Hyvin toimivien hakulausekkeiden muotoilu ja hakujen onnistumiseen vaikuttavat tekijät täys- ja osittaistäsmäyttävässä hakujärjestelmässä

Tampereen yliopisto
Informaatiotutkimuksen laitos
Informaatiotutkimuksen pro gradu
Huhtikuu 2005
Seija Mukala

Tampereen yliopisto

Informaatiotutkimuksen laitos

MUKALA, SEIJA: Hyvin toimivien hakulausekkeiden muotoilu ja hakujen onnistumiseen vaikuttavat tekijät täys- ja osittaistäsmäyttävässä hakujärjestelmässä

Pro gradu -tutkielma, 80 s., 8 liites.

Informaatiotutkimus

Huhtikuu 2005.

Pro gradu -tutkielman aihe on hyvin toimivien hakulausekkeiden muotoilu ja hakujen onnistumiseen vaikuttavat tekijät täys- ja osittaistäsmäyttävässä hakujärjestelmässä. Tutkielma perustuu empiiriseen aineistoon, joka tallennettiin informaatiotutkimuksen perusopintojen tiedonhaun harjoituskurssin yhteydessä Tampereen yliopiston informaatiotutkimuksen laitoksella syksyllä 2003. Informantit olivat harjoituskurssin opiskelijoita. Harjoituksissa käytettiin tiedonhakupeliä. Tutkielmassa analysoidaan ja vertaillaan tiedonhakupelin lokitiedostoon tallentuneiden täys- ja osittaistäsmäyttävien hakujen tuloksellisuutta ja hakuavainten käyttöä hakulausekkeissa. Tarkasteltavaan osa-aineistoon on valittu yhden hakuaiheen 30 tarkinta hakua sekä täys- että osittaistäsmäyttävästä hakujärjestelmästä. Lokitiedostosta eroteltujen parhaiden hakujen luokitteluperuste on hakutulosten tarkkuus. Täsmäytysmallien hakulausekkeet muotoiltiin eri tavalla. Harjoituksissa käytettiin täystäsmäyttävien hakulausekkeiden muotoilussa peräkkäisten fasettien strategiaa. Osittaistäsmäyttävät hakulausekkeet muodostettiin käyttämällä samoja hakuavaimia kuin täystäsmäytyksessä. Tarkastelen osittaistäsmäyttäviä hakuja Boolen hakujen pohjalta.

Opiskelijat kuvasivat kokemuksiaan tiedonhakupelin käytöstä harjoituskurssilla kerätyssä esseeaineistossa. Esseeaineiston kerronta on luokiteltu tiedonhakupelin palautemuotoja, teknisiä ongelmia, parannusehdotuksia ja hakuavainten valintaa koskeviin puheisiin. Esseeaineiston asiakokonaisuuksien luokittelussa on huomioitu myös opastuksen tarpeeseen, harjoitustilanteen ongelmiin sekä kritiikkiin liittyvät puheet. Empiirisen tutkimuksen metodit ovat esseeaineiston osalta sisällönanalyysi, hakulausekkeiden osalta hauissa käytettyjen fasettien analyysi. Esseeaineiston ja lokitietojen tarkastelussa käytetään kvalitatiivisia ja kvantitatiivisia menetelmiä.

Tutkittu osa-aineisto on pieni, joten saatujen tulosten perusteella ei voida tehdä pitkälle vietyjä johtopäätöksiä. Osittaistäsmäyttävät haut ovat tutkitussa osa-aineistossa yleisesti täystäsmäyttäviä tarkempia. Yhteinen piirre molemmille hakutyypeille on se, että toisistaan poikkeavilla hakulausekkeilla on saatu harjoituksissa tuloksia, joiden tarkkuusarvot ryhmässä ovat täsmälleen samoja. Tiedonhakupelin lokitiedoista voi suorittaa otannan, jonka pohjalta tutkitaan hakulausekkeiden muotoilun vaikutusta hakutuloksiin, kun tarkasteluun valitaan useita hakuja samoilta hakijoilta.

Sisällysluettelo

1. Johdanto

Tavoitteenani on tutkia empiirisesti hakulausekkeiden toimivuutta ja hakujen onnistumiseen vaikuttavia tekijöitä täys- ja osittaistäsmäyttävässä hakujärjestelmässä. Täys- ja osittaistäsmäytys edellyttävät erilaista hakulausekkeen muotoilua. Täystäsmäyttävässä hakujärjestelmässä hakutermit yhdistetään Boolen operaattoreilla, jolloin haun tulokseen saadaan vain hakulausekkeeseen täydellisesti täsmäävät dokumentit. TRIP-hakujärjestelmän täystäsmäyttävän hakulausekkeen pituus on rajallinen. Osittaistäsmäytyksessä hakutermit ilmaistaan sanalistana, jonka pituutta ei ole erikseen rajattu. Osittaistäsmäytyksessä voidaan kuitenkin hyödyntää hakulausekkeen muotoilussa fasetointiin perustuvaa rakennetta. Kuten täsmäytysmenetelmän nimi jo kertoo, osittaistäsmäytyksessä myös osittain hakulausekkeeseen täsmäävät dokumentit valikoituvat hakutulokseen.

Tämä tutkielma perustuu laajahkoon empiiriseen aineistoon, joka on tallennettu informaatiotutkimuksen perusopintojen tekstitiedonhaun harjoituskurssin yhteydessä Tampereen yliopiston informaatiotutkimuksen laitoksella syksyllä 2003. Harjoitukset tehtiin tiedonhakupelillä. Informantit ovat kurssin opiskelijoita, jotka perehtyivät tiedonhakupelin käyttöön yhteisissä sali- ja itsenäisesti toteutetuissa verkkoharjoituksissa. Harjoituksissa suoritettiin tiedonhakuja Alma Median lehtiartikkeleita sisältävistä tekstitietokannoista käyttämällä sekä täys- että osittaistäsmäyttävää hakujärjestelmää. Tutkimusaineisto koostuu syksyn 2003 tiedonhakupelin harjoitusten lokitiedoista. Lokitietojen ohella käytän myös opiskelijoiden hakuharjoitusten yhteydessä kirjoittamia esseitä, joita tarkastelen kerrontana. En käsittele työssäni tiedonhakupelin käyttöä oppimisen enkä oppimisympäristön näkökulmista. Kartoitan esseevastauksista perusopintojen tiedonhakukurssin opiskelijoiden kokemuksia täystäsmäyttävän TRIP- ja osittaistäsmäyttävän InQuery-hakujärjestelmän eroista sekä hyvin toimivien hakulausekkeiden löytymisestä. Hakuharjoitusten esseevastauksista saa hyvän käsityksen informaatiotutkimuksen perusopinto-opiskelijoiden noudattamista hakustrategioista ja hakutaktiikoista. Keskityn varsinaisessa lokianalyysissa maailmantalouteen liittyvään hakuaiheeseen. Käytettävissäni on edellä mainitun aineiston lisäksi hakuaiheesta laadittu valmis hakuavainten laatuluokitusanalyysi.

Selvitän, miten tiedonhakuharjoituksiin osallistuneet opiskelijat ovat onnistuneet soveltamaan kahden eri hakujärjestelmän hakumenetelmiä etsiessään relevantteja Etelä-Amerikan velkakriisiä koskevia lehtiartikkeleita tekstitietokannasta. Tutkin lokitietojen pohjalta hyvin onnistuneiden täys- ja osittaistäsmäyttävien hakulausekkeiden tehokkuudessa ja rakenteessa ilmeneviä eroja. Pohdin myös hakujen epäonnistumisen syitä.

2. Tutkimuskäsitteistö

Tekstitietokannoista suoritettava hakuprosessi on monimutkainen, joten hakujärjestelmässä on oltava monipuoliset hakumahdollisuudet. Tekstitietokantojen tiedonhakua helpottaa perusmuotoistettu hakemisto, jossa myös yhdyssanat on pilkottu osiin ja palautettu perusmuotoonsa. Näin sanan eri taivutusmuodot tekstissä löydetään yhdellä hakulausekkeen sanamuodolla. Koska käyttöliittymältä edellytetään käyttäjäystävällisyyttä, käyttöliittymän hahmottaminen ei saa olla liian vaikeaa. Käytön oppimisen tulee olla helppoa ja virhemahdollisuudet on minimoitava. Hakujärjestelmässä pitää olla mahdollisuus paitsi virheiden korjaamiseen, myös haun uudelleenmuotoiluun. Vaikka liittymään on sisällytetty automaattiset tuki- ja ohjaustoiminnot, on hakijalle itselleen jätetty mahdollisuus säännellä hakujärjestelmän toimintoja ja hakuprosessin kokonaisuutta. (Alaterä & Halttunen 2002, 37.)

2.1 Tiedonhakujärjestelmä

Tiedonhakujärjestelmän ydin on *täsmäytysalgoritmi*, joka laskee hakulausekkeen ja dokumentin esitysten samankaltaisuuden. Esityksestä käytetään kirjallisuudessa usein termiä *representaatio*. Hakujärjestelmä tunnistaa samankaltaisuuden perusteella dokumentin kuulumisen tai kuulumattomuuden tulosjoukkoon. Täsmäytysmenetelmä toisin sanoen vertaa tietokantaan tallennettujen dokumenttien representaatioita hakulausekkeen representaatioon. Täsmäytysmenetelmät ovat täydellinen täsmäytys ja osittaistäsmäytys. Täystäsmäyttävä hakujärjestelmä noudattaa Boolen logiikkaa. Hakuaiheeseen liittyvää tietoa etsitään täystäsmäytyksessä hakuavainten kattamalta yhteiseltä käsitteelliseltä alueelta, hakuavaimin määriteltyjen alueiden leikkauskohdasta tai sulkemalla eri alueita pois hakutuloksesta sen mukaan, mitä hakuavaimia dokumenteissa voidaan olettaa esiintyvän. Boolen hakujärjestelmät perustuvat hakulausekkeen ja dokumenttien *täydelliseen täsmäytykseen*, jolloin

relevantit dokumentit hajaantuvat satunnaisesti hakutuloksessa. Dokumentin tulee täyttää hakulausekkeen sille asettamat loogiset ehdot – hakulausekkeessa käytettyjen hakuavainten kombinaatiot. Hakujoukon dokumenttien järjestys on sattumanvarainen.

Osittaistäsmäyttävissä hakujärjestelmissä toimii tulosjoukon *relevanssilajittelu*, jossa dokumentit järjestetään automaattisesti oletetun relevanssin mukaiseen järjestykseen. Tämä on hyödyllistä, kun tietokannat ovat laajoja ja tulosjoukot suuria. (Alaterä & Halttunen 2002, 38 - 42.) Relevanssilajittelussa hakutulos lajitellaan dokumenttien vertailulukujen perusteella alenevan relevanssin järjestykseen, jolloin hakulauseketta parhaiten vastaavat dokumentit tulevat hakutuloslistan kärkeen. Täsmäytys perustuu dokumenttikohtaiseen painoarvoon ja hakuavainten painotukseen. Täsmäytys tapahtuu laskemalla hakulausekkeen ja dokumentin yhteisten sanojen painojen perusteella dokumentille vertailuluku. Dokumenttien indeksoinnissa ja vertailulukujen laskennassa käytetään tilastomatemaattisia menetelmiä. Dokumenttikohtaisten sanojen painoja laskettaessa huomioidaan *termifrekvenssi* ja *dokumenttifrekvenssi* – hakuavaimen esiintymistiheys dokumentissa sekä niiden dokumenttien määrä, joissa hakuavain esiintyy. Termifrekvenssi ja dokumenttifrekvenssi suhteutetaan laskennassa tietokannassa olevien dokumenttien määrään ja pituuteen. Eräs osittaistäsmäytyksen hakumenetelmistä perustuu todennäköisyyslaskentaan. Osittaistäsmäyttävissä hakujärjestelmissä voidaan käyttää myös Boolen operaattoreiden kaltaisia operaattoreita. Hakujoukko järjestyy todennäköisen täsmäävyyden perusteella. Mikäli relevanssipalaute on automaattinen, käyttäjä voi valikoida hakutuloksesta helposti itsensä kannalta hyödylliset dokumentit. Hakujärjestelmä tunnistaa valittujen dokumenttien piirteet ja muotoilee niiden perusteella uuden hakulausekkeen. (Alaterä & Halttunen 2002, 41 - 42.)

Lokitiedosto on tapahtumatiedosto, johon kaikki hakujärjestelmän käyttötiedot kirjautuvat automaattisesti. *Lokianalyysin* avulla voidaan tulkita tiedonhakupelin tapahtumatiedostoon kirjattujen hakulokien sisällön perusteella hakujärjestelmän käyttöä ja hakujen onnistumista käyttäjittäin.

2.2 Relevanssi

Relevanssin käsite on tiedonhaun evaluoinnin kannalta keskeinen. Relevanssin lajeja ovat *aiherelevanssi* ja *käyttäjärelevanssi*. Aiherelevanssilla tarkoitetaan hakulausekkeen ja dokumenttien kuvausten välistä täsmäävyyttä. Tekstitietokannoissa juuri dokumentin aiheenmukaisuus on relevanssin eksplisiittinen kriteeri (Kekäläinen & Järvelin 2002). Aiherelevanssin tutkiminen jättää tiedon käyttäjän huomioimatta (Alaterä & Halttunen 2002, 125 - 126). Käyttäjän kannalta ei kuitenkaan ole merkitsevää haussa löytyneiden dokumenttien mahdollisimman suuri määrä. Käyttäjän intressissä on hyödyllisten, käyttökelpoisten dokumenttien löytyminen tietokannasta. Käyttäjärelevanssilla tarkoitetaankin käyttäjän omaa arviota tiedontarpeensa ja löydetyn informaation suhteesta tietyllä hetkellä (Schamber 1990). Käyttäjärelevanssin arviointi on tiedonhakututkimuksessa ongelmallinen, koska käyttäjän reaktiot eivät sinällään ole toistettavissa koe- tai laboratorio-olosuhteissa. Silti tutkimuksessa tulisi pyrkiä huomioimaan molemmat relevanssin lajit (Alaterä & Halttunen 2002, 126 - 127).

