee vain ensimmäisen tulossivun tai listan kärjessä olevia dokumentteja (Costa & Silva 2011; Jansen et al. 2000). Toisin sanoen relevantti tulosdokumentti on sitä arvokkaampi, mitä korkeammalla se on tuloslistalla.

Perinteisessä tiedonhaun evaluointitutkimuksessa tuloksellisuuden mittaamisessa on voitu ottaa huomioon todella suuri määrä tulosdokumentteja (Keskustalo 2010, 11). Esimerkiksi kansainvälisen tiedonhakukonferenssin TREC:n evaluointitulokset perustuvat 1000 ensimmäisen tulosdokumentin huomioimiseen (Voorhees & Harman 2005, 58). Näin usean tulosdokumentin huomioiminen perustuu yleensä etukäteen mittavalla työllä laadittuun relevanssikorpukseen.

Käyttäjätutkimusten mukaan yleensä käyttäjä kuitenkin tarkastelee vain ensimmäisen tulossivun ensimmäisiä dokumentteja (Costa & Silva 2011; Jansen et al. 2000). Lisäksi tarkasteltavien tulosten määrän sijaan hakuaiheiden määrän kasvattamisella saadaan evaluointia parannettua tehokkaammin (Sanderson & Zobel 2005; Cartetette et al. 2008). Näiden huomioiden perusteella evaluoinnin mittareina kannattaisi suosia erityisesti ensimmäisiin dokumentteihin keskittyviä mittareita evaluoinnin tehokkuuden ja luotettavuuden parantamiseksi.

Tiedonhakututkimuksen yleisesti käytettävissä olevat mittarit ovat hyvin skaalautuvia, minkä vuoksi niiden soveltaminen onnistuu vaikka tarkasteltavien tulosdokumenttien määrä muuttuisi. Tutkimusasetelmassa luvussa 4.6 esitellään tässä tutkimuksessa käytetyt kumuloituvan hyödyn (CG[10]) ja tarkkuuden (P10) laskeminen kymmenelle ensimmäiselle tulosdokumentille.

3.4 Aikaisempi tutkimus

Tämän aliluvun ensimmäisessä osiossa perehdytään verkkoarkistoihin kohdistuvaan tiedonhaun tutkimukseen, joka toimi pohjustuksena tutkimuksen testikokoelmaa rakennettaessa. Toisessa osiossa käsitellään luonnollisen kielen hallintaan tarkoitettujen sovellutusten tuloksellisuutta tiedonhaussa mittaavaa tutkimusta.

3.4.1 Verkkoarkistoihin liittyvä tutkimus

Verkkoarkistojen tiedonhakua koskeva tutkimus käynnistyi vasta tämän vuosituhannen ensimmäisen vuosikymmenen lopulla. Ensimmäisessä tutkimuksessa kartoitettiin Portugalin kansalliskirjaston ylläpitämän verkkoarkiston käyttäjien tiedontarpeita (Costa & Silva 2010). Niitä selvitettiin lokianalyysin, verkkosivuilla täytettävän lomakkeen sekä käyttäjäkokeiden avulla. Toinenkin verkkoarkistojen tiedonhakua koskeva tutkimus tehtiin tutkimalla lokianalyysin avulla samaisen portugalilaisen verkkoarkiston käyttäjien tiedonhakukäyttäytymistä (Costa & Silva 2011). Näiden tutkimusten lisäksi on ilmestynyt joitain kirjastojen omia käyttäjätutkimuksia sekä verkkoarkistojen kehittämisen tueksi laadittuja käyttöskenaarioita (IIPC 2006). Costa ja Silva (2012) ovat lisäksi julkaisseet artikkelin saman portugalilaisen verkkoarkiston evaluoinnista.

Tiedontarpeiden luokitteluun Costa ja Silva (2010) käyttivät Rosen ja Levinsonin (2004) uudelleen määrittelemää Broderin (2002) jakoa. Sen mukaan web-tiedonhaussa tiedontarpeet voidaan jakaa

1. navigationaaliseen tiedontarpeeseen, jossa pyritään hakeutumaan tiedossa olevalle sivustolle.

2. informationaaliseen tiedontarpeeseen, jossa haetaan tietoa jostain aiheesta.

3. transaktionaaliseen tiedontarpeeseen, jossa tarkoituksena on suorittaa jokin tietty tehtävä.

Costan ja Silvan (2010) ensimmäinen tutkimus päätyi tulokseen, jonka mukaan verkkoarkistosta tietoa hakevan tiedontarve on yleensä (63,5 %) navigaationaalinen, mutta myös informationaalinen tiedontarve oli sangen yleinen (27,9 %). Sen sijaan transaktionaalinen tiedontarve oli jokseenkin harvinaisempi (8,6 %). Tämä jakauma eroaa huomattavasti web-tiedonhaun tiedontarpeiden jakaumasta. Web-tiedonhaun tarpeita selvittäneen lokianalyysitutkimuksen mukaan web-tiedonhaussa yleisin tiedontarve on informationaalinen. Web-tiedonhaussa myös transaktionaalinen tiedontarve on navigaationaalista tiedontarvetta yleisemmin esiintyvä. (Jansen et al. 2008.)

Costan ja Silvan (2010) tutkimuksessa selvitettiin myös verkkoarkistoon kohdistuneiden hakujen aiheita. Navigaationaaliset haut kohdistuivat yleisimmin sosiaalisiin (28,3 %), tietoteknisiin (14,5 %) tai opetustarkoituksellisiin (14,5 %) aiheisiin. Informaationaaliset haut taas koskivat pääasiassa henkilöitä (37,5 %), terveyttä (15,8 %) sekä viihdettä (10,6 %).

Costan ja Silvan (2011) **hakukäyttäytymistä** selvittävän tutkimuksen mukaan verkkoarkistoon tehdyistä istunnoista 65 % sisälsi vain yhden tehdyn kyselyn, kun taas web-tiedonhaussa yhden kyselyn pituiset istunnot ovat hieman harvinaisempia. Verkkoarkistosta saaduista hakutuloksista tarkasteltiin keskimäärin 1,4 tulokseksi saatua sivua, mikä on hieman vähemmän kuin web-tiedonhaun tulosten tarkastelussa. Merkittävin ero hakukäyttäytymisessä on kuitenkin hakuoperaattoreiden käyttämisessä. Verkkoarkistoon kohdistuneista hauista peräti 26 % sisälsi hakuoperaattorin käyttöä. Hakuoperaattoreiden käyttö oli merkittävissä määrin eli 20,2 % kyselyistä fraasioperaattorin (”) käyttöä. Fraasioperaattorin käyttö Costan ja Silvan (2011)

tutkimuksessa olikin huomattavasti yleisempää kuin Jansenin ja kumppaneiden (Jansen et al. 2000) web-tiedonhakukäyttäytymisen tutkimuksessa.

Erittäin mielenkiintoista verkkoarkistojen hakutulosten tarkastelussa oli se, että varhaisimpia verkkosivuja tarkasteltiin tuloksista huomattavasti eniten. Peräti 50 % kaikista klikatuista tulokseksi tulleista sivuista oli arkistoitu ensimmäisenä arkistointivuotena (Costa & Silva 2011). Tämän tuloksen perusteella voidaan ajatella, että tulosdokumentin ikä voi hyvinkin olla käyttäjälle tärkeä tekijä dokumentin relevanssia arvioidessa. Kokotekstikyselyissä tulokset olikin rajattu päättymisvuoden perusteella 24 % kyselyistä ja URL (Uniform Resource Locator, www-osoite) -kyselyissä 30 % kyselyistä, mutta aloitusvuotta ei ollut katsottu tarpeelliseksi rajata kuin 1 % kaikista kyselyistä.

3.4.2 Luonnollisen kielen hallinnan sovellusten tutkimus

Modernin suomen kielen käsittelymenetelmien tiedonhaun evaluointitutkimuksen perustana on Alkulan väitöskirjaan (2000) johtanut FULLTEXT-projekti. Alkula vertasi useiden suomen kielen käsittelyratkaisujen vaikutusta tiedonhaun tuloksellisuuteen sekä tehokkuuteen. Sittemmin erityisesti morfologian hallintaan keskittyviä tutkimuksia on tehty useita ja uusia menetelmiä on kehitetty. Seuraavaksi kuvataan lyhyesti yksittäisinä kappaleina edustavasti suomen kielen käsittelymenetelmiä evaluoivia tutkimuksia, joiden tulokset on koottu taulukkoon 1.

Kettunen (2004) vertasi FINTWOL-lemmausta, Snowball- ja MaxStemma-stemmauksia sekä käsittelemätöntä tekstiä TUTK-testikokoelmassa, jossa oli 30 hakuaihetta. Kettusen tutkimusarkkitehtuuri oli InQuery-hakujärjestelmä, joka on kehitetty Massachusettsin yliopistossa ja jonka täsmäytys perustuu klassiseen todennäköisyyslaskentaan (Callan et al. 1992). Kyselyt olivat pitkiä, eli ne sisälsivät hakuaiheen otsikkokentän ja kuvauskentän (topic, description). Tulosten mittarina toimi keskitarkkuus (AP).

Tomlinson (2004) evaluoi käsittelymenetelmiä tutkimusarkkitehtuurinaan Hummingbird SearcServer, joka perustuu alun perin Fulcrum Technologies -yrityksen kehittämään SearchServer-kerneliin. Hummingbird tukee luonnollisen kielen käsittelyä usealla eri kielellä suomen kieli mukaan luettuna. Tutkimuksessaan hän vertasi Hummingbirdin stemmausta, Humminbirdin sanakirjaan perustuvaa stemmausta sekä käsittelemätöntä tekstiä. Tutkimuksen aineistona oli CLEF 2003 -kokoelma. Kyselyiden hakusanat olivat

hakuaiheiden otsikko- ja kuvailukentän sanat. Mittareina eri indeksointimenetelmien vertailuun käytettiin tarkkuutta useassa eri katkaisupisteessä sekä keskitarkkuutta (AP).

Hollink ja kumppanit (Hollink et al. 2004) vertasivat Snowball-stemmauksen ja käsittelemättömän indeksoinnin lisäksi sumeita merkkijonomenetelmiä CLEF 2002 -kokoelmassa. Sumeina menetelminä olivat 4-, 5-, ja 6-grammit. Tutkimusarkkitehtuurina heillä oli FlexIR-järjestelmä, joka on kehitetty Amsterdamin yliopistossa ja perustuu vektorimalliin (Monz & de Rijke 2002). Tomlinsonin vertailun tapaan heidän vertailussa hakusanoina olivat otsikko- ja kuvailukentän sanat, mutta vertailulukuna oli keskimääräinen keskitarkkuus (MAP).

Airio (2006) selvitti vertailututkimuksessaan käsittelemättömän, Snowball-stemmatun sekä FINTWOL-lemmatun indeksoinnin tuloseroja. Tutkimusarkkitehtuurina hän käytti InQuery-hakujärjestelmää. Tomlinsonin tapaan Airio käytti tutkimuksessaan CLEF 2003 -aineistoa ja vertailulukuna keskitarkkuutta, mutta ei P10-tarkkuuksia. Aiempien tutkimusten tapaan käytettiin hakusanoina otsikko- ja kuvailukentän sanoja.

Kettunen ja Airio (2006) laajensivat luonnollisen kielen käsittelymenetelmien vertailua generatiivisiin menetelmiin. Heidän tutkimuksessaan verrattiin aiemmissa vertailuissa käytettyjen käsittelemättömän, Snowball-stemmatun sekä FINTWOL-lemmauksen lisäksi generatiivisia FCG9 ja- FCG12-menetelmiä. Airion aiemman tutkimuksen tapaan tutkimuksessa käytettiin CLEF 2003 -aineistoa, jonka lisäksi he käyttivät myös TUTK-testikokoelmaa. Vertailuluku oli aiemmasta tutkimuksesta vaihtunut MAP-lukuun. Myös Kettusella ja Airiolla oli käytössään InQuery-hakujärjestelmä. Kettunen ja kumppanit (Kettunen et al. 2007) laajensivat tutkimusta vertaamalla lisäksi lyhyitä kyselyitä.

Kettunen ja Arvola (2012) vertasivat kehittämäänsä generatiivista SWERG(+)-menetelmää FCG-menetelmiin, FINTWOL-lemmaukseen, Snowball-stemmaukseen ja käsittelemättömään esitykseen. Hakusanojen muotojen generoinnissa he käyttivät erikseen FINTWOL-menetelmän tuottamaa ensimmäistä lemmaa (taulukossa 1 ”ensim.”) ja kaikkia lemmoja (taulukossa 1 ”kaikki”). Aineistona heillä oli CLEF 2003 -testikokoelma. Kyselyt tehtiin erikseen pitkinä (topic, description) ja lyhyinä (topic). Hakujärjestelmänä toimi Lemur, joka on kehitetty Massachusetts Amherst -yliopiston ja Carnegie Mellon -yliopiston yhteistyössä ja perustuu avoimeen lähdekoodiin. Lemur

tarjoaa täsmäytykseen useita vaihtoehtoja, joista Kettunen ja Arvola käyttivät todennäköisyysperiaatteeseen perustuvaa täsmäytysmenetelmää.

Taulukko 1, luonnollisen suomen kielen käsittelymenetelmiä vertailevia tutkimuksia

Tutkimus	Järjestelmä	Aineisto	Menetelmä	Kysely	Mittari	Tulos
Kettunen	InQuery	TUTK	Fintwol	TD	AP	35.0
2004			MaxStemma	TD	AP	33.5
			Snowball	TD	AP	27.7
			Plain	TD	AP	18.9
Tomlinson	Hummingbird	CLEF 2003	Stemmaus + sanakirja (Sea...)	TD	P10	35.3
2004			Stemmaus (SearchServer)	TD	P10	27.8
			Ei käsittelyä	TD	P10	23.8
			Stemmaus + sanakirja (Sea...)	TD	AP	0.553
			Stemmaus (SearchServer)	TD	AP	0.422
			Ei käsittelyä	TD	AP	0.301
Hollink	FlexIR	CLEF 2002	Stemmaus (Snowball + split)	TD	MAP	0.3633
2004			Ei käsittelyä (split)	TD	MAP	0.3020
			4-grammi	TD	MAP	0.3536
			5-grammi	TD	MAP	0.3762
			6-grammi	TD	MAP	0.3560
Airio	InQuery	CLEF 2003	Lemmaus (FINTWOL + split)	TD	AP	50.5
2006			Stemmaus (Snowball)	TD	AP	48.5
			Ei käsittelyä	TD	AP	31
Kettunen	InQuery	TUTK	Lemmaus (FINTWOL + split)	TD	MAP	37.8
Airio			Stemmaus (Snowball)	TD	MAP	29.8
2006			FCG_3	TD	MAP	26.4
			FCG_9	TD	MAP	32.4
			FCG_12	TD	MAP	32.7
			Ei käsittelyä	TD	MAP	19.6
Kettunen	InQuery	CLEF 2003	Lemmaus (FINTWOL + split)	TD	MAP	37.6
Airio			Stemmaus (Snowball)	TD	MAP	34.7
2006			FCG_3	TD	MAP	24.2
			FCG_9	TD	MAP	33.7
			FCG_12	TD	MAP	34.0
			Ei käsittelyä	TD	MAP	22.7
Kettunen	InQuery	CLEF 2003	Lemmaus (FINTWOL + split)	T	MAP	42.8
Airio			Stemmaus (Snowball)	T	MAP	41.3
Järvelin			FCG_9	T	MAP	37.9
2007			FCG_12	T	MAP	38.1
			Ei käsittelyä	T	MAP	22.6
Kettunen	Lemur	CLEF 2003	Lemmaus (FINTWOL, kaikki)	TD	MAP	0.4021
Arvola			Lemmaus (FINTWOL, ensim.)	TD	MAP	0.5145
2012			Stemmaus (Snowball)	TD	MAP	0.4218
			SWERG+, kaikki	TD	MAP	0.4886
			SWERG+, ensimmäinen	TD	MAP	0.4775
			FCG_12, kaikki	TD	MAP	0.4487
			FCG_12, ensimmäinen	TD	MAP	0.4328
			FCG_6, kaikki	TD	MAP	0.4138
			FCG_6, ensimmäinen	TD	MAP	0.4101
			Ei käsittelyä	TD	MAP	0.3158
			Lemmaus (FINTWOL, kaikki)	T	MAP	0.4525
			Lemmaus (FINTWOL, ensim.)	T	MAP	0.4500
			Stemmaus (Snowball)	T	MAP	0.3251
			SWERG+, kaikki	T	MAP	0.4301
			SWERG+, ensimmäinen	T	MAP	0.4227
			FCG_12, kaikki	T	MAP	0.3815
			FCG_12, ensimmäinen	T	MAP	0.3777
			FCG_6, kaikki	T	MAP	0.3625
			FCG_6, ensimmäinen	T	MAP	0.3561
			Ei käsittelyä	T	MAP	0.2620

Yhteenvetona tuloksista voi todeta, että

1. kaikki menetelmät parantavat tuloksellisuutta käsittelemättömään verrattuna.
2. FINTWOL-lemmaus on tuloksellisin menetelmä.
3. n-grammit ovat suunnilleen yhtä tuloksellisia kuin Snowball-stemmaus.
4. generoivat menetelmät ovat vertailukelpoisia Snowball-stemmauksen kanssa.
5. generoitavien menetelmien tuloksellisuus paranee, mitä enemmän muotoja menetelmä tuottaa. Tuotteliaimmat menetelmät SWERG ja SWERG+ ovat Snowball-stemmausta tuloksellisempia.