Saracevic (1996) on jaotellut relevanssin käsitteen *algoritmiseen* relevanssiin, *aiherelevanssiin*, *kognitiiviseen* relevanssiin, *tilannerelevanssiin* ja *affektiiviseen* relevanssiin. Algoritmisella relevanssilla tarkoitetaan hakulausekkeen ja dokumentin suhdetta, jonka määrää hakujärjestelmässä käytetty algoritmi, kun hakujärjestelmä etsii tulosta hakuun. Aiherelevanssi on haun aiheen ja dokumentin aiheen suhde. Kognitiivinen relevanssi tarkoittaa käyttäjän tiedontarpeen ja dokumentin suhdetta. Kognitiivinen relevanssi liittyy dokumentin uutuuteen, laatuun ja informatiivisuuteen käyttäjän tiedontilan kannalta. Tilannerelevanssi määräytyy käyttäjän tilanteen, ongelman tai tehtävän mukaan. Dokumentin edellytetään olevan käyttäjälleen hyödyllinen esimerkiksi ongelmanratkaisussa. Affektiivinen relevanssi viittaa käyttäjän aikomuksiin, motivaatioon tai tunteisiin tiedonhaun kontekstissa (Alaterä & Halttunen 2002, 127). Borlund (2000, 29 - 30) huomauttaa, ettei Saracevicin luokituksen mukaista affektiivista, käyttäjän aikomuksiin, motivaatioon tai tunteisiin pohjautuvaa relevanssia voida pitää itsenäisenä relevanssin lajina. Affektiivisuus on Borlundin mukaan sidoksissa kaikkiin subjektiivisen relevanssin tyyppeihin. Tiedontarve ei voi sinänsä olla spesifi relevanssin laji tai piirre, sillä relevanssi on niiden dokumenttien ominaisuus, joilla tiedontarvetta tyydytetään.

2.3 Tiedonhaun evaluointitutkimus

Tiedonhaun *evaluointitutkimuksessa* arvioidaan hakujen onnistumista mittaamalla hakutuloksena löytyneiden dokumenttien relevanssia suhteessa saantikantaan. Hakusuoritusta voidaan mitata saantiin ja tarkkuuteen perustuvalla hakuaiheeseen liittyvällä *keskitarkkuusarvolla*, joka on haun tuloksellisuuden mittari. Täydellisen hakutuloksen saavuttaminen suuresta kokotekstitietokannasta on utopiaa, koska luonnollinen kieli on epätäsmällistä eikä hakulausekkeessa käytetyillä hakuavaimilla pystytä kattamaan tietokannan kaikkien relevanttien dokumenttien sisältöjä. Kun hakulausekkeen käsitteiden ja dokumenttien käsitteellisten esitysten välinen vastaavuus on puutteellinen, jää moni relevantti dokumentti löytämättä. Kyse on sekä haun suorittajan osaamisesta että indeksointikieleen liittyvistä ongelmista. Optimaalinen hakutulos edellyttää hakulausekkeessa käytettävien käsitteiden ja dokumenteissa esiintyvien käsitteiden sisällöllistä vastaavuutta. Hakujärjestelmien kehittämisen näkökulmasta on tärkeää hakujärjestelmän piirteiden ja apuvälineiden vaikutus relevanttien dokumenttien löytyvyyteen. Näitä apuvälineitä ovat muun muassa automaattinen tesaurus ja relevanssipalaute. Kun hakulausekkeet voidaan kirjata muistiin ja toistaa, on eri hakujärjestelmien ja niiden vaikuttavien piirteiden vertaaminen myös mahdollista.

2.4 Hakusuunnitelma

Täystäsmäytyksessä *hakusuunnitelman* arviointi tapahtuu *tyhjentävyyden*, *tarkkuuden* ja *kattavuuden* perusteella. Tyhjentävyydellä tarkoitetaan hakusuunnitelmaan sisältyvien hakuaihetta jäsentävien aspektien määrää. Täysin tyhjentävään hakusuunnitelmaan on käytetty hakuaiheen kaikkia aspekteja. Tyhjentävyyden käsite liittyy hakuaiheen aspekteja rajaaviin suhteisiin. Hakusuunnitelman tarkkuus liittyy käsitteiden hierarkkisiin suhteisiin. Kyse on hakusuunnitelman aspekteja kuvaavien käsitteiden täsmällisyydestä. Hakusuunnitelman kattavuus viittaa hakulausekkeessa tiedontarpeen ilmaisuun käytettyjä käsitteitä edustavien hakuavainten määrään. Täystäsmäytyksessä hakuavaimet kattavat hakuaihetta koskevan tiedon käsitteellistä aluetta. Hakusuunnitelma on sitä kattavampi, mitä useammilla käsitteillä hakusuunnitelman eri aspektien ulottuvuuksia kyetään kuvaamaan. Hakusuunnitelman kattavuus liittyy hakuaiheen eri aspektien välisiin ja niiden sisäisiin rinnakkaisiin

suhteisiin. Tyhjentävyys, tarkkuus ja kattavuus vaikuttavat hakujen saantiin ja tarkkuuteen. Kun hakusuunnitelmaan lisätään aspekteja, haun tyhjentävyys kohoaa. Tällöin saanti pienenee, tarkkuus kasvaa ja tulosjoukon koko supistuu. Tarkkuus saadaan paranemaan käyttämällä tarkempia käsitteitä, jolloin saanti pienenee, tarkkuus lisääntyy ja tulosjoukon koko supistuu. Hakusuunnitelmasta tulee kattavampi, kun hakua laajennetaan lisäämällä siihen käsitettä kuvaavia hakuavaimia. Tällöin saanti kohoaa, tarkkuus alenee ja tulosjoukon koko kasvaa. Hyvän saannin edellytys on hakusuunnitelman kattavuuden painottaminen, jolloin haussa käytettävien käsitteiden liiallisesta tarkkuudesta on tingittävä. Hyvä tarkkuus puolestaan edellyttää hakusuunnitelman käsitteiden tarkkuutta. Tällöin toimiva hakusuunnitelma voi olla tyhjentävä, muttei kattava. (Järvelin & Kekäläinen, 2002.)

Osittaistäsmäyttävässä hakusuunnitelmassa hakuavaimina käytetyillä käsitteillä ei ole rinnakkaisia tai rajaavia suhteita. Hakulausekkeeseen voidaan liittää hakuavaimiksi käsitteitä, jotka kuvaavat tiedontarpeen eri ulottuvuuksia. Hakufasetin lisääminen ei osittaistäsmäytyksessä pienennä tulosjoukkoa, mutta voi parantaa hakusuunnitelman tarkkuutta. (Järvelin & Kekäläinen, 2002.)

2.5 Hakustrategia

Hakustrategialla tarkoitetaan hakulausekkeen muotoilemista, hakusuunnitelmaa kokonaisuudessaan tai tapaa lähestyä haun suorittamista. Tunnettuja hakustrategian tyyppejä ovat pikahaku, lohkostrategia, helmenkasvatusstrategia, vuorovaikutteinen selailu sekä erilaiset fasetointiin perustuvat strategiat. *Fasettiperustaiset* strategiat ovat *peräkkäisten* fasettien strategia, *spesifein* fasetti *ensin* -strategia ja *pareittain yhdistettyjen* fasettien strategia. Strategioita on kehitetty ja käytetty ennen kaikkea täystäsmäytykseen perustuvissa hauissa. Vaikka täys- ja osittaistäsmäytyksen hakusyntaksi on erilainen, myös osittaistäsmäyttäviä hakuja voidaan jäsentää samaan tapaan hakuaiheen käsiterakennetta hyödyntämällä. Vaikutukset hakutuloksiin ovat erilaiset.

Peräkkäisten fasettien strategiassa haku aloitetaan suuren tulosjoukon tuottavalla fasetilla, minkä jälkeen hakua tarkennetaan rajaavilla faseteilla. Spesifeintä fasettia käytettäessä liian laajaa hakutulosta voidaan supistaa lisäämällä hakulausekkeeseen vähemmän spesifejä fasetteja. Kun fasetteja yhdistetään täystäsmäytyksessä käyttämällä haussa pareittain yhdistettyjen fasettien strategiaa, niin vähintään kahden hakukäsitteen on esiinnyttävä saaduissa hakutuloksissa. Jos kaikki fasetit yhdistettäisiin täystäsmäytyksessä samaan tiedonhakuun, haku muuttuisi liian spesifiksi, mikä johtaisi nollatulokseen. Tiedonhaun kuluessa tapahtuvat *siirrot*, joita käytetään haun toteuttamiseen, ovat *hakutaktiikaksi* nimitetty osa tiedonhaun strategiaa. Batesin (1987) jaottelussa tiedonhaun aloitusvaiheen taktiikoita ovat mm. hyvän hakulausekkeen säilyttäminen, hakuavainten etsiminen, uusien hakuavainten etsiminen löydetyn informaation pohjalta sekä laajempien, suppeampien tai rinnakkaisten hakuavainten käyttö. Tiedonhaun edetessä käytettäviä taktiikoita voivat olla käsitteiden lisääminen, vähentäminen tai poistaminen hakulausekkeesta, käsitteen laajentaminen synonyymien avulla tai haun tarkentaminen karsimalla rinnakkaisia hakuavaimia. (Alaterä & Halttunen 2002, 86 - 89.)

2.6 Hakulauseke

Hakulausekkeet jaotellaan rakenteellisten ominaisuuksiensa perusteella. Hakulausekkeesta käytetään kirjallisuudessa usein myös käsitettä hakukysely. Täystäsmäyttävässä haussa käytetään vahvoja *rakenteisia* hakulausekkeita, jotka noudattavat Boolen logiikkaa ja edellyttävät hakuavainten ja niitä yhdistävien operaattoreiden sekä dokumentin sisällön täydellistä täsmäävyyttä. Hakuoperaattoreiden käyttö ei ole välttämätöntä osittaistäsmäytyksessä. Osittaistäsmäyttävässä haussa voidaan käyttää joko *heikkoa*, *rakenteetonta* sanalistaa tai *vahvaa*, *rakenteista* hakulauseketta, joka perustuu fasettien käyttöön hakuavaimina. Osittaistäsmäyttävän vahvan hakulausekkeen hakutehoa lisäävät haun painottaminen relevanssipalautteen pohjalta sekä sanaliittojen käyttö hakuavaimina. Osittaistäsmäytyksessä voidaan hyödyntää samoja hakuavaimia kuin täystäsmäyttävissä hakulausekkeissa. Hakija valitsee osittaistäsmäytyksessä hakujärjestelmästä itse, kuinka monta tulosta haluaa tulosjoukkoon näkyviin hakuprosessinsa edetessä. Hakutulokseen valikoituvat myös hakulausekkeeseen osittain täsmäävät dokumentit. Hakutuloksen järjestys perustuu osittaistäsmäytyksessä

hakulausekkeen ja dokumentin täsmäävyyden asteeseen. Tulos lajittuu todennäköisen relevanssin mukaan laskevaan järjestykseen. (Järvelin & Kekäläinen, 2002.)

2.7 Hakutulos

Hakutulosta voidaan arvioida *saannin* ja *tarkkuuden*. Mainitut käsitteet perustuvat relevanssiin. Kokotekstitietokantaan voidaan luoda tiedonhaun evaluointia varten asiantuntijoiden suorittamien relevanssiarvioiden perusteella *saantikanta*, jonka avulla voidaan määritellä hakujen tuloksellisuus. Dokumenttitiedonhaun tehokkuutta mitataan saantia ja tarkkuutta ilmaisevilla tuloksellisuuden mittareilla. Saantiluvut ilmaisevat, millaisen osuuden hakujärjestelmä löytää tietokannan kaikista relevanteista dokumenteista. Tarkkuusluvut ilmaisevat hakujärjestelmän kykyä löytää tietokannasta vain siihen sisällytetyt relevantit dokumentit. Saanti voidaan käsittää relevantin dokumentin löytymisen todennäköisyydeksi, hakutuloksen tarkkuus taas voidaan tulkita löytyvän dokumentin relevanssin todennäköisyydeksi. Suuriin saantilukuihin päästään kattavalla haulla, joka ei ole liian tarkka. Hakujen tarkkuus kasvaa, kun haussa käytetyt käsitteet ovat tarkkoja.

Haku on toimiva, jos siinä on onnistuttu tavoittamaan aihevastaavuudeltaan hyvät dokumentit. Aihevastaavuudesta puhuttaessa käytetään termiä *aiherelevanssi*. Loppukäyttäjä päättää todellisessa hakutilanteessa, ovatko haussa löytyneet dokumentit hyödyllisiä hänen tiedontarpeensa näkökulmasta. *Käyttäjärelevanssi* määrittyy suhteessa loppukäyttäjän tyydyttyneeseen tai vastausta vaille jääneeseen tiedontarpeeseen. Haun saannin ja tarkkuuden laskemista varten on erityiset laskentakaavat. *Suhteellinen saanti* saadaan, kun verrataan haussa saatujen relevanttien dokumenttien määrää saantikannan kaikkien tehtäväkohtaisten relevanttien dokumenttien määrään. *Absoluuttisen saannin* edellytyksenä on, että saantikanta on tehty arvioimalla kaikki tietokannan dokumentit suhteessa hakuaiheeseen. Kun haussa saatujen relevanttien dokumenttien määrä jaetaan tällaisen saantikannan [relevanttien] dokumenttien määrällä, saadaan absoluuttinen saanti. Täydelliseen hakutulokseen – sataprosenttiseen saantiin – on mahdollista päästä vain pienissä dokumenttikokoelmissa. Sen käyttö kriteerinä saantia arvioitaessa ei tutkimuksen kannalta ole mielekäs, koska tietokantoihin sisällytettyjen dokumenttien määrä on yleensä suuri. (Alaterä & Halttunen 2002, 127 - 129.)

3. Tiedonhakupelin soveltaminen tutkimus- ja opetuskäyttöön

Haun aiheen täsmentäminen, hakukäsitteiden tunnistaminen ja hakulausekkeiden muotoilu edesauttavat tekstitietokannan relevanttien dokumenttien löytymistä. Mikäli hakujärjestelmä antaa ohjaavaa palautetta hakuprosessin edetessä, hakija pystyy hakulauseketta muuntelemalla parantamaan haun tuloksellisuutta. Hakujärjestelmän ja käyttäjän välinen vuorovaikutus toimii. Tiedonhaku sinänsä on heuristinen prosessi. Kun hakija sen lisäksi, että tuntee hakumenetelmiä, myös hyödyntää omaa kekseliäisyyttään ongelmanratkaisussa, hän voi löytää hakutulosta parantavan menetelmän arvioidessaan yrittämällä ja erehtymällä hakuprosessinsa etenemistä.