Tämä yhteenveto toimii vertailukohtana seuraavassa luvussa esitettäviin tämän tutkimuksen tuloksiin. Kohdat 1, 3, 4 ja 5 ovat verrattavissa tähän tutkimukseen. Kohdan 2 huomiota ei tässä tutkimuksessa verrata, koska se vaatisi perusmuotoindeksin käyttöä. Tässä aliluvussa esiteltyjen tutkimusten kaltaisia evaluointitutkimuksia ei ole aiemmin suoritettu suomenkielisellä verkkoarkistoaineistolla, johon mahdollisesti tulevaisuudessa kehittyy tiedontarpeita joiden täyttämiseksi tarvitaan kyseisessä kontekstissa tuloksellisia kokotekstitiedonhakumenetelmiä.

4 TUTKIMUSASETELMA JA TULOKSET

Tässä luvussa esitellään tiedonhaun laboratoriomallin mukainen tutkimusasetelma. Ensin esitellään tutkimuskysymys. Toiseksi esitellään kokoelma, eli Kansalliskirjaston Suomalainen verkkoarkisto. Kolmanneksi esitellään tutkimuksessa käytetty Apache Solr -hakujärjestelmä. Neljänneksi esitellään hakuaiheiden ja kyselyiden luominen sekä relevanssiarvioiden asteikko ja suorittaminen. Viidennessä aliluvussa perustellaan tutkimuksessa vertailtavat taivutusmuotoindeksiin sopivat generatiiviset menetelmät ja perustason menetelmä. Kuudentena esitellään tutkimuksessa käytetyt evaluoinnin mittarit esimerkkien kera sekä mittareiden arvojen tilastollista merkitsevyyttä selvittävät testit. Viimeiseksi esitellään tutkimuksen tulokset.

4.1 Tutkimuskysymys

Tässä tutkimuksessa mitataan tiedonhaun laboratoriomallin mukaan neljän eri kyselynkäsittelymenetelmän tuloksellisuutta Suomalaisesta verkkoarkistosta rakennetun testikokoelman taivutusmuotoindeksiä käyttäen. Tutkimuskysymys on: Miten kyselynkäsittelymenetelmien variointi vaikuttaa tiedonhaun tuloksellisuuteen?

4.2 Kokoelma

Tutkimuksen kokoelmana on Kansalliskirjaston keräämä ja ylläpitämä Suomalainen verkkoarkisto. Verkkoarkistoon on haravoitu suomalaisia verkkosivuja vuodesta 2006 alkaen. Suurin osa aineistosta on suomenkielistä ja ainakin hakuaiheiden kartoittamisessa ja tuloksissa olivat edustettuina vain suomenkieliset verkkosivut. Verkkoarkisto on käytettävissä tekijänoikeuslain rajauksen perusteella ainoastaan vapaakappalekirjastoissa, Eduskunnan kirjastossa sekä Kansallisessa audiovisuaalisessa instituutissa (KAVI).

Verkkoarkisto tiedonhaun evaluointitutkimuksen aineistona eroaa perinteisesti tutkimuksessa käytetyistä aineistoista. Yleensä aineistona on käytetty artikkelitietokantoja, joiden artikkeleilla on yleensä tunnistettavissa selkeä aihe. Verkkoarkistossa näin ei ole, vaan aineistossa on paljon esimerkiksi sanalistoja, keskustelua ja testisivuja. Tai sitten sivulla ei ole mitään selkeää tarkoitusta tai sen tarkoitus on hämätä internetin hakukoneita.

Tutkimuksessa kokoelma rajattiin käytännön syistä vuonna 2006 kerättyihin verkkosivuihin. Vuoden aikana kerättyjä sivustoja voidaan pitää sivujen laadun osalta edustavana otoksena. Lisäksi tutkimuksessa on osoitettu, että eniten selatut tulokset edustavat verkkoarkiston varhaisinta aineistoa eli tässä tapauksessa vuoden 2006 aikana kerättyä aineistoa. Indeksoituja sivuja oli yhteensä lähes 30 miljoonaa. Indeksointiprosessissa sivujen välistä linkkirakennetta ei otettu huomioon painotuksessa kuten ei myöskään sivun sisäistä rakennetta, vaan kaikkien html-tägien sisällöt indeksoitiin samalla tavalla. Ainoa käsittely oli useimmin esiintyvien sulkusanojen jättäminen pois indeksistä. Muuten sanat indeksoitiin ilman käsittelyä taivutusmuotoindeksiin. Tällainen menettely on tyypillistä verkkoarkistojen kokoteksti-indeksien rakentamisessa.

4.3 Hakukone

Suomalaisen verkkoarkiston kokotekstitiedonhaku on toteutettu Apache Solr -ohjelmistolla, joka perustuu avoimella lähdekoodilla toteutettuun Apache Lucene -tiedonhakukirjastoon. Myös tämä tutkimus on toteutettu Apache Solr -ohjelmistolla, koska se mahdollistaa suuren kokoelman käsittelyn tehokkaasti ja tarjoaa tuloksellisen osittaistäsmäyttävän hakumenetelmän joka perustuu vektorimalliin ja tf*idf-painokaavaan. Apache Lucene -menetelmän kaava tuloksen (score) laskemiselle dokumentille (d) hakutermeillä (q) on

$$score(q,d) = co(q,d) * qNorm(q) * \sum_{t \text{ in } q} (tf(t \text{ in } d) \cdot idf(t) 2 * t.gB() * norm(t,d)) \quad (3)$$

jossa

- Co(q,d) tarkoittaa, kuinka monta hakusanoista q esiintyy dokumentissa d.
- qNorm(q) on normalisointikerroin, joka tekee eri kyselyistä keskenään vertailukelpoisia, mutta ei vaikuta yksittäisen kyselyn tulosdokumenttien sijoitukseen.
- tf(t in d) kertoo, kuinka monesti hakusana t esiintyy dokumentissa d.
- idf(t) on käänteinen dokumenttifrekfenssi. Mitä harvemmassa dokumentissa sana t esiintyy, sitä paremman eli suuremman arvon dokumentti saa.

- t.gB() on sanalle t annettava painotuskerroin. Mikäli gB()-arvoa ei anneta erikseen, ei sillä ole vaikutusta painotukseen, mikä oli tilanne tässä tutkimuksessa.

- norm(t,d) on sanojen t esiintymisestä dokumentin d rakenteessa palkitseva kerroin. Tässä tutkimuksessa kaikki sanat indeksointiin samaan kenttään, joten tällä kertoimella ei ollut merkitystä.

Tarkempi kuvaus Lucene Solr -hakumenetelmästä ja yllä esitetyn funktion osista löytyy Apachen dokumentaatiosta (Apache Lucene 2015).

4.4 Hakuaiheet, kyselyt ja relevanssiarviot

Valmiita hakuaiheita menetelmien evaluointiin Suomalaisen verkkoarkiston dokumenteilla ei ollut, joten sellaiset täytyi laatia. Aiheita laadittiin 20 kappaletta, joista 16 päätyi lopulliseen testiin. Aiheita karsittiin, koska vastoin alustavia oletuksia tietyissä aiheissa niitä edustavien dokumenttien määrä testikokoelmassa oli vähäinen.

Aiheiden kartoittamisessa hyödynnettiin luvussa 3.4.1 esiteltyjä tutkimuksia, joissa kartoitettiin Portugalin kansalliskirjaston portugalilaisia verkkosivustoja sisältävän arkiston käyttäjien hakukäyttäytymistä ja kategorisia hakuaiheita. Tämän lisäksi aitoja hakutehtäviä kysyttiin ja saatiin Turun yliopiston digitaalisen kulttuurin tutkimukseen suuntautuneelta professori Jaakko Suomiselta sekä Tampereen yliopiston peli- ja internettutkimukseen suuntautuneelta professori Frans Mäyrältä. Tutkimukseen päätyneistä hakuaiheiden kuvaukset ja hakusanat löytyvät tutkimuksen liitteestä 1. Hakuaiheen osista vain hakusanat on tarkoitettu käytettäviksi kyselyitä suorittaessa.

Relevanssiarvio suoritettiin 4-portaisina (0–3) siten, että

- arvo 0 tarkoittaa dokumentin olevan aiheelle epärelevantti.

- arvo 1 tarkoittaa dokumentin olevan hieman relevantti.

- arvo 2 tarkoittaa dokumentin sisältävän huomattavan määrän relevanttia tietosisältöä.

- arvo 3 tarkoittaa dokumentin kertovan aiheesta kattavasti ja laadukkaasti.

Relevanssiarviot tehtiin perinteisesti ihmistyöllä tutkijan itse tekemänä. Arviot tehtiin pooling-menetelmää soveltamalla siten, että jokaisesta tutkimukseen suunnitellusta 4 menetelmästä otettiin arvioitavien dokumenttien joukkoon 10 ensimmäistä tulosdokumenttia per hakuaihe. Yhteensä arvioituja dokumentteja 16 hakuaiheessa oli 542 eli keskimäärin noin 34 dokumenttia per hakuaihe. Arvioitavien dokumenttien suurta määrää suhteessa pooling-menetelmän syvyyteen ja menetelmien määrään selittää yhden hakumenetelmän huono tuloksellisuus. Yksi menetelmä tuotti pääasiassa epärelevantteja tuloksia, kolmen muun menetelmän tuottaessa enemmän samoja relevantteja tuloksia.

Arvioiden sisäistä luotettavuutta testattiin arvioimalla uudelleen kolmen hakuaiheen pooling-menetelmällä saadut tulosdokumentit. Uudelleen arvioituja dokumentteja oli yhteensä 85. Toistetut relevanssiarviot eivät eronneet alkuperäisistä arvioinneista ollenkaan binäärisellä tasolla. Eli 0-arvon dokumentit olivat samat sekä alkuperäisessä että toistetussa arviossa. Sen sijaan 2-relevanssiarvon dokumenteista yksi todettiin uudelleen arvioinnissa 1-relevanssiarvoiseksi. Lisäksi yksi 2-relevanssiarvon dokumentti todettiin uudelleen arvioinnissa 3-relevanssiarvoiseksi.

Uudelleen arvioinnissa havaittu muutos relevanssiarvoissa ei vaikuttaisi P10-mittarin tuloksiin. Sen sijaan NDCG(10)-mittarilla vaikutusta olisi, mutta se olisi hyvin pientä. Tällaista luotettavuuden mittaamista voidaan pitää luonteeltaan ainakin suuntaa-antava.

4.5 Vertailtavat menetelmät

Tutkimuksessa verrattiin neljää eri menetelmää, jotka ovat perustasona toimineet täysin käsittelemättömät kyselyt, Frequent case generation 3 (FCG3), Simple word ending based rule generator (SWERG+) sekä Snowball-stemmaus yhdistettynä villiin korttiin. Tutkimukseen valittiin nämä menetelmät, koska niitä voidaan soveltaa taituvutusmuotoindeksiin ja ne edustavat monipuolisesti sovellettavissa olevia menetelmiä.

Perustasona toimivat käsittelemättömät kyselyt ovat tässä tapauksessa perusmuodossa olevia hakusanoja. Hakusanat tulivat valmiiksi perusmuodossa hakuaiheiden rakentamisen tuloksena. Teoriassa aidompana perustasona voitaisiin pitää varsinaisen käyttäjän hakutilanteessa järjestelmälle antamia taivutusmuotoja, mutta sellaisia ei tutkimuksessa voitu jokaisella hakuaiheelle tuottaa. Kuitenkin käsittelemättömistä

hakuaiheista saaduista perusmuodossa olevista hakusanoista muodostettavien kyselyjen tuloksista voidaan nähdä, miten lemmattavat kyselyt toimisivat tässä testikokoelmassa taivutusmuotoindeksin kanssa.

FCG3-menetelmä generoi kolminkertaisen määrän hakusanoja perustasoon verrattuna, eli perusmuodon lisäksi yksikön genetiivin ja partitiivin. Menetelmä valittiin tutkimukseen, koska kokoelman dokumenttien taivutusmuotojen oletettiin yleensä täsmäävän sen generoiviin muotoihin. Lisäksi FCG3-menetelmän oletettiin olevan selkeästi erilainen ja tuottavan erilaisia tuloksia kuin perustason.

SWERG+-menetelmä tuottaa FCG3-menetelmään verrattuna huomattavasti enemmän muotoja ja sen pitäisi kattaa kaikki kokoelmassa merkittävissä määrin esiintyvistä taivutusmuodoista. SWERG+-menetelmän nähtiin edustavan paljon muotoja generoivia menetelmiä ja sillä oletettiin olevan selkeä ero FCG3-menetelmään.

Snowball-stemmaus yhdistettynä villiin korttiin tuottaa kaikki mahdolliset sanan vartalon sisältävät muodot. Tämän lisäksi menetelmä saa hakusanat täsmäämään kaikkiin sanoihin, jotka alkavan hakusanan vartalolla. Menetelmän nähtiin myös antavan osviittaa siitä, miten villin kortin soveltaminen testikokoelmassa vaikuttaa tiedonhaun tuloksellisuuteen.

Alla on esitetty yhden hakuaiheen kyselyt jokaiselle menetelmälle esimerkkinä niiden tuottamista kyselyistä.

- Käsittelemätön: urheilu, liikunta, rasismi

- FCG3: urheilu, urheilun, urheilua, liikunta, liikunnan, liikuntaa, rasismi, rasismin, rasismia

- SWERG+: urheilu, urheilut, urheilun, urheilujen, urheilua, urheiluja, urheilussa, urheiluna, urheiluksi, urheilulla, urheilulta, urheilulle, urheilutta, urheiluissa, urheiluina, urheiluiksi, urheiluilla, urheiluilta, urheiluille, urheiluitta, urheilusta, urheiluista, urheiluun, urheiluihin, urheilt, urheiln, urheila, liikunta, liikunnat, liikunnan, liikunnassa, liikunnana, liikunnaksi, liikunnalla, liikunnalta, liikunnalle, liikunnatta, liikunnissa, liikunnina, liikunniksi, liikunnilla, liikunnilta, liikunnille, liikunnitta, liikunnasta, liikunnista, liikunnoissa, liikunnoina, liikunnoiksi, liikunnoilla, liikunnoilta, liikunnoille, liikunnoitta,

liikunnoista, liikuntien, liikuntia, liikuntiin, liikuntojen, liikuntoja, liikuntoihin, liikuntoissa, liikuntoina, liikuntoiksi, liikuntoilla, liikuntoilta, liikuntoille, liikuntoitta, liikuntoista, liikuntaa, liikuntaan, liikuntat, liikuntan, liikuntassa, liikuntana, liikuntaksi, liikuntalla, liikuntalta, liikuntalle, liikuntatta, liikuntasta, liikuntt, liikuntn, liikuntl, rasismi, rasismeja, rasismeissa, rasismeina, rasismeiksi, rasismeilla, rasismeilta, rasismeille, rasismeitta, rasismeista, rasismeihin, rasismejä, rasismeissä, rasismeinä, rasismeillä, rasismeiltä, rasismeittä, rasismeistä, rasismit, rasismin, rasismien, rasismiä, rasismissä, rasisminä, rasismiksi, rasismillä, rasismiltä, rasismille, rasismittä, rasismistä, rasismiin, rasisll, rasismia, rasismissa, rasismina, rasismilla, rasismilta, rasismitta, rasismista, rasismt, rasismn, rasismä, rasisma

- Snowball-stemmaus yhdistettynä villiin korttiin: urheilu*, liikun*, rasism*

Menetelmiä ja niiden taustalla olevaa teoriaa on esitelty enemmän tutkimuksen luvussa 3.2.

4.6 Evaluointi

Evaluoinnissa tarkastellaan pooling-menetelmällä kerättyjä ja arvioituja kymmentä ensimmäistä tulosta per menetelmä per hakuaihe. Tuloksellisuuden mittareina ovat tarkkuus kymmenen ensimmäisen tulosdokumentin jälkeen (P10) ja kumuloituva hyöty kymmenen ensimmäisen tulosdokumentin jälkeen (CG[10]). P10-lukuarvo kuvaa menetelmän tuloksellisuutta binäärisellä asteikolla. CG(10)-lukuarvo taas kuvaa menetelmän tuloksellisuutta neliportaisella asteikolla. Tuloksista P10 esitetään 16 hakuaiheen P10-arvojen keskiarvona ja CG(10) normalisoituna ja diskontattuna (NDCG[10]) 16 hakuaiheen keskiarvona, jotta menetelmiä voidaan verrata ja esittää kahdella menetelmäkohtaisella tunnusluvulla. Tässä luvussa esitettävä kumuloituvan hyödyn laskemisesimerkki mukailee Järvelinin ja Kekäläisen (2002) esimerkkiä.

Taulukossa 2 on esimerkkinä esitetty kolmen menetelmän hakuaiheelle q saamat tulokset järjestyksessä ensimmäisestä tuloksesta kymmenenteen aiemmin luvussa esitetyn 4-portaisen relevanssin mukaan.

Taulukko 2, esimerkkitulokset kolmelle eri menetelmälle yhdelle hakuaiheelle

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
menetelmä1	3	3	2	1	3	0	0	0	1	0
menetelmä2	3	3	2	1	3	0	0	0	1	1
menetelmä3	1	3	1	1	0	0	1	0	0	0

Taulukon 2 arvot täytyy muuttaa binäärisiksi P10-arvon laskemiseksi. Tässä tutkimuksessa käytettiin ns. liberaalia tulkintaa, jonka mukaan relevanteiksi todetaan kaikki dokumentit jotka eivät olleet epärelevantteja. Binäärinen esitys on esitetty taulukossa 3.