Tiedonhakupeli on opetuskäyttöön laadittu web-pohjainen sovellus, joka kehitettiin Tampereen yliopiston informaatiotutkimuksen laitoksella alun perin tiedonhaun tutkimusta ja sen ilmiöiden analysointia silmällä pitäen. Tiedonhakupeli on ollut käytössä 1998 lähtien. Tiedonhakupelin tarkoitus on tukea ja edistää tiedonhaun menetelmien omaksumista. Uusien sovellusten kehittämisessä on huomioitu sekä pedagoginen että käyttäjänäkökulma. Käyttäjän näkökulmasta on tärkeää tiedonhakupelin ohjeiden selkeys, helppokäyttöisyys ja hakujen onnistumista arvioiva palautejärjestelmä. Tiedonhakupelin käyttöä on perusteltu sillä, että se simuloi todellisia hakuolosuhteita. Pelaaminen ei ole itsetarkoitus. Tärkeämpää on tutustuttaa käyttäjä tiedonhaun menetelmiin. Tiedonhakupeli antaa palautteen harjoitushakujen onnistumisesta ja mahdollistaa hakulausekkeen muotoilun uudelleen, kun tavoitteena on hakutuloksen parantaminen. Käyttäjä tavallaan oppii yrittämisen ja erehtymisen menetelmällä. Tiedonhakupelin interaktiivisuudella tarkoitetaan hakujärjestelmän, tekstitietokantojen ja hakua suorittavan ihmisen välistä vuorovaikutusta. Tiedonhakupelin opetussovelluksia on tutkittu, niistä on julkaistu artikkeleita ja tehty väitöskirja (Halttunen 2004) sekä pro graduja (Makkonen 2002; Laakkonen 2003; Pennanen 2003). Tutkimuksessa on keskitytty tiedonhakupelin opetussovellusten ohella myös käyttäjävirheiden analyysiin.

Tiedonhakupelissä käyttäjä muotoilee itse hakulausekkeen, tarkistaa, että muotoilu on hakukielen syntaksin mukainen ja toteuttaa haun. Hakulauseke ohjautuu hakujärjestelmään, järjestelmä käy tulosjoukkoa läpi, vertaa löytyneitä dokumentteja harjoitustietokannan kaikkiin relevantteihin aihedokumentteihin sekä arvioi suoritetun

haun saannin ja tarkkuuden. Hakujärjestelmä kirjaa automaattisesti tiedot käyttötapahtumista lokitiedostoon, josta tapahtumatiedot ovat analysointia varten saatavilla. Kyse on hakuprosessin ohella myös hakujärjestelmän suorituskyvystä. Dokumenttien automaattisessa indeksoinnissa käytetyt hakuavaimet vaikuttavat osaltaan haun lopputulokseen. Kun luonnollinen kieli on monitulkintaista ja siihen sisältyy runsaasti sanojen johdoksia, taivutusmuotoja ja yhdyssanoja, niin tiedonhaun käsitteellinen suunnittelu hankaloituu.

Tiedonhakupeli voidaan ymmärtää kielipeliksi, jonka siirrot tapahtuvat hakuavainten avulla. Pelissä on logiikka, strategia ja säännöt, joiden mukaan edetä. Wittgenstein näkee kielipelin yleensä välineenä, jonka avulla voidaan tutkia kieltä yksinkertaistetussa tilanteessa. Kielessä merkitykset sisäistetään asioita ja ilmiöitä nimeämällä ja sijoittamalla nimetty merkityskontekstiinsa. Koska merkitykset muodostuvat ja sisäistetään kielenkäytön konteksteissa, eri tilanteisiin soveltuvat eri kielipelit. Kielipelin lausemuotona voi olla komento, käsky, kiitos tai tervehdys, joka sisältää merkityksen. Ymmärtäminen edellyttää sekä kykyä käyttää kieltä että tunnistaa kielen erilaisia käyttöyhteyksiä, sillä käsitteillä on myös käyttöyhteytensä mukaan määrittyneitä merkityksiä. Wittgensteinin ajattelussa kielipeli on primitiivinen kielilaji, joka mahdollistaa asioiden yksinkertaistetun ja yhdenmukaisen esittämisen. (von Wright 1982, 233 - 238.) Kielellisten ilmausten käyttöä ja kielenkäyttäjän tekoja maailmassa kutsutaan kielipelin siirroiksi. Tiedonhaussa kielen monimerkityksisyys aiheuttaa ongelmia sekä käsitteiden sisällöllisen määrittelyn että hakutuloksen relevanssin näkökulmasta.

Tiedonhakupeli on työkalu, jonka avulla hakutulosta voidaan arvioida ja hakulauseketta muokata edelleen optimaalisen hakutuloksen saavuttamiseksi. Hakija saa hakusuorituksesta välittömän palautteen; saanti ja tarkkuus tulevat näytölle saantipiiraana ja tarkkuuskäyrinä. Haun suorittaja voi tutkia hakulausekkeen muuntelun vaikutuksia hakutuloksiin. Tiedot suoritetuista hauista tallentuvat ohjelmaan automaattisesti. Tiedonhakupelin palautetoimintojen avulla hakija pystyy tarkistamaan onnistuneimpien hakujensa tarkkuuden suhteessa suurempaan hakutulosjoukkoon. Tiedonhakupelin toiminta perustuu paljolti käyttäjän intuitioon. Valittuaan haun aiheen, tietokannan ja käytettävän tiedonhakujärjestelmän käyttäjä siirtyy hakuohjelmassa muotoilemaan hakulauseketta ja syöttää haun kohteena

olevaan hakujärjestelmään. Hakujärjestelmä prosessoi hakulausekkeen, poimii siihen täsmäävät dokumentit ja lataa haun tulokset sekä saantia ja tarkkuutta ilmaisevat luvut käyttäjän nähtäväksi näyttöruudulle. Saanti- ja tarkkuuslukujen ohella näkyvissä ovat myös aiemmin suoritettujen hakujen parhaat tulokset, joten käyttäjä voi seurata suoraan hakutulostensa kehittymistä. Tiedonhakupeliin on sisällytetty kunniataulu, jonne pääseminen ilmaisee haun erinomaista onnistumista. Kunniataulun tarkoitus on motivoida käyttäjää kehittymään tiedon hakijana. Ohjelma tallentaa kaikkien hakujen lokitiedot, jotka näin ovat käytettävissä myöhempää tarkastelua varten.

Tiedonhakupelissä haut tehdään Alma Median lehtiartikkeleita sisältävästä rakenteettomasta kokotekstitietokannasta. Tiedonhaussa käytetään sekä TRIP- että InQuery-hakujärjestelmää, joten tiedonhakupelin molemmilla hakujärjestelmillä saatujen hakutulosten vertailulle on hyvät lähtökohdat. Pelistä on kehitetty useita uusia versioita, joiden pohjalta suunnitellaan tiedonhaun itsenäisen verkko-opiskelun mahdollistavaa versiota. Tiedonhakupelin prototyypissä pelin aloitussivulla on hakutehtävän kuvaus, hakukenttä ja linkitys hakukielen oppaaseen. Tiedonhakupeli näyttää hakujen saannin ja tarkkuuden automaattisesti.

Tiedonhakututkimuksessa on keskitytty yleisesti hakumenetelmien keskimääräisen tuloksellisuuden mittaamiseen, mutta yksittäistenkin hakujen analysointi voi tuoda tutkimukseen oman kontribuutionsa. Hakujen analysointiin soveltuva tiedonhakupeli on tiedonhaun tutkimus- ja opetustoimintaa yhdistävä työväline. Tiedonhakupelin komponentit ovat: sarja hakuaiheita, mahdollisuus hakuavainten ideointiin edellisten hakujen tulosdokumenttien pohjalta, sekä hakijan suoriutumista hakutehtävästä mittaava ja visualisoiva moduuli.

3.1 Tiedonhakupelin hakujärjestelmät

TRIP-tiedonhallintajärjestelmä on TietoEnatorin tekstitiedonhallintaa varten kehittämä ohjelmisto, joka on myös integroitavissa muihin ohjelmistoihin. Siihen sisältyy web-pohjainen tiedonhaun sovellus, joka perustuu Boolen täystäsmäyttävään tiedonhakumalliin. Boolen logiikan mukaisesti muodostetussa hakulausekkeessa voidaan käyttää hakuavaimia, Boolen operaattoreita, sulkumerkkejä ja läheisyysoperaattoreita. Kun hakutulokseen saadaan vain hakulausekkeeseen

täydellisesti täsmääviä dokumentteja, on virheiden mahdollisuus jokseenkin suuri. Tiedonhakupelissä on hyödynnetty TRIP-hakujärjestelmää ja Alma Median lehtiartikkelitietokantaa.

TRIP-hakujärjestelmän tavoin myös osittaistäsmäyttävä InQuery-hakujärjestelmä toimii vuorovaikutteisesti hakijan ja hakujärjestelmän välillä. Hakulausekkeen muotoilussa voi käyttää luonnollista kieltä tai kyselykieltä. Vuorovaikutteisen hakujärjestelmän etu on siinä, että hakulauseketta voi muuntaa ja toistaa. Tiedonhakupelissä hakutulokset käsittelevä arviointiohjelma tuo näkyviin tulosten saannin ja tarkkuuden, mikä helpottaa hakutuloksen kokonaisarvioinnin suorittamista. Hakujärjestelmästä tehdyt haut ovat joko luonnollisen kielen hakuja tai rakenteisia hakuja. Pystyäkseen käsittelemään luonnollisella kielellä tehdyn hakulausekkeen hakuohjelma muuntaa sen rakenteiseen muotoon. Hakulausekkeen syöttäminen suoraan hakujärjestelmään rakenteisessa muodossa auttaa hakijaa määrittelemään hakulausekkeessa käytettyjen hakuavainten väliset suhteet, mikä saattaa johtaa parempaan hakutulokseen. Rakenteisen hakulausekkeen muotoilu edellyttää rakenteisen kyselykielen ja sen hakuoperaattorien hallintaa.

InQuery-hakujärjestelmä tunnistaa summaoperaattorit (myös painotetut hakuoperaattorit), läheisyysoperaattorit, yhdistävät, rajaavat ja kieltävät Boolen operaattorit sekä fraasihakuun, kappalehakuun ja synonyymihakuun sopivat hakuoperaattorit. Monipuolisia hakuoperaattoreita käytettäessä hakulausekkeen teho paranee. Hakujärjestelmän parametreja vastaavat operaattorit tukevat, rajaavat ja täsmentävät perushakua. Tiedonhakupelin hakujärjestelmien aihevastaavuudeltaan parhaita dokumentteja seuraavien dokumenttien ryhmät ovat tutkimuksen kannalta kiinnostavia.

3.2 Lokitiedostojen käyttö tiedonhakututkimuksessa

Hakuharjoituksissa suoritetut haut tallentuvat tiedonhakupelin lokitiedostoon, josta ne voidaan poimia tarkasteltavaksi eri perustein. Lokitiedosto on toisin sanoen tapahtumatiedosto, jonka pohjalta pystytään selvittämään sekä osittais- että täystäsmäyttävän hakujärjestelmän käyttöä ja eri hakujärjestelmistä tehtyjen hakujen onnistumista käyttäjä- ja hakuaihekohtaisesti. Lokitiedoista näkyy hakuyrityksen

numero, käyttäjätunniste, hakuaihe ja hakujärjestelmä. Hakuloki näyttää myös haun keskitarkkuusarvon, tulosjoukon koon ja jokaisen hakujärjestelmään syötetyn hakulausekkeen muodon. Hakulausekkeen kehittyminen sekä pyrkimys parempaan ja kattavampaan hakutulokseen voidaan jäljittää hakulokeista käyttäjittäin. Vaikka hakujen aloitusohjeet ovat tiedonhakupelissä samat kaikille, hakijakohtaiset erot tulevat esiin hakustrategian ja hakuavainten valinnassa sitä mukaa, kun hakulausekkeet monimutkaistuvat. Tiedonhakupelin lokitiedoista käy ilmi yksittäisen käyttäjän muodostamien hakulausekkeiden paremmuusjärjestys. Toimivia hakustrategioita on useita ja hyviä hakutuloksia saavutetaan monin eri tavoin muotoilluilla hakulausekkeilla.

Hakujen lokitiedostoja voidaan hyödyntää paitsi kyselykeskeisessä kokeellisessa tiedonhakututkimuksessa, myös tutkittaessa web-hakujärjestelmien käyttöä ja käyttäjien tiedonhakukäyttäytymistä tiedontarpeiden näkökulmasta. Kokeellisen tiedonhakututkimuksen painopistealueisiin kuuluu toisaalta tiedonhaun oppimisen prosessi, toisaalta taas tiedonhakujärjestelmien kehittäminen ja niiden käyttäjäystävällisyyden parantaminen. Informaatiotutkimuksen perusopintojen tiedonhakukurssin harjoitusten lokitiedot ja opiskelijoiden harjoitusten yhteydessä kirjoittamat esseevastaukset kertovat, kuinka opiskelijat ovat hahmottaneet hakutehtävät ja mitä he ovat ajatelleet tai erityisesti pohtineet pyrkiessään mahdollisimman hyvään lopputulokseen tiedonhakupelin aihehauissa.

4. Tekstitiedonhaun tutkimustraditio

Tiedonhaun tutkimuksen yhteydessä käytetään käsitteitä *kyselykeskeinen* ja *käyttäjäkeskeinen* tutkimus. Tiedonhaun kysely- ja käyttäjäkeskeisten kokeellisten tutkimusten päämääränä on mitata erilaisten hakujärjestelmien suorituskykyä, tehokkuutta ja tuloksellisuutta. Hakujärjestelmät hakuominaisuuksineen, hakuanalyysin menetelmät sekä harjoitus- ja testitietokantojen käyttö edustavat tiedonhakututkimuksen tyypillisiä koeasetelmia. Optimaalisten hakujen vertailu valottaa erilaisten hakulausekkeiden käyttäytymistä hakujärjestelmässä. Hakujärjestelmien toimivuutta ja tehokkuutta tekstitiedonhaussa on tutkittu paljon. Tavoitteena on yleisesti ollut hakujärjestelmien kehittäminen ja hakutulosten

tarkkuuden parantaminen empiirisesti tutkitun tiedon avulla. Hakutulokseen vaikuttavia tekijöitä, koejärjestelyitä ja harjoitustietokantoja voidaan käsitellä ja muunnella monin tavoin. Millainen on toimiva hakulauseke? Miten hakulausekkeen hakuavainten muotoilun, lausekkeen rajaamisen tai laajentamisen avulla voidaan vaikuttaa hakutehoon? Tekstitiedonhaussa käytetään vuorovaikutteisia menetelmiä ja palautejärjestelmiä. Tekstitiedonhaun varhaisessa tutkimuksessa käytettiin pieniä dokumenttitietokantoja tiedonhaun tekniikoiden kokeilemiseen. Tiedonhaun evaluointitutkimuksessa arvioidaan hakujen onnistumista ja hakujärjestelmän tehokkuutta tietokannasta löytyvien dokumenttien aihevastaavuuden, tarkkuuden ja saannin perusteella. Vertaileva arviointi perustuu teorian ja käytännön yhteensovittamiseen, kun tavoitteena on parempien hakutulosten saavuttaminen.