Taulukko 3, esimerkkitulokset kolmelle eri menetelmälle yhdelle hakuaiheelle binäärisinä

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.		p10
menetelmä1	1	1	1	1	1	0	0	0	1	0		0.6
menetelmä2	1	1	1	1	1	0	0	0	1	1		0.7
menetelmä3	1	1	1	1	0	0	1	0	0	0		0.5

P10-arvon laskeminen yhdelle hakuaiheelle on yksinkertaista. Siihen sovelletaan tarkkuuden laskemiseen käytettyä kaavaa

$$\text{tarkkuus} = \frac{\text{relevanttien tulosdokumenttien määrä}}{\text{tulosdokumenttien määrä}} \tag{4}$$

jonka perusteella menetelmä1:n P10-arvo on (1+1+1+1+1+0+0+0+1+0) / 10 = 0,6.

CG(10)-arvon laskeminen yhdelle hakuaiheelle tapahtuu kaavalla

$$CG[i] = \begin{cases} G[i], & \text{jos i = 1} \\ CG[i-1] + G[i], & \text{jos i > 1} \end{cases} \tag{5}$$

jonka perusteella taulukon 2 tuloksille saadaan vektorit, joiden arvot on esitetty riveittäin taulukossa 4.

Taulukko 4, CG-hakuvektorien arvot

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
v_{cg_men1}	3	6	8	9	12	12	12	12	13	13
v_{cg_men2}	3	6	8	9	12	12	12	12	13	14
v_{cg_men3}	1	4	5	6	6	6	7	7	7	7

Kumuloituvassa hyödyssä arvot nousevat tasaisesti riippumatta tulosdokumentin sijoituksesta tuloslistalla. Dokumentin sijainti tuloslistalla voidaan ottaa huomioon diskonttaamalla, jolloin mittari palkitsee sitä paremmin mitä lähempänä tuloslistan alkua relevantin dokumentit ovat. Diskontattu kumuloituva hyöty voidaan laskea kaavalla

$$DCG[i] = \begin{cases} CG[i], & \text{jos } i < b \\ DCG[i-1] + G[i]/\log_b(i), & \text{jos } i \geqslant b \end{cases} \quad (6)$$

jossa b on logaritmin kantaluku. Kaavasta kannattaa huomioida, että diskonttausta ei kuitenkaan pidä suorittaa kantalukua pienemmän sijaluvun tuloslistan dokumenteille, koska se mitätöisi niiden arvon. B:n arvolla 2 taulukon 4 kumuloituvat arvot on esitetty diskontattuna taulukossa 5.

Taulukko 5, DCG-hakuvektorien arvot

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
v_{dcg_men1}	3	6	7.26	7.76	9.05	9.05	9.05	9.05	9.37	9.37
v_{dcg_men2}	3	6	7.26	7.76	9.05	9.05	9.05	9.05	9.37	9.67
v_{dcg_men3}	1	4	4.63	5.13	5.13	5.13	5.49	5.49	5.49	5.49

Jotta (diskontattua) kumuloituvaa hyötyä voitaisiin verrata yli hakualueiden, voidaan arvo normalisoida välille 0–1. Normalisointi tehdään muodostamalla ihanteellinen vektori, jossa hakuaiheen tulosdokumentit on järjestetty relevanssin perusteella. Ihannevektorin muodostamisessa tulee käyttää kaikkia saantikannan relevantteja dokumentteja. Mikäli oletetaan, että menetelmien 1–3 kymmenen ensimmäistä tulosta muodostavat saantikannan ja että niiden relevantit dokumentit ovat päällekkäisiä, saadaan ihannevektoriksi v_i, ja ihanteelliseksi kumuloituvan hyödyn vektoriksi v_{cd_i} sekä ihanteelliseksi diskontatuksi kumuloituvan hyödyn vektoriksi v_{dcg_i}. Oletuksien mukaan

saantikannassa on siis 3 3-relevanssinarvon dokumenttia, 1 2-relevanssiarvon dokumentti ja 4 1-relevanssiarvon dokumenttia. Ihannevektorit on esitetty taulukossa 6.

Taulukko 6, ihannevektorien arvot

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
v_i	3	3	3	2	1	1	1	1	0	0
v_{cg_i}	3	6	9	11	12	13	14	15	15	15
v_{dcg_i}	3	6	7.89	8.89	9.32	9.71	10.07	10.4	10.4	10.4

Taulukossa 5 esitetyt menetelmien saamat (diskontatun) kumuloituvan hyödyn vektorit voidaan normalisoida ihannevektoria hyödyntäen kaavalla

$$v_{ndcg\_menx} = v_{dcg\_menx}[1]/v_{dcg\_i}[1], v_{dcg\_menx}[2]/v_{dcg\_i}[2], \ldots, v_{dcg\_menx}[k]/v_{dcg\_i}[k] \quad (7)$$

jonka perusteella menetelmien normalisoitujen vektorien arvoiksi saadaan taulukossa 7 esitetyt arvot.

Taulukko 7, NDCG-hakuvektorien arvot

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
v_{ndcg_men1}	1	1	0.92	0.87	0.97	0.93	0.9	0.87	0.9	0.9
v_{ndcg_men2}	1	1	0.92	0.87	0.97	0.93	0.9	0.87	0.9	0.93
v_{ndcg_men3}	0.33	0.67	0.59	0.58	0.55	0.53	0.55	0.53	0.53	0.53

Tulosten merkitsevyys testattiin Friedmanin testillä, joka on tarkoitettu useammalle kuin kahdelle toisistaan riippumattomalle otokselle, joiden arvot ovat lohkoittain eikä arvojen vaadita toteuttavan normaalijakaumaa (Conover 1971, 265). Lisäksi menetelmien parittaisvertailua varten suoritettiin Wilcoxonin merkittyjen sijalukujen testi. Wilcoxonin testi on tarkoitettu kahdelle otokselle, joiden arvot ovat lohkoittain eikä niiden vaadita toteuttavan normaalijakaumaa (Conover 1971, 206).

4.7 Tulokset

P10-arvojen keskiarvot 16 hakuaiheelle neljälle tutkimuksessa verratulle menetelmälle esitetään taulukossa 8. FCG3 osoittautui P10-mittarilla selkeästi parhaaksi menetelmäksi saaden keskiarvoksi 0,694. SWERG+ oli keskiarvolla 0,500 vain niukasti

parempi käsittelemätöntä vaihtoehtoa, joka sai keskiarvoksi 0,481. Snowball-stemmaus yhdistettynä villiin korttiin osoittautui tulokseltaan huonoksi menetelmäksi saaden keskiarvoksi vain 0,255.

Taulukko 8, P10- ja NDCG(10)-keskiarvot 16 hakuaiheelle

Menetelmä	**P10**	**NDCG(10)**
Käsittelemätön	0.481	0.497
FCG3	0.694	0.598
SWERG+	0.500	0.470
Snowball ja villikortti	0.225	0.145

Myös NDCG(10)-arvojen vertailussa FCG3 osoittautui selkeästi parhaaksi menetelmäksi keskiarvolla 0,598. Toisin kuin P10-mittarilla käsittelemätön vaihtoehto oli niukasti SWERG+:a parempi. Käsittelemätön sai NDCG(10)-keskiarvoksi 0,497 SWERG+:n saadessa 0,470. Snowball-stemmaus yhdistettynä villiin korttiin oli taas tuloksellisesti heikoin saaden keskiarvoksi vain 0,145. NDCG(10)-arvojen keskiarvot 16 hakuaiheelle neljälle tutkimuksessa verratulle menetelmälle esitetään taulukossa 8.

Tulosten merkitsevyyden määrittämiseksi käytettiin Heikkilän (2008, 195) esittelemää luokitusta, jonka mukaan p-arvon ollessa enintään 0,001 tilastollinen merkitsevyys on erittäin merkitsevä. Enintään 0,01 on merkitsevä, enintään 0,05 on melkein merkitsevä ja enintään 0,1 suuntaa-antava. Friedmanin testissä saatiin P10-arvojen p-arvoksi 0,000 ja NDCG(10)-arvojen p-arvoksi 0,000. Molemmilla mittareilla saatuja eroavaisuuksia voidaan pitää tilastollisesti luokituksen mukaan erittäin merkitsevänä. Friedmanin testin perusteella voidaan todeta, että suurella todennäköisyydellä ainakin jonkin menetelmäparin välillä on tilastollisesti merkitseväksi todettu tuloksellisuusero. Toisin sanoen mitattu ero ei johtune sattumasta.

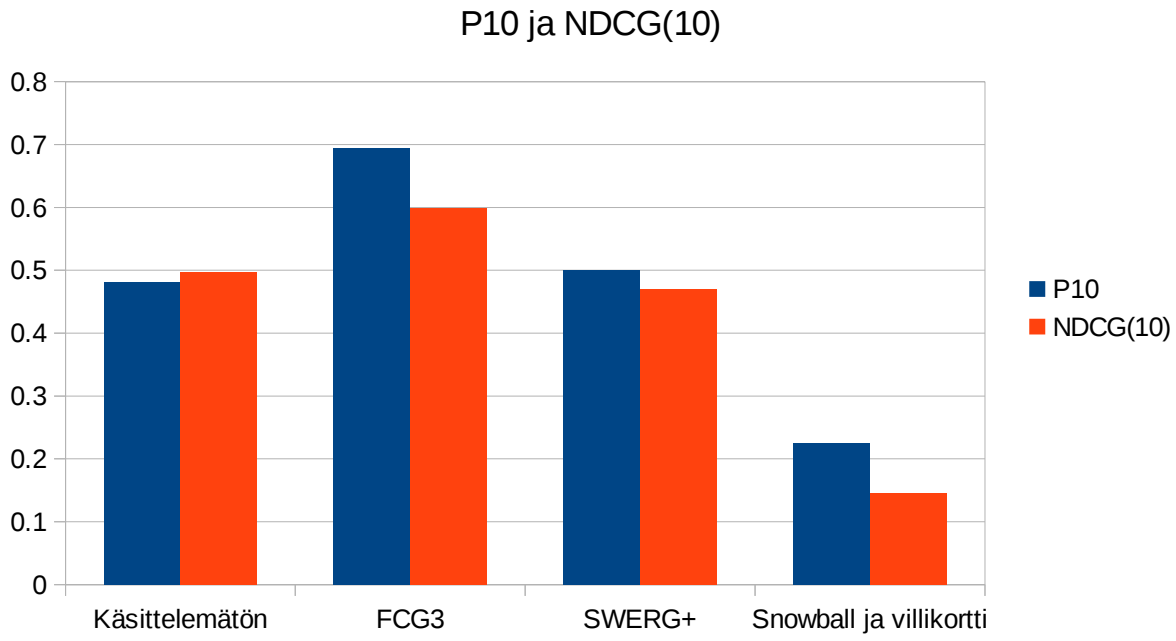
Koska Friedmanin testissä löytyi merkitsevyyseroja, täytyi suorittaa Wilcoxonin parivertailu. Sen perusteella P10-arvoja verrattaessa FCG3-menetelmän eroa perustasona toimineeseen käsittelemättömään vaihtoehtoon voidaan pitää tilastollisesti merkitsevänä p-arvon (0,002) ollessa alle 0,01. NDCG(10)-arvojen eroa FCG3-menetelmän ja käsittelemättömän välillä voidaan pitää suuntaa-antavana p-arvon (0,098) ollessa alle 0,1. Samoin P10-arvoja verrattaessa FCG3-menetelmän eroa SWERG+-menetelmään voidaan pitää tilastollisesti suuntaa-antavana p-arvon ollessa

0,012. Lisäksi kaikkien muiden menetelmien eroa Snowball-stemmauksen ja villin kortin yhdistelmään voidaan pitää tilastollisesti vähintään merkitsevänä paitsi P10-arviolla käsittelemättömään verrattuna jolloin ero on melkein merkitsevä. Muita parivertailuja ei voida pitää tilastollisesti merkitsevinä. Parivertailun p-arvot on esitetty taulukossa 9.

Taulukko 9, Wilcoxonin parivertailun p-arvot (σ) P10- ja NDCG(10)-arvoille 16 hakuaiheelle

Pari	**σP10:**	**σNDCG(10)**
Käsittelemätön – FCG3	0.002	0.098
Käsittelemätön – SWERG+	0.703	0.679
Käsittelemätön – Snowball + villi kortti	0.014	0.001
FCG3 – SWERG+	0.012	0.148
FCG3 – Snowball + villi kortti	0.001	0.000
SWERG+ - Snowball + villi kortti	0.001	0.001

Tilastotestaus mittaa kuitenkin vain sitä, johtuvatko erot sattumasta vai ovatko erot todennäköisesti vaikuttavia. Käytännössä tuloksia voidaan verrata soveltamalla tulosten vaikutusta tiedonhaun tuloslistaan. Tutkimuksen tulosten perusteella FCG3-menetelmä tuottaisi keskimäärin 7 relevanttia dokumenttia 10 ensimmäisen tuloksen joukkoon eli ensimmäiselle tulossivulle. Käsittelemätön vaihtoehto ja SWERG+-menetelmä sen sijaan tuottaisivat keskimäärin 5 relevanttia dokumenttia 10 ensimmäisen tuloksen joukkoon. Heikoiten tuloksellisesti pärjännyt Snowball-stemmaus yhdistettynä villiin korttiin tuottaisi keskimäärin vain 2 relevanttia tulosta ensimmäiselle tulossivulle. P10- ja NDCG(10)-arvot on havainnollistettu pylväinä kuvassa 6.



Kuva 6, P10- ja NDCG(10)-keskiarvot 16 hakuaiheelle

Toistaiseksi esitellyt tulokset kertovat tutkimuksesa käytettyjen hakuaiheiden keskimääräisistä tuloksista. Kuitenkin yksittäisissä hakuaiheissa ja käytännössä yksittäisissä hauissa tulokset voivat poikeata keskimääräisistä tuloksista hakuaiheen erityispiirteiden takia. Tämän takia tuloksia on hyvä tarkastella myös hakuaiheittain ja selvittää erityisesti miksi joidenkin hakuaiheiden tulokset mahdollisesti poikkeavat keskiarvosta. Hakuaiheittain tuloksia esitellään taulukossa 10, josta näkyy menetelmittäin P10-tulos jokaiselle 16 hakuaiheelle ja menetelmän sijoitus keskinäisessä vertailussa.

Taulukko 10, P10-tulokset hakuaiheittain

Hakuaihe	**Käsittelem.**		**FCG3**		**SWERG+**		**Snowball+vk**	
	P10	sija	P10	sija	P10	sija	P10	sija
1	0.7	2	0.9	1	0.5	3	0	4
2	0.4	4	0.5	2	0.7	1	0.5	2
3	0.2	2	0.6	1	0.1	3	0	4
4	0.5	4	0.9	3	1	1	1	1
5	0.2	2	0.7	1	0.2	2	0.1	4
6	0.2	3	0.4	1	0.4	2	0	4
7	0.8	2	0.9	1	0.8	2	0	4
8	0.5	2	0.6	1	0.3	3	0.2	4
9	0.8	3	1	1	0.9	2	0.8	3
10	0.5	2	0.9	1	0.2	3	0.2	3
11	0.5	3	0.8	1	0.7	2	0.1	4
12	0	3	0.2	2	0.3	1	0	3
13	0.1	3	0.4	1	0.2	2	0.1	3
14	0.6	3	0.9	1	0.8	2	0.2	4
15	0.7	1	0.6	2	0.5	3	0.1	4
16	1	1	0.8	2	0.4	3	0.3	4
keskiarvo	0.481		0.694		0.500		0.225	

Hakuaiheittain tuloksia tarkastelu tukee havaintoa, että FCG3-menetelmä on yleensä menetelmistä paras tai hyvin lähellä parasta menetelmää. Poikkeuksia on kuitenkin olemassa. Hakuaiheissa 2 ja 16 FCG3-menetelmä häviää hakuaiheessa parhaalle menetelmälle enemmän kuin kymmenyksellä vaikkakin vain kahdella kymmenyksellä. Hakuaiheessa 2 on huomattava, että SWERG+-menetelmä ja Snowball-stemmaus yhdistettynä villiin korttiin ovat vähintään yhtä hyviä tai parempia kuin FCG3 käsittelemättömän ollessa heikoin menetelmä. Tässä hakuaiheessa hieman

poikkeuksellinen tulos selittyy mahdollisesti sillä, että aiheen kysely oli pitkähkö ja kolme sanaa (twitter, jaiku, qaiku) neljästä ovat erisnimiä jotka viittaavat ainoastaan hakuaiheessa tarkoitettuihin sanoihin. Sen sijaan hakuaiheessa 16 on huomattava, että FCG3 on kuitenkin selvästi parempi kuin selvästi enemmän muotoja generoivat SWERG+-menetelmä ja Snowball-stemmaus yhdistettynä villiin korttiin, vain yhtä muotoa edustavan käsittelemättömän kyselyn ollessa paras. Sama trendi on havaittavissa kyselyssä 15, jonka kysely on aiheen 16 tapaan pitkähkö ja sisältää erisnimiä. Näissä kyselyissä erisnimet ja sanojen lukumäärä luultavasti takaavat hyvän tuloksellisuuden maltillisella muotojen generoinnilla.

Hakuaiheen 2 lisäksi Snowball-stemmaus yhdistettynä villiin korttiin saa poikkeuksellisen hyvän tuloksen myös hakuaiheissa 4 ja 9. Näille poikkeuksille syy on mahdollisesti siinä, että niissä on keskeisessä osassa yhdyssana. Yhdyssanoja sisältävien kyselyjen on havaittu toimivan hyvin stemmauksen tai runsaasti muotoja generoivien käsittelyjen yhteydessä (Järvelin et al. 2015).