Hakujärjestelmien tehokkuus ja hakujen tuloksellisuus ovat kriteerit, joiden pohjalta järjestelmäsovelluksia on pyritty kehittämään. Salton (1970; 1973; 1986) on tekstitiedonhaun tutkimuksen uranuurtaja, jonka tiedonhaun tekniikoihin liittyviin kokeellisiin tutkimuksiin on usein viitattu. Salton tutki automaattisen tekstiperustaisen hakujärjestelmän toimivuutta pienissä dokumenttitietokannoissa, muttei ottanut tutkimuksissaan huomioon suurten tekstimäärien hallintaan liittyviä ongelmia. Saltonin mukaan automaattinen tekstiperustainen hakujärjestelmä oli joka tapauksessa kilpailukykyinen verrattaessa sitä *intellektuaaliseen* indeksointiin perustuvaan hakujärjestelmään. Intellektuaaliseen indeksointiin perustuva hakujärjestelmä pystyy vastaamaan hakulausekkeeseen, jonka muotoilussa on käytetty vain dokumenttien asiasanoituksessa esiintyviä hakuavaimia. Blair ja Maron (1990) pyrkivät osoittamaan Saltonin tekstitiedonhakua koskevien argumenttien heikot kohdat omilla vasta-argumenteillaan. He esittivät perusteluissaan, että aiherelevanssiltaan hyvissä dokumenteissa mahdollisesti käytetyt luonnollisen kielen sanat tai fraasit ovat hankalasti ennakoitavissa, koska luonnollisen kielen vaihtelu on suurta. Suurista tekstitietokannoista löytyy irrelevantteja dokumentteja, joissa hakuavain esiintyy, vaikka ei täsmääkään hakuaiheeseen. Blair ja Maron (1990, 439) näkevät liian laajan hakutulosjoukon suurten tekstitietokantojen ongelmana. Standardoidun hakustrategian tyypillisiin piirteisiin kuuluu rajatun hakutuloksen tuottaminen tekstitiedonhaussa: dokumentteja löytyy vähemmän, joskin hakutulos tarkentuu. Sormunen puhuu artikkelissaan (2001) täystäsmäyttävien hakujen ankkuritermien käyttöön liittyvästä erityisestä piirteestä suurissa tekstitietokannoissa. Ankkuritermien vaikutus ilmenee

tiedonhakijan pyrkimyksenä pitäytyä hakuprosessin aikana alkuperäisissä hakuavaimissa, vaikka se johtaa tiukasti fokusoituihin hakuihin ja pienentää saantia. Sormunen huomauttaa, ettei Blairin ja Maronin (1985; 1990) tutkimuksissa ole tarkasteltu täystäsmäyttävää hakujärjestelmää ja Boolen hakuoperaattoreiden toimimista siinä järjestelmälähtöisestä näkökulmasta. Järvelin ja Kekäläinen (2002) puhuvat tekstitiedonhaun yhteydessä sumeista hauista. Tekstitiedonhaku on sumeaa, koska hakulausekkeen muotoilussa on vaikea tietää, millaista sanastoa jonkin aiheen kuvaamiseen on käytetty dokumenttien teksteissä.

Hakujärjestelmästä voidaan käyttää *kielipeli* -metaforaa: hakujärjestelmä näyttäytyy tällöin erilaisten argumenttien elinympäristönä, jossa pätevät omat säännöt. Blair ja Maron (1985) ovat huomauttaneet, ettei tekstitiedonhaun kontekstissa ole helppoa löytää läheskään kaikkia aihevastaavuudeltaan relevantteja kokotekstitietokantaan sisällytettyjä dokumentteja. Koko hakujärjestelmä on tavallaan perustunut otaksumalle, että tiedonhakija pystyy ennakoimaan täsmälleen ne hakusanat ja fraasit, joita dokumenteissa esiintyy ja löytää sen vuoksi vain sisällöltään relevantit dokumentit. Aikaisemmassa tutkimuksessa on todettu, että tekstitiedonhaku tuottaa hyvän hakutuloksen vain pienissä tietokannoissa. Hakujärjestelmä ei indeksointikielen toimivuudesta ja luonnollisen kielen käsittelykyvystä huolimatta pysty tunnistamaan tietokannan kaikkia relevantteja dokumentteja, kun tietokanta on laaja. Laaja hakutulos ei ole tarkka, koska siihen sisältyy hälyä – aihevastaavuudeltaan huonoja osumia. Epätarkan ja liian lavean hakutuloksen yhteydessä puhutaan tulokseen liittyvästä redundanssista. Vasta tiedon loppukäyttäjä tunnistaa tulosjoukosta itselleen tärkeää informaatiota sisältävät dokumentit.

4.1 Tiedonhakujärjestelmän tehokkuuden arviointi

F. W. Lancasterin, Richard L. Rapportin ja J. Kiffin Penryn EARS-tutkimus (1972) on esimerkki luonnolliseen kieleen perustuvan online-tiedonhakujärjestelmän tehokkuuden arvioinnista. Tutkimuksessa suoritettiin hakuavain- ja vapaatekstihakuja lääketieteellisiä tiivistelmiä sisältävästä tietokannasta. Koehenkilöt olivat erikoisalan asiantuntijoita. Hakutulosten perusteella tehtiin saantia ja tarkkuutta koskevat arviot. Samalla arvioitiin hakujärjestelmän käyttäjien tyytyväisyyttä saatuihin tuloksiin ja analysoitiin hakujen onnistumiseen tai epäonnistumiseen vaikuttaneita tekijöitä.

EARS-tutkimus on koeasetelmaltaan klassinen ja metodologialtaan tyypillinen tehokkuuden arviointitutkimuksen esimerkki, jossa huomio kiinnitettiin myös käyttäjän ja hakujärjestelmän väliseen vuorovaikutukseen.

David C. Blair ja M.E. Maron käsittelivät IBM:n STAIRS-tekstitiedonhakujärjestelmään liittyvässä tutkimuksessaan (1985) dokumenttien löytyvyyttä suuren dokumenttikokoelman rajatulta osa-alueelta. Tutkimuskohteena oli erikoisalan dokumentteja sisältävä kokoelma, jota asiantuntijat hyödynsivät työssään. Blair ja Maron painottivat relevanttien dokumenttien linkittämisen tärkeyttä hakujärjestelmän kehittämistyössä. IBM:n STAIRS (Storage And Information Retrieval System) on kokotekstidokumentteja sisältävä nopea ja tiedonkäsittelyn kapasiteetiltaan suuri hakujärjestelmä. Siinä oli tutkimusajankohtana noin 350 000 sivua erilaisia asiakirjoja, muistioita, pöytäkirjoja ja raportteja. STAIRS-hakujärjestelmän tehokkuutta tutkineessa Blairin ja Maronin arvioinnissa (1985) pyrittiin mittaamaan tiedonhaussa löytyneiden dokumenttien hyödyllisyyttä noudattamalla vuorovaikutteisen tiedonhaun periaatteita. Hakujen saannin ja tarkkuuden mittauksessa tavoiteltiin mahdollisimman suurta objektiivisuutta.

Varhaiset tutkijat Swanson (1960) ja Salton (1970) olivat optimistisia kokotekstitiedonhaun tuloksellisuuden suhteen. Blair ja Maron (1985) pohtivat, millä periaatteella dokumenttikokoelma tulisi järjestää, jotta yksittäisten dokumenttien löytyvyyttä kokotekstitietokannasta voitaisiin parantaa. Automaattisen tekstitiedonhaun perusidea oli, että hakujärjestelmän jokainen dokumentti pystyttäisiin löytämään tietokantaan tallennettujen dokumenttien joukosta dokumentissa esiintyvien yksittäisten sanojen perusteella. Automaattiseen indeksointiin perustuvan hakujärjestelmän oli ajateltu tekevän indeksointia suorittavan ihmisen työpanoksen tarpeettomaksi. Blair ja Maron totesivat omassa tutkimuksessaan STAIRSin hakutehon huonoksi. Tenopirin (1985) tutkimuksessa käytettiin yhtä hakuaihetta, josta tehdyt haut olivat kokoteksti- ja tiivistelmähakuja. Sen ohella hakulausekkeen muotoilussa käytettiin kontrolloitua hakusanastoa ja otsikkohakua. Tenopirin tutkimuksessa hakutuloksen aihevastaavuuden arvioinnin suorittivat asiantuntijat. Tietokannan relevanteista dokumenteista jäi suuri osa löytämättä. STAIRS-hakujärjestelmän tehokkuustutkimuksessa käytetyt vuorovaikutteisen tiedonhaun menetelmät ovat perusta eri hakujärjestelmien keskinäiselle vertailulle arvioitaessa

relevanttien dokumenttien löytymistä – saantia. Tiedonhaun tutkimuskirjallisuudessa on usein viitattu juuri STAIRS-tutkimukseen, jossa on kyse dokumenttien intellektuaalisen sisällön ja hakukielen syntaksin sekä hakulausekkeen sisällön vastaavuudesta.

David C. Blair puhuu artikkelissaan (1996) tiedonhaun tehokkuudesta ja dokumenttien löytyvyyden ongelmista. Hakijan hakutuloksen relevanssia koskevassa omassa arviossa on subjektiivisuutta ja epävarmuutta aiheuttavia tekijöitä. Milloin hakija lopettaa? Milloin tulos on riittävä? Mistä hyödyllisiä dokumentteja kannattaa etsiä? Miten relevanssia tulee arvioida? Missä vaiheessa arviointi on syytä lopettaa? Millaista metodologiaa voidaan soveltaa tietokannan relevanttien löytymättömien dokumenttien etsimiseen? Jos tiedonhaun saantia ja tarkkuutta mittaavan tutkimuksen koeasetelmaan sisältyy runsaasti epävarmuutta aiheuttavia tekijöitä, ei tutkimustakaan voida pitää luotettavana. Kun tietokanta sisältää runsaasti dokumentteja, niiden hakeminen edellyttää systemaattista hakustrategiaa. Hakustrategian systemaattisuus taas mahdollistaa eri hakujärjestelmistä saatujen tulosten vertailun. Arvioinnissa ja mahdollisessa hypoteesin testauksessa kannattaa kiinnittää huomio tuloksiin, joiden valossa hakujärjestelmän tehokkuus ei vastaakaan odotuksia. Blair käyttää artikkelissaan myös dokumenttitietokantoihin liittyvästä redundantista informaatiosta nimitystä häly. Hän toteaa hälyn alentavan haun tehokkuutta huomattavasti suurissa tekstitietokannoissa.

Sormunen on palannut artikkelissaan (2001) Blairin ja Maronin tutkimukseen, jossa tutkijat totesivat Boolen haut tehottomiksi suurissa kokotekstitietokannoissa. Blair ja Maron perustivat päätelmänsä hakujen suorittajien hakukäyttäytymiseen. Kun hakujen suorittajat pitäytyivät ennakoimissaan hyödyllisissä hakuavaimissa, haut jäivät kapea-alaisiksi ja tulosjoukko pieneksi. STAIRS-tutkimuksessa ei hyödynnetty täystäsmäyttävän hakujärjestelmän perustoimintoja. Sormusen mukaan STAIRS-tutkimuksen lähtökohta oli väärä. Sormunen tutki empiirisesti Boolen hakujen toimivuutta suurissa ja pienissä tietokannoissa lähtökohtanaan hakujärjestelmä. Hakujärjestelmään voidaan sisällyttää myös hakulausekkeen optimointialgoritmi. Koska Boolen haussa edellytetään tekstin ja hakuavaimen täydellistä täsmäävyyttä ja relevanttien dokumenttien pääkäsitteet ovat usein implisiittisiä, Sormunen on esittänyt ratkaisuksi fasettiperustaista osittaistäsmäyttävää hakumenetelmää. Tämä parantaa

hakutuloksen tarkkuutta tilanteessa, jossa täystäsmäyttävä haku tuottaa suuren, mutta epätarkan tulosjoukon. Tutkimuksen kysymyksenasettelun avulla selvitettiin Boolen hakujen tehokkuutta sekä suuressa että pienessä tietokannassa, kun hakutulosjoukko on suuri ja hakulausekkeet on muotoiltu tietokannan suhteen optimaalisesti. Edelleen tutkittiin, missä määrin optimaalisesti muotoillut hakulausekkeet olivat erilaisia suurissa tietokannoissa. Myös suuria ja pieniä tietokantoja varten optimoiduissa Boolen hauissa havaitut yhtäläisyyksiä ja eroja selittävät relevanttien dokumenttien piirteet tutkittiin suorittamalla tietokannan relevanttien dokumenttien fasettianalyysi. Kun täydellistä hakulauseketta ei ole, niin tietokannan kaikkien relevanttien dokumenttien löytyminen edellyttäisi useiden fasettiperustaisten hakujen suorittamista samasta hakuaiheesta (Sormunen, 2001). Haut eivät yleensä ole riittävän kattavia löytämään aihevastaavuudeltaan parhaita dokumentteja.

Fasettianalyysin löydöksiä sovellettiin sekä tarkkojen että epätarkkojen hakujen tuloksena löytyneiden dokumenttien vertailuun. Tuloksista tulkittiin optimoitujen hakulausekkeiden vaikuttavia rakenteellisia ominaisuuksia. Menettely tukee relevanttien dokumenttien hakuominaisuuksien vertailua. Hakutulokseen todettiin vaikuttavan paitsi tietokannan koon, myös tietokantaan sisältyvien relevanttien dokumenttien määrän. Mikäli dokumenteissa ei ollut eksplisiittisiä hakukäsitteitä, hakutuloksen tarkkuus oli vähäinen, vaikka tulosjoukko oli suuri. Pienistä tietokannoista suoritettujen hakujen tulokset olivat tarkempia. Hakutuloksen tarkkuuden todettiin laskevan huomattavasti sekä suurissa että pienissä tietokannoissa tulosjoukon koon kasvaessa. Mitä enemmän relevantteja dokumentteja löytyy, sitä jyrkemmin haun tuloksen tarkkuus laskee. Mikäli haussa saadun tulosjoukon koko on suuri, fasettiperustainen osittaistäsmäyttävä uusi haku voi tuottaa tarkemman tuloksen. Sormusen tutkimus vahvisti Blairin ja Maronin päätelmiä Boolen hakujen tehottomuudesta suurissa kokotekstitietokannoissa. Sormunen kuitenkin huomauttaa, että hänen tutkimuksensa havainnot perustuvat tiedonhakujärjestelmän teknisten ominaisuuksien analyysiin – Boolen hakujen täsmäyttämiseen tekstitietokannan sisältöä vastaavaksi. Tutkimustulos osoittaa täystäsmäyttävän hakumenetelmän toimivuuden rajat, vaikka Boolen hakujen avulla voidaankin esittää tiedontarpeen semanttiset aspektit monipuolisesti. Boolen haut perustuvat fasettirakenteeseen. Hakulausekkeen fasettirakenteen on todettu kohottavan tarkkuutta todennäköisyyteen perustuvissa hauissa ja hakujen laajennuksissa (Kekäläinen ja Järvelin 1998;

Sormunen et al. 2001) sekä sanakirjaperustaisessa kieltenvälisen tiedonhaun tutkimuksessa (Pirkola 1998). Mainituissa tutkimuksissa on verrattu rakenteettomia ja rakenteisia osittaistäsmäyttäviä hakuja ja osoitettu rakenteisten hakujen parempi toimivuus.