5 KESKUSTELU JA YHTEENVETO

Suomenkielistä verkkoarkistoa ei ole aiemmin käytetty tiedonhaun evaluointitutkimuksen kokoelmana. Tässä tutkimuksessa verkkoarkistosta rakennettua testikokoelmaa käytettiin neljän kyselynkäsittelymenetelmän evaluoinnissa taivutusmuotoindeksissä. Menetelmien evaluoinnissa saatiin seuraavat tulokset:

1. FCG3-menetelmä paransi tuloksia verrattuna perustason käsittelemättömiin kyselyihin.

2. SWERG+-menetelmä ei parantanut tuloksia verrattuna perustason käsittelemättömiin kyselyihin.

3. Snowball-stemmaus yhdistettynä villiin korttiin heikensi tuloksia verrattuna perustason käsittelemättömiin kyselyihin.

Aiemmasta luvussa 3.4.2 esitellystä artikkelikokoelmissa suoritetusta evaluointitutkimuksesta tulokset erosivat seuraavilta kohdin:

1. Runsaasti taivutusmuotoja generoiva SWERG+-menetelmä ei ollut parempi kuin niukasti muotoja generoiva FCG3-menetelmä.

2. SWERG+-menetelmä ei ollut parempi kuin perustason käsittelemättömät kyselyt.

SWERG+-menetelmän tuloksellista heikkoutta verrattuna aiempaan tutkimukseen voidaan mahdollisesti selittää käytetyn kokoelman erityispiirteillä. Artikkelikokoelmien dokumentteja voidaan pitää laadukkaina, ja siten niissä useassa eri muodossa esiintyvät sanat kertovat oletettavasti paljon hakuaiheesta. Sen sijaan verkkoarkistossa dokumenttien laatu on vaihtelevampaa ja useassa eri muodossa esiintyvät sanat eivät välttämättä kerro enää varsinaisesta hakuaiheesta. Tätä oletusta tukee myös se, että Snowball-stemmaus yhdistettynä villiin korttiin tuotti heikkoja tuloksia. On kuitenkin hyvä huomioida, että esimerkiksi yhdyssanojen osalta Snowball-stemmaus yhdistettynä villiin korttiin toimi hyvin.

Vastaavasti FCG3-menetelmän hyvää tuloksellisuutta verrattuna aiempaan tutkimukseen voidaan mahdollisesti pitää SWERG+-menetelmän heikkouden tapaan riippuvaisena

kokoelman erityispiirteistä. FCG3-menetelmän generoimat muodot (nominatiivi, genetiivi, partitiivi) muodostavat suomen kielessä lauseen subjektin. Subjekti yleensä kertoo mistä asiassa lauseessa on kyse (Lahtinen 2000, 51). Asian selvittämiseksi tarkemmin pitäisi tutkia lauserakenteen huomioon ottavien hakumenetelmien tuloksellisuutta verkkoarkistoaineistolla.

Tutkimuksen tuloksia tarkastellessa kannattaa pitää mielessä, että tutkimuksessa käytettiin vain 16 hakuaihetta. Luotettavampien tutkimustulosten saamiseksi olisi jatkossa syytä vertailla menetelmiä useammalla hakuaiheella. Tällöin sekä tilastollinen merkitsevyys olisi paremmin todennettavissa että evaluointia voitaisiin pitää luotettavampana (Carterette et al. 2008; Sanderson & Zobel 2005). Toisaalta jo suuntaa-antavien tulosten perusteella voidaan olettaa verkkoarkiston kokoelmana aiheuttavan erilaisia vaatimuksia hakumenetelmille kuin esimerkiksi artikkelikokoelmat asettavat. Tutkimustuloksia voidaan mahdollisesti pitää mielenkiintoisina muihin verkkoarkiston kaltaisiin vaihtelevan laadukkaita dokumentteja sisältäviin kokoelmiin hakumenetelmiä kehitettäessä ja valittaessa.

Tutkimuksen tuloksia hyödyntäessä on syytä muistaa, että tutkimuksessa vertailtiin ainoastaan luonnollisen suomen kielen käsittelymenetelmiä ja tarkemmin rajattuna taivutusmuotoindeksiin sopivia generatiivisia menetelmiä. Tuloksia ei siis välttämättä kannata soveltaa suoraan esimerkiksi web-tiedonhaun järjestelmiin, joissa tulosten järjestämiseen vaikuttaa yleensä esimerkiksi myös PageRank ja tiedonhakijan hakuhistoria. Sen sijaan tulokset ovat oletettavasti mielenkiintoisempia sellaisen tutkimuksen näkökulmasta, jossa ei olla lähtökohtaisesti kiinnostuneita (ainoastaan) PageRankin kaltaisten mittareiden korottamista suosituista sivustoista. Tutkimuksessa mielenkiintoiseksi voidaan kokea myös tai erityisesti muiden mittareiden perusteella vähemmän suositut sivustot, milloin niiden löytymiseksi luonnollisen kielen käsittelyn osalta tuloksellisten menetelmien merkitys korostuu.

Tämä tutkimus on vasta ensimmäinen suomenkielistä verkkoarkistoa kokoelmana käyttävä tutkimus, joka toivottavasti antaa vihjeitä jatkotutkimukselle. Verkkoarkistojen tiedonhaun kehittäminen on toistaiseksi ollut maltillista arkistoinnin keskittyessä pääasiassa aineiston hankintaan ja säilytykseen. Tämä on tapahtunut aineiston saatavuuden ja haun kustannuksella resurssien ollessa vaatimattomia suhteessa aineiston kokoon ja monimuotoisuuteen. Tarve verkkoarkistojen tiedonhakumenetelmien

evaluoinnille ja kehittämiselle kasvanee, kun tulevaisuudessa verkkoarkistojen asema tutkimuksen aineistona muuttuu vahvemmaksi.

LÄHTEET

Agichtein, E., Brill, E. & Dumais, S. 2006. Improving web search ranking by incorporating user behavior information. Teoksessa Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06). New York, NY, USA: ACM, 19–26.

Airio, E. 2006. Word Normalization and Decompounding in Mono- and Bilingual IR. Information Retrieval 9 (3), 249–271.

Alkula, R. 2000. Merkkijonoista suomen kielen sanoiksi: Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja hakuun. Acta electronica Universitatis Tamperensis 761. Tampere: Tampereen yliopisto.

Alonso, O., Gertz, M. & Baeza-Yates, R. 2007. On the value of temporal information in information retrieval. ACM SIGIR Forum 41 (2), 35–41.

Apache Lucene 2015
<https://lucene.apache.org/core/4_6_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html> (Viitattu 28.11.2015)

Bailey, P., Craswell, N., Soboroff, I., Thomas P., de Vries, A., P. & Yilmaz E. 2008. Relevance assessment: are judges exchangeable and does it matter. Teoksessa Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). New York, NY, USA: ACM, 667–674.

Borlund, P. 2000. Evaluation of interactive information retrieval systems. Turku: Åbo Akademi University Press.

Broder, A. 2002. A taxonomy of web search. ACM SIGIR Forum 36, (2), 3–10.

Callan, J., P., Croft, W., B & Harding, S., M. 1992. The INQUERY Retrieval System. Teoksessa Tjoa, A, M, & Ramos, I. (toim.) Database and Expert Systems Applications. Vienna: Springer, 78–83.

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A. & Allan, J. 2008. Evaluation over thousands of queries. Teoksessa Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). New York, NY, USA: ACM, 651–658.

Chowdhury, G. G. 2010. Introduction to modern information retrieval. 3. painos. London: Facet Publishing.

Cleverdon, C., W. & Keen, M. 1966. Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 2, Test results. <http://hdl.handle.net/1826/863> (Viitattu 28.11.2015)

Clough, P. & Sanderson, M. 2013. Evaluating the performance of information retrieval systems using test collections. Information research 18 (2) artikkeli 582.

Conover, W. J. 1971. Practical nonparametric statistics. New York: Wiley.

Costa, M. & Silva, M., J. 2010. Understanding the information needs of web archive users. Teoksessa Proceedings of the 10th International Web Archiving Workshop. 9–16.

Costa, M. & Silva, M., J. 2011. Characterizing search behavior in web archives. Teoksessa Proceedings of the 1st International Temporal Web Analytics Workshop. CEUR Workshop Proceedings 707.

Costa, M. & Silva, M., J. 2012. Evaluating web archive search systems. Teoksessa Wang, X., S., Cruz, I., Delis, A. & Huang, G. (toim.) Lecture Notes in Computer Science 7651. Heidelberg: Springer, 440–454.

Croft, B., Metzler, D., & Strohman, T. 2010. Search Engines: Information retrieval in practice. Boston: Addison-Wesley.

Efron, M. & Winget, M. 2010. Query polyrepresentation for ranking retrieval systems without relevance judgments. Journal of the American society for information science and technology 61 (6), 1532–2890.

Heikkilä, T. 2008. Tilastollinen tutkimus. 7. uudistettu painos. Helsinki: Edita.

Hollink, V., Kamps, J., Monz, C. & de Rijke, M. 2004. Monolingual document retrieval for European language. Information Retrieval 7 (1), 33–52.

IIPC Access Working Group. 2006. Use cases for access to internet archives. <http://www.netpreserve.org/sites/default/files/resources/UseCases.pdf> (Viitattu 28.11.2015)

Ingwersen, P. & Järvelin, K. 2005. The turn : integration of information seeking and retrieval in context. Dordrecht: Springer.

Jansen, B., J., Booth D., L. & Spink, A. 2008. Determining the informational, navigational, and transactional intent of Web queries. Information processing & management 44 (3), 1251–1266.

Jansen, B., J., Spink, A. & Saracevic, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. Information processing and management: an international journal 36 (2), 207–227.

Järvelin, A, Keskustalo, H., Sormunen, E., Saastamoinen, M. & Kettunen, K. 2015. Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. Journal of the association for information Science and Technology (verkkojulkaisu). <http://onlinelibrary.wiley.com/doi/10.1002/asi.23379/abstract> (Viitattu 28.11.2015)

Järvelin, K. & Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20 (4), 422–446.

Karlsson, F. 2008. Yleinen kielitiede. Uudistetun laitoksen 3. painos. Helsinki: Gaudeamus Helsinki University Press.

Kekäläinen, J. & Järvelin, K. 2002. Evaluating information retrieval systems under the challenges of interaction and multi-dimensional dynamic relevance. Teoksessa Bruce, H., Fidel, R., Ingwersen, P. & Vakkari P. (toim.) Proceedings of the 4th CoLIS Conference. Greenwood Village, CO: Libraries Unlimited, 253–270.

Keskustalo, H. 2010. Towards simulating and evaluating user interaction in information retrieval using test collections. Acta electronica Universitatis Tamperensis 1012. Tampere: Tampere University Press.

Kettunen, K. 2004. Covering the morphological variation of Finnish query nouns in a probabilistic best-match system. Teoksessa The first baltic conference, human language technologies - The Baltic perspective. Riga, Latvia. April 21–22, 2004, 73–80.

Kettunen, K. 2007. Reductive and generative approaches to morphological variation of keywords in monolingual information retrieval. Acta Universitatis Tamperensis 1261.

Kettunen K. 2008. Automatic generation of frequent case forms of query keywords in text retrieval. Advances in natural language processing. Lecture Notes in Computer Science 5221, 222–236.

Kettunen, K and Airio, E. 2006. Is a morphologically complex language really that complex in full-text retrieval?. Teoksessa Salakoski T., Ginter F., Pyysalo S. & Pahikkala T. (toim.) Advances in Natural Language Processing, LNAI 4139, Heidelberg: Springer-Verlag, 411–422.

Kettunen, K. & Airio, E. & Järvelin, K. 2007. Restricted inflectional form generation in management of morphological keyword variation. Information Retrieval 10 (4–5), 415–444.

Kettunen K. & Arvola P. 2012. Generating variant keyword forms for a morphologically complex language leads to successful information retrieval with Finnish. Teoksessa Salampasis M. & Larsen B. (toim.) Multidisciplinary Information Retrieval : 5th Information Retrieval Facility Conference IRFC 2012 Vienna, Australia, July 2–3, 2012 Proceedings. Lecture Notes in Computer Science 7356. Heidelberg: Springer, 113–126.

Kinney, K., A., Huffman, S., B. & Zhai J. 2008. How evaluator domain expertise affects search result relevance judgments. Teoksessa Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08), 591–598.

Lahtinen, T. 2000. Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods. <http://urn.fi/URN:ISBN:951-45-9640-4> (Viitattu 28.11.2015)

Lancaster F., W. 1971. The cost-effectiveness analysis of information retrieval and dissemination systems. Journal of the American society for information science 22 (1), 12–27.

Lingsoft 2015. <http://www.lingsoft.fi> (Viitattu 28.11.2015)

Mizzaro S. 1997. Relevance: The whole history. Journal of the American society for information science 48 (9), 810–832.

Monz, C. & de Rijke, M. 2002. Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. Teoksessa Peters, C., Braschler, M., Gonzalo, J. & Kluck, M. (toim.) Lecture notes in computer science 2406. Heidelberg: Springer, 262–277.

Porter, M., F. 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Robertson, A., M. & Willett, P. 1998 Applications of n-grams in textual information systems. Journal of documentation 54 (1), 48–67.

Rose D. & Levinson, D. 2004. Understanding user goals in web search. Teoksessa Proceedings of the 13th International Conference on World Wide Web, 13–19.

Salton, G. 1986. Another look at automatic text-retrieval systems. Communications of the ACM 29 (7), 648–656.

Salton, G. & McGill, M., J. 1987 Introduction to modern information retrieval. 3. painos. Auckland: McGraw-Hill.

Salton, G., Wong, A. & Yang C., S. 1975. A vector space model for automatic indexing. Communications of the ACM 18 (11), 613–620.

Sanderson, M. & Braschler, M. 2009. Best practices for test collection creation and information retrieval system evaluation. <http://www.seg.rmit.edu.au/mark/publications/my_papers/T-CLEF-test-collection-report.pdf> (Viitattu 28.11.2015)

Sanderson M. & Zobel J. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. Teoksessa Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05), 162–169.

Saracevic, T. 1996. Relevance reconsidered. Teoksessa Proceedings of the second conference on conceptions of library and information science, Copenhagen, Denmark, 201–218.

Snowball 2015. <http://snowball.tartarus.org> (Viitattu 28.11.2015)

Soboroff, I., Nicholas, C. & Cahan P. 2001. Ranking retrieval systems without relevance judgments. Teoksessa Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). ACM: New York, NY, USA, 66–73.

Sormunen, E., Kekäläinen, J., Koivisto, J. & Järvelin, K. 2001. Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. Journal of documentation 57(3), 358–376.

Spärck-Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28 (1), 11–21.

Tomlinson, S. 2004. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServerTM at CLEF 2003. Teoksessa Peters, C., Gonzalo, J., Braschler, M. & Kluck, M. (toim.) Comparative evaluation of multilingual information access systems. Lecture notes in computer science 3237, 286–300.

Vakkari, P. 2001. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. Journal of documentation 57 (1), 44–60.

Voorhees, E., M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. Teoksessa Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM: New York, NY, USA, 315-323.

Voorhees, E., M. 1999. Natural language processing and information retrieval. Teoksessa Pazienza M., T. (toim.) Information extraction: Towards scalable, adaptable systems. London: Springer-Verlag, 32–48.

Voorhees E., M. & Harman D., K. (toim.). 2005. TREC : experiment and evaluation in information retrieval. Cambridge (Mass.): MIT Press, cop.

Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments?. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM: New York, NY, USA, 307–314.

LIITE 1: HAKUAIHEET

1

T: sosiaalinen media

D: Sosiaalisesta mediasta kertovien sivujen asennoituminen aiheeseen.

2

T: mikroblogi jaiku qaiku twitter

D: Sivuja joissa arvioidaan tai esitellään mikroblogipalveluita (jaiku qaiku twitter).

3

T: talous tulevaisuus

Millaisena talouden tulevaisuus nähdään.

4

T: verkkomainonta

D: Verkkomainonnasta kertovien sivujen asennoituminen ja käsitys verkkomainonnasta.

5

T: urheilu liikunta rasismi

D: Urheilussa esiintyvästä rasismista kertovia sivuja.

6

T: nokia laatu puhelin

D: Sivuja joissa otetaan kantaa Nokian puhelimien laatuun.

7

T: sims peli

D: Sims-peleihin liittyvää keskustelua.

8

T: sasi kolari onnettomuus

D: Kansanedustaja Sasin kolarionnettomuudesta kertovia sivuja.

9

T: tallinna pronssisoturi

D: Tallinan pronssisoturipatsaan siirrosta ja sen lievelilmiöistä kertovia sivuja.

10

T: tupakkalaki kielto

D: Uudesta tupakkalaista ja tupakoinkiellosta kertovia sivuja.

11

T: auringonpimennys

D: Auringonpimennyksestä kertovia sivuja.

12

T: makasiini tulipalo

D: Makasiinien tulipalosta kertovia sivuja.

13

T: italia ranska finaali jalkapallo berliini

D: Vuoden 2010 jalkapallon mm-kilpailujen finaalista kertovia sivuja.

14

T: salmonella epidemia

D: Salmonellaepidemioista kertovia sivuja.

15

T: maj marma ukkonen

D: Maj Karman Ukkonen-albumista kertovia sivuja.

16

T: tarja halonen kampanja tuki

D: Tarja Halosen kampanjan tuesta kertovia sivuja.