5. Hakulausekkeiden takautuva arviointi ja hakujen tuloksellisuus vuorovaikutteisissa hakujärjestelmissä

Referoin seuraavassa osuudessa Sormusen konferenssiartikkelia (2002), jossa käsitellään vuorovaikutteisen tiedonhaun takautuvan arvioinnin menetelmää, InQuery-hakujärjestelmän rakennetta ja toimintaa sekä eri tavoin muotoiltujen hakulausekkeiden optimaalisen tuloksellisuuden mitattavuutta koeolosuhteissa. Kokeellisessa tiedonhakututkimuksessa tiedonhakupeli toimii hakuanalyysissa välineenä, jonka avulla voidaan jäljittää ja tunnistaa sekä hakutulokseen vaikuttavia myönteisiä tekijöitä että virhelähteitä.

Täystäsmäyttävien ja osittaistäsmäyttävien hakulausekkeiden takautuvan eli retrospektiivisen arvioinnin menetelmä perustuu eri tiedonhakumallien pohjalta kokeellisesti suoritettujen hakujen tulosten vertailuun. Kokeessa keskityttiin vertailemaan Boolen operaattoreita käyttämällä saatujen täystäsmäyttävien sekä osittaistäsmäyttävien rakenteisten ja rakenteettomien hakujen optimaalista tuloksellisuutta. Kokeen tulos osoitti oikeiksi Boolen hakujen tarkkuudessa korkeiden saantilukujen kohdalla aiemmin tunnistetut ongelmat. Erityyppiset hakulausekkeet voidaan nähdä kokeellisen arvioinnin menetelmien haasteena. Riippuen tiedonhakumallista, johon hakujärjestelmä perustuu, hakulauseke voidaan ilmaista käyttämällä Boolen operaattoreita, luonnollisen kielen lauseita tai sanajoukkoja. Perinteinen koeasetelma määräytyy yleensä käytettävän tiedonhakumallin mukaisesti. Keskeinen ongelma on vertailullisesti sopivan mittausmenetelmän löytäminen. Boolen operaattoreita hyödyntäen muotoillun haun tuloksena löytyy hakulauseketta vastaava joukko järjestämättömiä dokumentteja. Suoritusta mitataan saantiin ja tarkkuuteen perustuvalla hakuaiheeseen liittyvällä *keskitarkkuusarvolla*. Osittaistäsmäyttävässä hakujärjestelmässä dokumentit järjestetään annetun jaotteluperiaatteen mukaiseen järjestykseen (*optimaalinen aihevastaavuus*). Tietokannan dokumenttien relevanssille

lasketaan todennäköisyydet ja lajitellaan tulosjoukko niiden mukaan käyttäjän kannalta alenevan relevanssin järjestykseen (*relevanssilajittelu*). Eri tiedonhakumallien avulla saavutettujen tulosten keskinäinen vertaileminen saattaa olla ongelmallista.

5.1 Kyselykeskeinen ja käyttäjäkeskeinen tiedonhakututkimus

Täystäsmäyttävien ja osittaistäsmäyttävien hakujen tutkimuksessa on käytetty sekä *kyselykeskeistä* että vertailevaa *käyttäjäkeskeistä* kokeellista tutkimusta. Kyselykeskeisessä tutkimuksessa tulee kiinnittää huomio tiedonhakujärjestelmien eroihin. Tuloksellisuuden vertailu onnistuu parhaiten, mikäli hakulausekkeet on muotoiltu erikseen eri täsmäytysmallien avulla suoritettavaksi. Tästä on myös empiiristä näyttöä (Tenopir & Shu 1989). Käyttäjäkeskeisissä koeasetelmissa on vertailtu mm. koehenkilöiden Boolen operaattoreilla toimivassa täystäsmäyttävässä ja osittaiseen täsmäytykseen perustuvassa hakujärjestelmässä suorittamien hakujen tuloksellisuutta. Kokeissa on mitattu saantia ja tarkkuutta sekä löytyneiden dokumenttien arvoa suhteessa tulosjoukon kokoon ja hakuaiheisiin keskimäärin. Dokumentit voidaan arvottaa myös niiden uutuuden mukaan, vaikka dokumenttien ikää ei niiden aiherelevanssin kriteerinä yleisesti hyväksytäkään. Käyttäjät voivat muotoilla kaikkiin täsmäytysmenetelmiin sopivia hakulausekkeita, mutta käyttäjistä riippuvien muuttujien vaikutusta hakujärjestelmästä riippuviin muuttujiin ei pystytä kontrolloimaan. Käyttäjistä riippuvien muuttujien kontrollointi kääntyy ongelmaksi, mikäli tutkimuksessa on tarkoitus selvittää tiedonhakujärjestelmän erilaisiin täsmäytysmalleihin perustuvia ydinominaisuuksia. Eri hakujärjestelmien edellyttämien hakulausekkeen muotoilussa ilmenevien systemaattisten erojen analysointi tuottaa myös ongelmia. Kyselykeskeisten ja käyttäjäkeskeisten kokeellisten tutkimusten päämäärä on yhtenevä: tutkimuksissa on pyritty mittaamaan eri hakujärjestelmien suorituskykyä, tehokkuutta ja tuloksellisuutta. Tiedonhaun asiantuntijoita kiinnostavia kysymyksiä ovat:

- millaisissa tilanteissa osittaistäsmäyttävä haku toimii parhaiten?
- milloin täystäsmäyttävä haku on toimivampi?
- millainen hakulausekkeen muotoilun strategia on soveltuvin käytettäväksi täys- ja osittaistäsmäytyksessä?

Web-hauissa käyttäjä voi valita, käyttääkö hän yksinkertaisempaa hakuavainlistoihin perustuvaa bag-of-words -hakua vai vaativampaan tiedonhakuun tarkoitettua rakenteisen hakulausekkeen mallia.

Sormusen artikkelissa (2002) käydään läpi hakujen kontrolloitujen muuttujien vertailumenetelmää täys- ja osittaistäsmäyttävissä hakumalleissa ja sen ohella menetelmää soveltaneesta tapaustutkimuksesta saatuja tuloksia. Tutkimusmetodin periaatteet ovat:

- hakulausekkeen muotoilu ja optimointi erikseen kutakin hakujärjestelmää varten, jotta hakutulosten vertailtavuus paranisi
- hakutulokseen vaikuttavien kontrolloimattomien muuttujien osuuden vähentäminen minimiin suorittamalla aihekohtaiset haut yhtenäisen haku-suunnitelman mukaan
- tulosten optimaalisuus edellyttää relevanssitietojen saatavuutta ja käyttöä hakulausekkeen muotoiluprosessissa; metodi on takautuva
- tiedonhakupeli on interaktiivinen työväline, jonka avulla käyttäjä muotoilee tutkittua hakutyyppiä vastaavan optimaalisen hakulausekkeen

Koska täsmälleen samoin muotoillun hakulausekkeen syöttäminen vertailtaviin hakujärjestelmiin voi estää optimaalisen hakutuloksen saavuttamisen, hakulausekkeen muotoilussa käytetään hakuaiheen eri aspekteja kuvaavia käsitteitä. Toisaalta kattava hakusuunnitelmakin edustaa vain yhtä hakuaiheen tulkintaa. (Sormunen 2002.)

5.2 Hakulausekkeiden optimointi ja hakujen tuloksellisuuden testaaminen

Tässä esiteltävän tutkimuksen (Sormunen 2002) tarkoitus oli vertailla Boolen hakujen sekä rakenteisten ja rakenteettomien todennäköisyyteen perustuvien hakujen tehokkuutta ja tyypillisiä piirteitä. Haut pyrittiin optimoimaan erikseen relevanssitietojen perusteella. Tutkimuksessa selvitettiin hakusuorituksen yleisiä piirteitä, haun kattavuutta ja laajuutta sekä hakusuorituksen piirteitä kattavuudeltaan eri tasoisissa hauissa. Tutkimus toteutettiin koeolosuhteissa InQuery-hakujärjestelmässä suomenkielisiä sanomalehtiartikkeleita sisältävästä kokotekstitietokannasta, josta tutkimusta varten valittiin rajattu osa. Testitietokannassa oli joukko aiheita, joiden tarkkuus ja relevanssikategoriat määriteltiin aiemmissa

tutkimuksissa saatujen täystäsmäyttävien ja osittaistäsmäyttävien hakujen tulosten perusteella. Testitietokantaan liitettiin myös asiantuntevien hakuanalyytikkojen valmiiksi muotoilemia hakulausekkeita, joiden avulla pyrittiin tunnistamaan testitietokannan kaikkien hakuaiheiden kaikki löydettävissä olevat fasetit.

Testaus suoritettiin systemaattiseen otantaa perustuvasta artikkelikokoelmasta, testiotoksesta, johon sisältyi kaikkiaan 661 dokumenttia 18 hakuaiheesta. Testiotoksen kaikille relevanteille dokumenteille tehtiin tekstianalyysi. Näin löydettiin dokumenteista kaikki hakulausekkeen muotoilussa käyttökelpoiset fasetit. Testiotoksen relevanssitiedot sekä artikkeleissa esiintyvät fasetit ja hakuavaimiksi soveltuvat ilmaukset olivat luotettavia. Näin varmistettiin hakuavainten esiintyminen ainakin joissain testiotokseen kuuluvista dokumenteista ja hakuaiheista. Kokeen päämääränä oli myös puitteiden luominen oikeassa käyttötilanteessa suoritettavaa tutkimusta varten, jossa käyttäjät yrittävät löytää parhaita mahdollisia hakuavaimia ilman ulkopuolista ohjausta. Taannehtivassa arvioinnissa pyritään sulkemaan pois dokumenttien ennakoimattomat piirteet käyttämällä hakulausekkeen optimointiin harjoitustietokantaa ja haun tuloksellisuuden testaamiseen erillistä testitietokantaa. Testiotoksen hakuaiheet jaettiin kahteen ryhmään – harjoitustietokantaan ja testitietokantaan. Harjoitustietokannassa oli 335 artikkelia, kun taas tuloksellisuutta mitattaessa käytettävään testitietokantaan sisällytettiin 326 artikkelia. Hakulausekkeen muotoilussa tarvittavat hakuavaimet valittiin faseteista seuraavasti:

- luotiin lista kaikista ilmauksista, joita fasetti harjoitustietokannan dokumenteissa edustaa
- jätettiin pois monimutkaiset fraasit, joilla ei ollut vakiintunutta merkitystä
- jätettiin pois kaikki ilmaukset, jotka esiintyivät ainoastaan yhdessä relevantissa dokumentissa

Testausryhmään kuului kolme InQuery-hakujärjestelmään ja tiedonhakupelin mahdollisuuksiin hakujen analysoinnissa hyvin perehtynyttä asiantuntijaa. Tallennettua tekstitiedostoa ja hakuavainten ominaisuuksia oli muokattu parantamaan käyttäjäystävällisyyttä testaajien näkökulmasta. Hakulausekkeen optimointia testattiin kahdella tasolla. Ensin rajattiin hakuaiheiden määrä kuuteen aiheeseen ja hakuun käytettävä aika kuuteen tuntiin yhtä hakuaihetta kohti. Toisessa vaiheessa jokaisen testattavaksi annettiin kahden muun testaajan optimointeja kolmesta hakuaiheesta.

Menettelyn avulla tarkistettiin optimoitujen tulosten mahdolliset syntaktiset ja tekniset virheet. Samalla pyrittiin löytämään entistä optimaalisempia hakulausekkeita hakujen kattavuuden kaikilla tasoilla testaamalla kymmenen haun eri versioita kaikilla kattavuuden tasoilla hakulauseketyypeittäin. Tiedonhakupeli oli erillisenä käytössä optimoitaessa hakulausekkeiden eri tyyppejä, jolloin hakija pystyi suorittamaan vertailuja hakukategorian sisällä, muttei eri kategorioiden välillä. Hakujen tuloksellisuuden vertailussa käytettiin mittarina keskitarkkuutta. Tiedot hakuihin käytetystä ajasta, käyttäjätunnisteet ja keskitarkkuus tallentuivat lokitiedostoon. Parhaiden hakulausekkeiden kattavuus oli tarkistettavissa lokitiedostosta. Hakulausekkeista kävi ilmi sekä fasettien käyttö että käytettyjen fasettien määrän vaikutus hakutulokseen. Toteutettiin laaja testaus. Kokeessa suoritettujen hakujen määrä vaihteli hakuaiheesta riippuen 77 - 3050 hakuun. Hakulausekkeiden kaksivaiheisen optimoinnin hyöty oli se, että voitiin havaita syntaksivirheitä, jotka vaikuttivat oleellisesti optimoinnin tulokseen. Pyrkimyksenä oli edelleen paljastaa yksittäisen hakijan tehtäväsuorituksen ”sokeat pisteet” ja todettiin, että hakijaa vaihtamalla hakutulos parani huomattavasti hakutyypistä riippumatta. Kaikissa tapauksissa parantuneella hakutuloksella ei kuitenkaan ollut käytännön merkitystä. (Sormunen 2002, 6 - 13.)

5.2.1 *Takautuva arviointi ja tiedonhakupelin käyttö hakulausekkeiden analyysissa*

Sormunen (2002) on hyödyntänyt Harterin (1990) näkemyksiä takautuvasta arvioinnista tutkiessaan Boolen hakujen tuloksellisuutta ja rakennetta ja laajentanut menetelmän käyttöä sekä täystäsmäyttävien että osittaistäsmäyttävien hakujen vertailuun. Osittaistäsmäyttävät haut voidaan jakaa rakenteensa perusteella *heikkoihin* ja *vahvoihin* hakuihin. Bag-of-words -haut edustavat heikkoa rakennetta, koska hakuavainten väliset suhteet eivät ole niissä näkyvissä. Vahva hakurakenne puolestaan osoittaa myös hakuavainten keskinäiset suhteet. Osittaistäsmäyttävässä hakulausekkeen muotoilussa voidaan käyttää joko hakuoperaattoreita tai luonnollisen kielen ilmauksia.

Hakujen optimointiprosessin perustana on vuorovaikutteinen tiedonhakupelillä suoritettu hakuanalyysi. Erillisten harjoitus- ja testikokoelmien avulla voidaan välttää hakulausekkeiden optimoinnissa esiintyviä ongelmia, vaikka erillisten kokoelmien käyttö lisääkin työn määrää. Tutkimuksen kannalta tärkeää on suunnittelun ohella

väliin tulevien muuttujien vaikutuksen vähentäminen. Hakusuunnitelma tehdään hakulausekkeen muotoilun osalta ilmaisutasolla. Avaintekijöihin kuuluu myös hakulausekkeen optimointiprosessin kontrolloitavuus. Mitä enemmän vapautta optimointiprosessiin sisältyy, sen vaikeampaa on tehdä empiirisistä tuloksista päteviä päätelmiä. Tiedonhakupelin käytöstä hakulausekkeiden optimoinnissa ei ole tehty kattavaa arviointia. Tiedonhakupelin käytön etu on se, että menetelmä tukee myös hakutulosten yksityiskohtaista analyysia. Hakulausekkeen optimoinnin käänteinen puoli on se, että hakijat alkavat optimointiyrityksissään keskittyä pelaamiseen ja parempien pisteiden toivossa unohtavat testin alkuperäisen tarkoituksen. Niillä hakuavainten kattavuuden tasoilla, joilla saavutettiin suurin tehokkuus, tehtiin enemmän hakuja kuin muilla kattavuustasoilla.

5.3 Optimaalisten hakulausekkeiden tuloksellisuudesta ja rakenteesta

Testeissä ilmeni, että rakenteiset osittaistäsmäyttävät haut tuottivat hieman muita hakuja paremman tuloksen. Niiden keskitarkkuus oli 0.07 Boolen hakujen ja 0.06 rakenteettomien osittaistäsmäyttävien hakujen yläpuolella. Ero ei ollut tilastollisesti merkitsevä. Kun haun tuloksena löytyneiden dokumenttien määrä oli pienimmillään, Boolen hakujen tarkkuus oli samaa luokkaa kuin osittaistäsmäyttävien hakujen tarkkuuskin. Mutta kun löytyneiden dokumenttien määrä oli suurimmillaan, Boolen haut eivät olleet yhtä tehokkaita kuin osittaistäsmäyttävät haut. Boolen hauissa havaittu alhainen tarkkuus suurimpien tulosjoukkojen kohdalla oli osittaistäsmäyttävien hakujen tarkkuuteen verrattuna tilastollisesti merkitsevä. Tulos tuki osittain testihypoteesia, jonka mukaan rakenteisten hakulausekkeiden oletettiin tuottavan muita hakukategorioita parempia tuloksia osittaistäsmäytyksessä, koska niissä yhdistyvät dokumentin relevanssilajittelu ja hakulausekkeen rakenne. Boolen hakujen tulos selittyy testiympäristöllä olosuhteissa, joissa dokumenttien löytyvyys on heikko. InQuery järjestää dokumentit tulosjoukon sisällä niiden aihevastaavuuden mukaan. Kun saantiluvut ovat korkeimmillaan, Boolen hakujen tarkkuus laskee, koska täystäsmäytyksessä jotkut relevantitkin dokumentit jäävät löytymättä.

Vaikeimmin löydettävissä oleville dokumenteille on tunnusomaista se, ettei niissä ole hakufasettien sisältöön liittyviä käsitteitä tai tekstissä käytetyt ilmaukset eivät vastaa hakulausekkeessa käytettyjä hakuavaimia. Sopivien ilmausten esiintyminen

vaikeimmin löydettävissä dokumenteissa on vähäistä eikä haku tavoita tekstin asiasisältöä. Dokumenttien relevanssia ei voida tällaisissa tapauksissa arvioida niissä esiintyvien hakuavainten perusteella.

5.3.1 *Optimaalisten hakulausekkeiden rakennepiirteitä*

Haun kattavuudella tarkoitetaan fasetointiin perustuvassa hakulausekkeen muodostuksessa käytettyjen fasettien lukumäärää. Mitattaessa haun laajuutta fasettia kohden käytetään mittarina hakuavainten keskimääräistä lukumäärää. Boolen hakujen suuri kattavuus verrattuna rakenteisiin ja rakenteettomiin osittaistäsmäyttäviin hakuihin oli tilastollisesti merkitsevä. Hakujen kattavuudessa mitatut erot eivät olleet merkitseviä osittaistäsmäyttäviä rakenteisia ja rakenteettomia hakuja vertailtaessa. Rakenteettomat haut olivat keskimäärin laajempia kuin rakenteiset haut. Boolen haut olivat keskimäärin melko laajoja. Mainituilla havainnoilla ei voitu osoittaa olevan tilastollista merkitsevyyttä. Kun Boolen hauissa fasettien käyttö edellyttää täystäsmäytyvyyttä, ei Boolen hakujen alhainen kattavuus ollut yllätys. Optimaalinen haku oli kattavampi ja hakujen keskitarkkuus korkeampi, kun hauissa käytettiin yhden fasetin asemesta useampia fasetteja. Rakenteisissa ja rakenteettomissa osittaistäsmäyttävissä hauissa pitäisi käyttää lähes kaikkia hakukelpoisia fasetteja, jotta hakutulos olisi paras mahdollinen. Hakujen optimointiprosessissa jäivät käyttämättä fasetit, jotka olivat yleisluonteisia tai muuten vaikeasti ilmaistavissa hakulausekkeen puitteissa. Jos hauissa käytetään vain yhtä fasettia, ei eri hakutyyppien välillä ole järin suuria eroja saannissa ja tarkkuudessa. Monifasettikyselyssä tuloksena saatujen dokumenttien relevanssi on listauksen kärjessä hyvä, listauksen lopussa vähäinen.

Optimaalisten hakulausekkeiden vertailu niiden kattavuuden eri tasoilla edesauttaa eri-tyyppisten hakulausekkeiden käyttäytymisen ymmärtämistä hakujärjestelmässä. Kun pyritään löytämään kaikki relevantit dokumentit, voi yksittäiseen fasettiin perustuva haku tuottaa hyvän tuloksen siinä, missä fokusoitu osittaistäsmäyttävä hakukin, mikäli hakuavainten kattavuus on hyvä dokumentin sisältöön nähden. Tarkkuus oli melko alhainen kaikkien hakutyyppien osalta kaikilla kattavuustasoilla, kun löytyneiden dokumenttien määrä oli suurin. Rajaavan ja-operaattorin käyttö Boolen hauissa alensi tarkkuutta, samoin haun liian laaja kattavuus oli Boolen hakujen tehokkuuden ja toimivuuden kannalta riski. Kun hakufasetteja oli useampia, Boolen hakujen tarkkuus

laski muiden hakujen tarkkuusarvojen alapuolelle, vaikka hakumenetelmä sinänsä oli hyvä. Perinteisten Boolen hakujen tarkkuuden hajonnan aiheutti täystäsmäytyvyyden vaatimus. Sormunen (2002) viittaa artikkelissaan aiempiin tutkimuksiin, joissa oli osoitettu, että haun laajentamisesta on etua rakenteisessa haussa, kun taas rakenteettoman haun tehokkuus pyrkii alenemaan haun laajetessa. Ideaalisissakin olosuhteissa testauksissa saatu tarkkuus oli 0,10 tuntumassa. Tämän mukaan laajojen web-tietokantojen hakuennuste olisi huono.

Sormunen (2002) mainitsee, ettei rakenteisten ja rakenteettomien optimaalisten hakujen tarkkuudessa tai kattavuudessa ole huomattavia eroja. Tutkimuksessa esitellään havainnon tueksi hakufraasien ja läheisyysoperaattorien käytön sekä toisaalta fokusoitujen hakuavainten käytön vaikutuksia haun tarkkuuteen ja kattavuuteen. Kun hakufraasit pilkottiin yksittäisinä sanoina käytettäviksi hakuavaimiksi, havaittiin, ettei rakenteettomien hakujen keskitarkkuuden aleneminen ollut tilastollisesti merkitsevää. Tästä voitiin päätellä, ettei ennakoivan ja takautuvan arvioinnin tuloksissa ilmenevä ristiriitaisuus ollut selitettävissä läheisyysoperaattorien käytöllä. Takautuvassa arvioinnissa johtopäätökset hakulausekkeiden toimivuudesta tehdään jo saatujen tulosten perusteella. Kun hakulausekkeen muotoilussa käytettiin fokusoituja hakuavaimia, kaikilla hakuavaimilla ja -faseteilla oli esitys kaikissa harjoitustietokannan relevanteissa dokumenteissa. Optimointiprosessissa käytettiin vain haun tehokkuutta lisääviä hakuavaimia. Hakuavainten käytöstä aiheutuva häly oli pyritty minimoimaan lisäämällä optimaalisten hakulausekkeiden fasetteihin laajoja hakuavaimia, jolloin Boolen hakujen tarkkuus aleni rakenteettomien hakujen tarkkuutta nopeammin.

6. Tutkimusasetelma

Tässä aineistopohjaisessa tutkimuksessa selvitetään, millaiset käyttäjien muotoilemat hakulausekkeet ovat tehokkaimpia täys- ja osittaistäsmäyttävässä haussa. Aineistona on tiedonhaun perusteiden kurssilla syksyllä 2003 kerätty harjoitusaineisto. Harjoituksissa käytettiin tiedonhakupeliä, jonka avulla opiskelijoita perehdyttiin tekstitiedonhakuun täys- ja osittaistäsmäyttävästä hakujärjestelmästä. Tarkoituksena on tutkia hakulausekkeiden toimivuutta ja hakujen onnistumiseen vaikuttavia tekijöitä

eri hakujärjestelmissä empiirisesti harjoitusaineiston ja opiskelijoiden esseevastausten pohjalta. Esseeaineistosta tutkitaan, miten opiskelijat ovat kertoneet hakuharjoituksiin liittyvistä kokemuksistaan ja asennoitumisestaan tiedonhakupeliin yleensä. Aineiston tulkinnassa käytetään sekä kvalitatiivisia että kvantitatiivisia menetelmiä.

Tiedonhakupeliä käytetään tekstitiedonhaun opetuksessa ja tutkimuksessa. Hakuja tehdään Aamulehden tekstiarkistosta ja Kauppalehden artikkelitietokannasta sekä TUTK-tekstitietokannasta. Aamulehden tekstiarkistossa on artikkeleita vuosilta 1997 - 99, Kauppalehdestä on saatavilla artikkeleita vuosilta 1997 - 2000. TUTK-tekstitietokantaan sisältyy sanomalehtiartikkeleita Aamulehdestä, Keskisuomalaisesta ja Kauppalehdestä 1990-luvun alusta.

Tiedonhakupelillä kahta erilaista hakumenetelmää käyttämällä tehdyistä harjoituksista on tallennettu lokitiedot. Myös opiskelijoiden viimeisen harjoituskerran yhteydessä kirjoittamat esseet on tallennettu. Opiskelijat ovat kuvanneet esseissään hakuharjoitusten yhteydessä esiintyneitä ongelmia ja hakulausekkeiden kohentamiseen liittyviä ideoitaan. Aineistoon sisältyy TRIP-hakujärjestelmällä suoritettuja täystäsmäyttäviä Boolen hakuja 4408 kpl. InQuery-hakujärjestelmällä tehtyjä osittaistäsmäyttäviä hakuja on 4446 kpl. Kukin hakija on tehnyt molemmilla hakujärjestelmillä useita hakuyrityksiä pyrkiessään parantamaan yksittäistä hakulauseketta. Lokitiedoista tehdyistä Excel-taulukoista voidaan nähdä käyttäjittäin mitä hakuja on tehty, missä järjestyksessä haut on tehty, ja mikä haku on milläkin hakujärjestelmällä onnistunut parhaiten yksittäisen harjoitustehtävän kohdalla. Käytössäni on informaatiotutkimuksen perusopintojen syksyn 2003 tiedonhakukurssin harjoitusten lokitiedosto, jossa hakuaiheina olivat Etelä-Amerikan velkakriisi, Helsingin poliisisurmat, kalaruokareseptit, kilpaveneily ja tekstiiliteollisuus. Lokitiedoston kolmen ensiksi mainitun hakuaiheen tehtävistä on laadittu erikseen yhteenveto hakijakohtaisista parhaista hauista.

Olen valinnut tarkasteltavakseni Etelä-Amerikan velkakriisi -aiheesta tehdyt hakulausekkeet. Hyödynnän työssäni lokitietojen ohella valmista hakuavainten laatuluokitusta, jossa on eroteltu tämän hakuaiheen pääkäsitteet ja luokiteltu hakuavaimet ankkuritermeihin, kapeisiin termeihin sekä hakulausekkeessa hyvin tai huonosti toimineisiin hakutermeihin. Hakuavainten laatuluokitus perustuu niiden

hakuaiheeseen liittyvien dokumenttien määrään, joissa vastaava käsite esiintyy. Laatuluokituksessa hakuavaimista käytetään nimitystä ”termi”. Harjoitustietokannassa on yhteensä 51 relevanttia dokumenttia. Ankkuritermi esiintyy vähintään 34:ssä, hyvä hakutermi 17 - 33 dokumentissa. Tällöin ankkuritermeillä on edustus 25 %:ssa ja hyvällä hakutermillä 10 %:ssa kaikista relevanteista dokumenteista. Kapealla hakutermillä on edustus enintään 16 dokumentissa. Huonosti toimivien hakutermien edustus relevanteissa dokumenteissa on vähäinen. Ankkuritermi on hakuaiheen yleisin käsite. Sen hakuteho on hyvä, kun haussa pyritään mahdollisimman suureen saantiin. Hakutulos ei välttämättä ole tarkka. Hyvän hakutermin hakualue ei ole yhtä laaja. Kapea hakutermi kohdentaa ja tarkentaa hakua.

Tiedonhakupelin avulla on mahdollista löytää kaikkiaan 51 hakuaihetta käsittelevää relevanttia dokumenttia. Tiedot aiheesta tehtyjen parhaiden hakujen tuloksellisuudesta on esitetty lokitiedostossa kahtena erillisenä tehtäväryhmänä. Ensimmäisessä tehtäväryhmässä on sekä TRIPillä että InQuerylla suoritettujen parhaiden yksilökohtaisten hakujen tuloksellisuustiedot. Näitä hakuja on yhteensä 168 – 84 hakijan parhaiten onnistunut haku kummastakin hakujärjestelmästä. Tuloksellisuustiedoista näkyy hyvin onnistuneen hakusuorituksen tarkkuusarvon ohella myös hakujärjestelmäkohtainen keskitarkkuusarvo, joka on ensimmäisessä ryhmässä TRIPillä 0,151 ja InQuerylla 0,315. Parhaiden hakujen tuloksellisuustaulukossa on näkyvissä hakujen suora ja prosentuaalinen tarkkuusjakauma. Jälkimmäisessä tehtäväryhmässä hakuaihe on sama, mutta tehtävänasettelun lähtökohta hieman erilainen. Hakuharjoitusten yhteydessä Etelä-Amerikan velkakriisi -aiheesta tehtyjen parhaiden hakujen määrä on jälkimmäisessä tehtäväryhmässä 116 – 58 hakijan onnistunein haku TRIPistä ja InQuerysta. Tuloksellisuustietoihin on kirjattu hakusuoritusten tarkkuusarvot. Täystäsmäyttävien TRIP-hakujen keskitarkkuus on tässä ryhmässä 0,12 ja osittaistäsmäyttävien InQuery-hakujen keskitarkkuus 0,515. Ensimmäisessä ryhmässä TRIP-hakujen keskitarkkuus on alle puolet (0,48) InQuery-hakujen keskitarkkuudesta, jälkimmäisessä ryhmässä TRIPin keskitarkkuus on jäänyt alle neljäsosaan (0,23) InQueryn keskitarkkuudesta.

Analysoin luvussa 7 tiedonhakukurssin harjoituksiin osallistuneiden opiskelijoiden esseevastausten sisältöä. Mitä hakujärjestelmän piirteitä tiedonhakupelin käyttäjät ovat hyödyntäneet harjoitellessaan tekstitiedonhakua? Millaisia virheitä he ovat tehneet?

Miten he ovat pyrkineet korjaamaan hakulausekkeitaan saavuttaakseen parempia ja tarkempia hakutuloksia? Miten tiedonhakupelin kahden hakujärjestelmän ja yksittäisen käyttäjän välinen interaktiivisuus toimii? Miten se vaikuttaa hakutuloksiin? Olen luokitellut luvussa 7 tapahtuvaa tarkastelua varten opiskelijoiden esseevastausten asiakokonaisuudet tiedonhakupelin palautemuotoja, teknisiä ongelmia, tehtyjä parannusehdotuksia ja hakuavainten valintaa koskeviin puheisiin. Luvun lopussa käsittelen opiskelijoiden näkemyksiä opastuksen tarpeesta hakuharjoitusten yhteydessä, harjoitustilanteen yleisiä ongelmia sekä opiskelijoiden esittämää harjoitustilanteen kritiikkiä. Seitsemännen luvun viimeisessä alaluvussa esitän yksittäisen hakijan selontekoon perustuvan kuvauksen tiedonhakuprosessin etenemisestä.

Käsittelen luvussa 8 Etelä-Amerikan velkakriisiin liittyvän hakuaiheen lokitietoja. Rajaan aihettani poimimalla lokitiedoista 84 parhaan täys- ja osittaistäsmäyttävän haun hakulausekkeet, jolloin tarkasteltavassa osa-aineistossa on 168 hakulauseketta tuloksellisuustietoineen. Informaatiotutkimuksen perusopintojen tiedonhakukurssin harjoitusten lokitiedostosta eroteltujen parhaiden hakujen luokittelu perustuu saatujen hakutulosten tarkkuuteen. Tarkoitukseni on tutkia, miten eri tavoin muotoillut hakulausekkeet käyttäytyvät tiedonhakupelin hakujärjestelmissä. Pyrin tunnistamaan hyvin toimivia hakuavaimia Etelä-Amerikan velkariisi -aiheesta tehtyjen parhaiden hakujen tarkkuuden pohjalta.

7. Kokemuksia tekstitiedonhausta tiedonhakupelillä

Opiskelijat ovat kuvanneet esseissään monipuolisesti kokemuksiaan tekstitiedonhausta tiedonhakupelillä. Selonteoista saa hyvän käsityksen opiskelijoiden mielessä hakuharjoitusten eri vaiheissa liikkuneista asioista. Tiedonhaun harjoituskurssilla tehtiin hakuja kahdella erilaisella hakujärjestelmällä. TRIP-hakujärjestelmässä haut perustuvat täydelliseen täsmäävyyteen, InQuery-hakujärjestelmän hauissa noudatetaan osittaistäsmäyttävää menetelmää. Kuten käyttöjärjestelmissä yleensä, myös tiedonhakupelissä on hakuhistoria, johon tallentuu hakulausekkeita. Hakulausekkeiden toimimattomuus on oma ongelmakenttänsä, mikä käy ilmi sekä tiedonhakupelin hakuharjoitusten yhteydessä suoritettujen hakujen tuloksista että opiskelijoiden kirjallisista selonteoista. Hakuhistoriaa pidetään yleisesti kätevänä toimintona, koska

se auttaa hahmottamaan hakulausekkeita paremmin ja myötävaikuttaa niiden kehittämiseen.

Tiedonhakupelin käyttäjien motivoituminen hakutehtävien tekemiseen liittyy olennaisesti optimaalisten hakuavainten etsimiseen ja hyvin toimivien hakulausekkeiden muotoiluun, joten kokoan työn tässä osuudessa yhteen ensin hakuharjoituksiin osallistuneiden opiskelijoiden myönteisiä ja kielteisiä ajatuksia tiedonhakupelin käytöstä. Esseeaineistoon sisältyy runsaasti tiedonhakupelin palautemuotojen käytettävyyttä, pelissä ilmenneitä teknisiä ongelmia ja opiskelijoiden peliin tekemiä parannusehdotuksia sekä hakuavainten valintaa koskevia pohdintoja.

7.1 Opiskelijoiden asennoituminen tiedonhakupeliin

Tiedonhakuharjoitusten yhteydessä opiskelijoita pyydettiin kertomaan positiivisimmat ja negatiivisimmat kokemuksensa tiedonhakupelin käytöstä ja palautemuodoista. Aineistossa on 80 opiskelijan kirjalliset vastaukset. Tehtäviä tehtiin sekä saliharjoituksissa että itsenäisinä verkkoharjoituksina. Tiedonhakupelin avulla suoritettaviin harjoituksiin suhtauduttiin yleensä myönteisesti. Tiedonhakupeliä pidettiin hyvänä keksintönä. Eräs osallistujista mainitsi positiivisena kokemuksenaan sen, että tiedonhakupeliin voi syntyä riippuvuus. Pitää vain ensin tajuta sen juoni. Optimaalisten hakuavainten ja käsitteiden etsintä tulee mielenkiintoisemmaksi, kun peliin pääsee sisälle. Hakupelin pelaaminen on mukavaa ja voi tuottaa onnistumisen tunteita. Pelatessa saa käyttää omia aivojaan. Tiedonhakupelissä yhdistyvät huvi ja hyöty. Opiskelijat tunnistivat itsessään oppimisen ilon, sillä oppiminen tapahtuu leikin ja pelin varjolla yrittämällä ja erehtymällä. Taidot tavallaan hiotaan pelin etenemisen myötä. Peli nähdään hauskana tapana oppia. Peliä pidetään helppokäyttöisenä, selkeänä ja nopeana. Tiedonhakupeli on oivallinen ympäristö, joka harjaannuttaa käyttäjäänsä muodostamaan kattavia ja hyviä hakulausekkeita sekä huomaamaan tiedonhaun ongelmakohtia. Peli tarjoaa mielekkään ja mukavan tavan opetella tiedonhakua ja hakulausekkeiden muodostamista. Hakujärjestelmän antama reaaliajassa toimiva palaute hakusuoritusten onnistumisesta on toteutukseltaan hyvä. Opiskelijoiden mielestä on hienoa, kun huomaa soveltaneensa pelin hakujärjestelmään hyviä fasetteja ja osuvia hakuavaimia. Hakuavainten ideointi rinnastetaan aivojumppaan. Kun hakija näkee valitsemiensa hakuavainten vaikutukset

hakutulokseen, niin samalla oma kehittyminen ja edistyminen tiedonhakijana havainnollistuu. Tiedonhakutaitojaan pystyy arvioimaan, sillä tiedonhakupelin harjoituksissa saa pelin palautejärjestelmän avulla käsityksen suoriutumisestaan. Vaikka aiheet ovat outoja, tietoa löytyy ja hakuavaimia kehittelemällä hakutulokset kohentuvat. Haun onnistuminen on osoitus hakijan omien ideoiden onnistumisesta. Tiedonhakupeli on hyödyllinen tiedonhaun osa-alueiden opiskelussa ja edesauttaa osaltaan myös hakutoimintojen syvällisempää ymmärtämistä. Saliharjoituksissa törmää usein aikapulaan. Verkkoharjoitukset taas suovat aikaa syventyä tehtäviin, mikä osaltaan myötävaikuttaa positiivisten kokemusten syntyyn. Tiedonhakupeli on käyttäjäystävällinen ja erinomainen tiedonhaun opetus- ja havainnollistamisväline. Tallennettu hakuhistoria on hyödyllinen ja auttaa oppijaa etenemään hakusuorituksessaan. Tiedonhakupelin interaktiivisuus nähdään yleisesti myönteisenä ominaisuutena. Harjoittelun ja toiston kautta tulee samalla kokemusta erilaisista tiedonhakutilanteista. Hyvänä puolena mainittiin se, että harjoitusten yhteydessä on mahdollisuus tutustua kahden erilaisen hakujärjestelmän käyttöön.

Vaikka opiskelijoiden asenteet tiedonhakupeliä kohtaan olivat esseissä yleisesti myönteisiä, niissä tuotiin esiin myös negatiivisia seikkoja. Melko monet opiskelijat tosin kertoivat, ettei heillä ollut erityistä negatiivista kommentoitavaa tiedonhakupelin harjoituskäytöstä. Haun jumiutuminen tuntui turhauttavalta, vaikka se saattoi johtua hakijan taidoista, ei tiedonhakupelistä sinänsä. Monet mainitsivat erityisesti TRIP-hakujärjestelmän ja Boolen täystäsmäyttävien hakujen ongelmallisuuden. TRIP-haut olivat turhauttavia ja vaikeita. Aikaa kului keksimiseen, sillä hakutaktiikan löytäminen edellytti erilaisten vaihtoehtojen kokeilemista. TRIP-hakujärjestelmässä harmia aiheutti myös hakulausekkeen rajattu pituus. Hakutulos ei tuntunut paranevan hakuavaimia vaihtelemalla eikä hakuavainten katkaiseminenkaan kohentanut tilannetta. Rinnakkaisten hakuavainten lisääminen hakulausekkeeseen aiheutti haun tuloksellisuuden alenemisen. Ongelmia ilmeni, kun hakukoneet reistailivat ajoittain. Hakulausekkeen kirjoitussääntöjä pidettiin vaikeasti omaksuttavina ja toimivan hakulausekkeen kehittely tuotti melkoisesti hankaluuksia, jos hakuaihe oli tuntematon. Opiskelijat ihmettelivät, mitä tekivät väärin, kun hakutulos ei monien yritysten jälkeenkään muuttunut paremmaksi. Pienet ja vaikeasti tunnistettavat huolimattomuusvirheet saattavat vaikuttaa radikaalisti hakutulokseen. Löydetyt dokumentit eivät aina edes liity haettuun aiheeseen. Hakuohjeisiin pitäisi perehtyä heti

alussa, sillä perustiedon puuttuminen on turhauttavaa. Monet pitivät InQueryn osittaiseen täsmäytykseen perustuvaa hakujärjestelmää TRIPiä tuloksekkaampana. Kun hakija halusi siirtyä hakujärjestelmästä toiseen, piti käytössä oleva järjestelmä aina sulkea ensin. Tämä antoi aihetta pohtia, miksi tiedonhakupelin kahta hakujärjestelmää ei voida käyttää samanaikaisesti.

Opiskelijoiden asenteissa on mielenkiintoista se, miten itsestään selvästi he puhuvat tiedonhakupelistä nimenomaan pelinä. Eräs hakija kertoo, ettei ole saanut otetta koko peliin, koska osa pelin toiminnoista on jäänyt hänen käsityskykynsä ulkopuolelle. Peli on saanut hakijan tuntemaan itsensä maailman huonoimmaksi hakujen tekijäksi ja kysymään, millaisia ”keskivertopelaajan” tulokset yleensä ovat olleet. Peli ei ole aina toiminut odotetulla tavalla. Pelatessa oivaltaa oman tyhmyytensä ja hitautensa. Peli ei toimi loogisesti, koska hakijan itsensä mielestä hyvät hakuavaimet tuottavat huonon hakutuloksen. Eräs hakijoista kertoi uskoneensa aina, että hänen henkilökohtainen sanavarastonsa on laaja ja monipuolinen. Hakijaa ihmetytti, ettei hänen käsityksensä kielestä mitenkään käynyt yhteen tiedonhakupelin hakulogiikan kanssa. Opiskelijoiden esittämistä ajatuksista ja kommenteista voidaan päätellä, että tiedonhakupelin käyttö sinänsä on herättänyt hakuharjoituksissa sekä positiivisia että negatiivisia tunteita.

7.1.1 *Tiedonhakupelin palautetoiminnot*

Mitä opiskelijat ajattelivat tiedonhakupelin palautetoiminnoista – dokumenttipalkista, saantipiirakasta, visualisoinneista ja kunniataulusta? Tiedonhakupelin palautemuotoja pidettiin yleensä mielekkäinä. Pelin antama palaute on hakijalle ensiarvoisen tärkeää. Joku kaipasi palautelomaketta omia kommenttejaan varten tiedonhakupelin harjoitusten yhteyteen sen sijaan, että palautteen antoon oli mahdollisuus vasta jälkeenpäin. Dokumenttipalkkia pidettiin hyvänä ratkaisuna, koska dokumenttien relevanssin sai konkreettisesti näkyviin, dokumenttien tekstejä pääsi tutkimaan ja etsimään niistä käyttökelpoisia hakuavaimia, mikä puolestaan mahdollisti hakujen parantamisen. Useissa vastauksissa korostui dokumenttipalkin hyödyllisyys, koska se oli suoraan apuna hakutehtävissä. Joissain vastauksissa dokumenttipalkki mainittiin palautemuodoista parhaimpana ja sitä seurattiin eniten hakutehtävien teossa. Tosin vastakkaisiakin mielipiteitä esitettiin: dokumenttipalkista ei ollut käytännön hyötyä tehtävien kannalta.

Dokumenttipalkki oli saantipiirakan ohella havainnollisuutensa ja selkeytensä vuoksi tiedonhakupelin useimmin käytetty palautemuoto. Hakijan mukaan dokumenttipalkki on tärkeä palautteen antaja, jos vain ensin huomaa lukea ohjeen. Varsinkin InQueryn osittaistäsmäyttävissä harjoitushauissa dokumenttipalkin relevanssipalaute oli havainnollinen. Konkreettisten relevanttien dokumenttien selaaminen auttoi hakijoita löytämään uusia hakuavaimia ja muotoilemaan paremmin toimivia hakulausekkeita. Hakijapalautteen mukaan dokumenttipalkki on hyvä, koska sen avulla pääsee näkemään relevanttien dokumenttien sijainnin ja määrän. Palkki on hyödyllinen arvioitaessa haun onnistumista kokonaisuutena. Palkki tosin antaa hyvää palautetta ja hausta saa sen avulla kokonaiskäsityksen, mutta se on toisaalta myös turhauttava. Informatiivisena dokumenttipalkki antaa osviittaa haulle ja selventää peliä.

Saantipiirakkaa ei huomattu, siitä ei välitetty tai sitten se ei hakijan mielestä ollut tarpeellinen. Sitä ei katsottu ollenkaan, siitä ei pidetty tai sitten sen nähtiin helpottavan haun saannin ja tarkkuuden tajuamista. Palaute oman haun onnistumisesta oli selkeä. Piirakka havainnollisti relevanttien dokumenttien määrän kerralla ja oli siksi hyödyllinen. Pidettiinpä sitä myös tiedonhakupelin palautemuodoista tärkeimpänä. Jotkut pitivät saantipiirakkaa turhana, koska saanti näkyy myös prosenttilukuina. Saantipiirakka nähtiin joko visuaalisena hienosteluna tai hyvänä pikapalautteena. Tiedonhakupelin käyttäjien mielipiteissä ilmeni selvästi hajontaa. Saantipiirakan sanottiin olevan joko erittäin havainnollisen tai aina iloisen yllätyksen. Se auttoi näkemään haun suunnan ja onnistumisen. Erään hakijan mielestä saantipiirakan ohella tiedonhakupelin palautemuodoksi olisi aivan yhtä hyvin sopinut tarkkuuspiirakka. Saantipiirakkaa pidettiin ihan kivana: graafinen esitys puoltaa aina paikkaansa.

Miten tiedonhakupelin visualisointeja – saantipiirakkaa ja tarkkuuskäyriä – yleensä kommentoitiin? Visualisointi oli hyödyllinen, selkeä ja monipuolinen. Toisaalta visualisointeja ei katsottu eikä käytetty, koska niistä ei pidetty tai välitetty. Visualisoinnit olivat epäselviä tai muuten vaikeaselkoisia. Ei niitä osannut käyttää. Numeroinformaatio ja dokumenttipalkin tiedot olivat riittävät. Visualisointi oli kummallinen ja herätti ihmettelyä, vaikka näyttikin hienolta. Tiedonhakupelin visualisointeja saatettiin vaihtoehtoisesti vain vilkaista, koska haluttiin tietää, miltä ne näyttivät. Ajateltiin, että visualisointi voisi toimia kannustimenakin. Visualisointi oli

hyödyllinen etenkin eri hakutavoilla toteutettujen hakujen saannin ja tarkkuuden vertailussa.

Tiedonhakupelin palautetoiminnoista mielipiteitä jakaa ehkä eniten kunniataulu. Opiskelijoiden vastauksissa kunniataulua pidetään joko hupina tai sitten sen oletetaan ”addiktoivan” käyttäjää eli aiheuttavan peliriippuvuutta. Pelimetaforahan esiintyy selkeästi hakusuorituksen analyysimenetelmän nimessä. Pelin pelaamisen ideaan liitetään opiskelijoiden vastauksissa usein ”kilpailuvietti”. Tiedonhakupelin kunniataulu tavallaan ruokkii kilpailuviettiä. Ainakin aluksi taulu lisää ”kilpailuhenkeä”. Kaikki hakijat eivät suinkaan ole vakuuttuneita kunniataulun hyödyllisyydestä, vaikka joidenkin mielestä sen aikaansaama kilpailuhenki on sinänsä hyvänlaatuista. Tiedonhakupeli kunniatauluineen on kuin mikä tahansa muu peli. Taulu kyllä kasvattaa ”taistelumieltä”, mutta kilpailua lisäävänä se on myös hermostuttava. Kunniataulun nähdäänkin korostavan tiedonhakupelin pelillistä luonnetta. Taulu on pelissä lähinnä hauska lisä, josta ei välttämättä koidu hyötyä hakijalle tiedonhaun näkökulmasta. Se on melko tarpeeton, mutta ”kiva” olemassa, mikäli välttämättä haluaa kilpailla. Joka tapauksessa kunniataulu konkretisoi haun tuloksellisuutta.

”Kannustamiseen” ja ”pärjäämiseen” liittyvät opiskelijoiden kommentit edustavat samaa merkityskenttää kuin vastauksissa esiintyvät muut pelin ja pelaamisen ideaa korostavat kuvaukset. Kunniataulua pidetään vastauksissa hyvänä kannustimena tai hyvänä porkkanana. Jos hakijalla on yleensä kilpailuviettiä, kunniataulu voi olla hänelle aikamoinen lisäkannuste. Siitä näkee, miten on itse pärjännyt muihin tiedonhaun harjoitusryhmän hakijoihin verrattuna. Myös ”tavoitteiden asettaminen” ja ”tulokset” mainitaan muutamien opiskelijoiden vastauksissa heidän pohtiessaan kunniataulun merkitystä tiedonhakupelin palautetoimintojen kokonaisuudessa. Kunniataulun avulla voi asettaa itselleen tavoitteita, koska siitä myös näkee, millaisiin tuloksiin tulisi suunnilleen päästä. Kunniataululla pistäytyminen näyttää olevan esseevastauksessa mainitsemisen arvoinen seikka. Parhaimmillaan kunniataulu innostaa hakemaan ja antaa pontta omien hakujen parantamiselle. Kunniataulu täyttää joidenkin opiskelijoiden mielestä sille asetetun tehtävän motivoidessaan hakijoita yrittämään parempia tuloksia hakuharjoituksissa. Haun arvioinnissa taululla ei sen sijaan ole järin suurta merkitystä. Kaikki eivät pidä taulua kovin informatiivisena,

tarpeellisena tai tärkeänä, saati sitten välttämättömänä, joten sitä ehkä vilkaistaan vain uteliaisuudesta. Monissa kommenteissa tuodaan spontaanisti esiin mielipide kunniataulusta tiedonhakupeliin itsestään selvästi kuuluvana osana. Harva perustelee näkemystään tarkemmin. Mielipiteet kunniataulusta palautemuotona vaihtelevat hyvästä tai loistavasta ideasta epätoivoa aiheuttavaan ja masentavaan toimintoon, jonka hämärää logiikkaa ei pysty mitenkään ymmärtämään.

7.1.2 *Tiedonhakupelin palautemuotojen toimivuutta kuvaavat laadulliset määreet esseeaineistossa*

Olen esittänyt koosteessani huomioita opiskelijoiden vastauksissa ilmenneistä mielipiteistä ja näkemyksistä. Useissa vastauksissa ei ole erityisemmin perusteltu tiedonhakupelin palautetoimintoihin liittyviä käsityksiä. Opiskelijat esittivät aiheesta pääasiasiassa yleisiä kommentteja. Esseeaineistosta löytyi ainoastaan muutamia palautetoimintojen positiivisia tai negatiivisia piirteitä korostavia puheenvuoroja, joissa käsitysten perusteluun oli kiinnitetty huomiota. Esimerkkejä kunniataulun ideaa koskevista, osin perusteluistakin kommenteista olivat:

1. Kunniataulu on oikein kiva. Se ei ole tavoitteena kovin realistinen.
2. Kunniataulu on hyvä, koska siitä pystyy seuraamaan muiden onnistumista ja vertaamaan tuloksia omiin suorituksiinsa.
3. Kunniataulu on hyvä. Miksei molemmille hakujärjestelmille voi olla omaa taulua?
4. Kunniataulu on huono, koska se näyttää useita saman ihmisen tekemiä hakuja.
5. Ei kunniataulusta ole hyötyä, ellei näe, millä [haku]lausekkeilla parhaat tulokset on saatu.

Eräs hakija mainitsi päässeensä kunniataululle ainoastaan InQuery-hakujärjestelmällä tekemiensä hakujen ansiosta. Aineistossa oli kymmenen vastausta, joissa ei mitenkään eritelty tai kommentoitu tiedonhakupelin palautemuotoja. Muutamissa vastauksissa tuotiin silti esiin yleisluonteisia tiedonhakupeliin liittyviä kommentteja. Graafisia kuvioita pidettiin sinänsä informatiivisina. Hakuhistorian pitäisi näyttää arvot tai listata [haku]tulokset parhaasta huonoimpaan. Palautteesta on helppo huomata, missä on itse mennyt vikaan. Tehtävät olisi hyvä käydä luennolla etukäteen läpi eri variaatioineen. Näin olisi harjoituksissa helpompi tutustua tiedonhakupelin suomiin mahdollisuuksiin.

Dokumenttipalkki mainittiin useissa vastauksissa tiedonhakupelin monipuolisena ja parhaiten hakutehtävien tekoa tukevana palautemuotona. Dokumenttipalkille annettiin

80 kirjallisessa vastauksessa yhteensä 73 laatumäärettä, joista lähes kaikki olivat positiivisia. Opiskelijat ovat kuvanneet kirjoittamissaan vastauksissa tiedonhakupelin hyödyllisintä palautemuotoa seuraavalla tavalla:

Dokumenttipalkki palautemuotona:	Maininnat:
hyvä/paras	17
hyödyllinen	17
tärkeä	8
hakusanojen ja aihedokumenttien löytyminen	8
havainnollinen	7
selkeä	7
ei hyödyllinen	2
konkreettinen	1
toimiva	1
Yhteensä:	73

Taulukko 1: Dokumenttipalkki tiedonhakupelin palautemuotona

Tiedonhakupelissä on mahdollisuus siirtyä tarkastelemaan hakuharjoituksissa löytyneiden relevanttien dokumenttien sisältöä dokumenttipalkin kautta. Relevanssipalautteen hyödyntäminen puolestaan edesauttaa hakuharjoituksissa uusien hakuavainten löytämistä ja hakulausekkeen kehittämistä toimivampaan suuntaan. Tässä mielessä hakuharjoitusten voidaan sanoa simuloivan todellisia hakutilanteita. Tiedonhakupelin hakuharjoituksissa tekstitietokantojen käyttöalue on rajattu ja pelin hakuaiheet esitetty hakujärjestelmässä määrämuotoisina.

Tiedonhakupelin visualisoinneista saantipiirakka osoittautui hyväksi ja toimivaksi palautemuodoksi. Saantipiirakka mainittiin opiskelijoiden kirjoittamissa 80 esseevastauksessa 15 kertaa samassa yhteydessä dokumenttipalkin kanssa. Molemmat palautemuodot ovat havainnollisia ja tukevat hyvin toisiaan. Vaikka eräässä esseessä oli maininta saantipiirakan antaman palautteen vähäisemmästä hyödystä dokumenttipalkin palautteeseen verrattuna, niin kuitenkin oltiin yksimielisiä siitä, että tiedonhakupelin parhaat apuneuvot olivat dokumenttipalkki ja saantipiirakka yhdessä. Palautemuodot edesauttavat hakulausekkeen toimivuuden ymmärtämistä, koska hakija näkee helposti, konkreettisesti ja nopeasti oman hakunsa onnistumisen tai epäonnistumisen. Opiskelijoiden esseevastauksissa liitettiin dokumenttipalkin ja saantipiirakan toinen toistaan tukeviin palautemuotoihin niiden yhteisvaikutuksesta juontuvia laatumääreitä seuraavasti:

Dokumenttipalkki ja saantipiirakka yhdessä:	Maininnat:
hyödyllinen	4
havainnollinen	3
hyvä/paras	2
katsotuin	2
tukevat toisiaan	2
tärkeä	2
selkeä	1
Yhteensä:	16

Taulukko 2: Dokumenttipalkki ja saantipiirakka yhdessä tiedonhakupelin palautemuotoina

Opiskelijoiden 80 esseevastauksen joukossa oli yhteensä 14 vastausta, joissa ei ollut lainkaan mainintaa saantipiirakasta tiedonhakupelin palautemuotojen käsittelyn yhteydessä. Kaksi vastaajaa jätti saantipiirakan katsomatta. Eräs vastaaja kertoi, ettei tutkinut saantipiirakan palautetta usein. Toinen kertoi, ettei välittänyt tästä toiminnosta. Yksi vastaaja mainitsi, ettei löytänyt saantipiirakkaa tiedonhakupelin palautetoiminnoista. Eräs vastaajista piti tiedonhakupelin saantipiirakkaa visuaalisena hienosteluna, kun taas toinen kyseenalaisti täysin tämän toiminnon tarpeellisuuden. Saantipiirakkaan liittyvät laatumääreet ovat silti suurimmaksi osaksi positiivisia. Saantipiirakan sanottiin auttavan hahmottamaan haun onnistumista yleensä ja helpottavan saanti- ja tarkkuusarvojen tajuamista.

Saantipiirakka palautemuotona:	Maininnat:
havainnollinen	10
hyödyllinen	9
hyvä/paras	8
katsotuin	5
tärkeä/tärkein	5
turha	5
pidetty	5
konkreettinen	2
selkeä	2
tarpeen	2
helppo	1
Yhteensä:	54

Taulukko 3: Saantipiirakka tiedonhakupelin palautemuotona

Tiedonhakupelistä on kerrottu esseissä, että sen graafisen palautteen kaikki muodot yhdessä ovat toimivia. Graafiset kuviot ovat hyvin informatiivisia, koska ne antavat visuaalisessa muodossa tietoa haun onnistumisesta. Joillekin ihmisille kuviot ovat numeroita havainnollisempia. On hieman tulkinnanvarainen asia, mitä opiskelijat tarkoittavat puhuessaan esseissään visualisoinnista, visualisoinnit-valinnasta tai visualisointi-osuudesta. Onko puheessa kyse tiedonhakupelin koko grafiikasta dokumenttipalkkeineen, kunniatauluineen ja saantipiirakoineen? Vai tarkoitetaanko mahdollisesti visualisoinnit-valinnan saanti- ja tarkkuuskäyrää, jonka avulla tietoa hakeva henkilö pystyy seuraamaan oman hakuntuloksensa kehitystä parhaaseen hakuun verrattuna. Mitä tiedonhakupelin visualisointeihin sisältyy? Mitä visualisoinneista ajatellaan yleisesti? Vaikka visualisoinnit ovat käteviä, niiden tarpeellisuudesta esitetään monia näkemyksiä. Opiskelijoiden esseevastauksista näkyy jokseenkin vahva mielipiteiden jakautuminen: visualisointeja ei tarvita lainkaan tai sitten niitä pidetään tiedonhakupelin palautemuodoista hyödyllisimpänä. Eräässä vastauksessa visualisointeihin toivottiin lyhyttä selitystä parhaan haun tuloksellisuustiedot osoittavasta vertailuviivasta, jotta oman haun saannin ja tarkkuuden hahmottaminen kyseisen palautemuodon avulla helpottuisi.

Visualisoinnit palautemuotona:	Maininnat:
hyödyllinen	11
ei käytetty/ei osattu käyttää	10
tuloksia havainnollistava	8
vertailua helpottava	7
sekava	5
vähiten hyötyä	5
ei pidetty	1
epäkäytännöllinen	1
hienon näköinen	1
hyvä	1
kiva	1
kätevä	1
käytännöllinen	1
mielenkiintoinen	1
paljon käytetty	1
vaikeahko tulkita	1
Yhteensä:	56

Taulukko 4: Visualisoinnit tiedonhakupelin palautemuotona

7.1.3 *Tiedonhakupelin tekniset ongelmat*

Tiedonhakupelissä ilmenneistä teknisistä ongelmista ei puhuttu mitään 40 esseevastauksessa. Ohjelman toiminnassa ilmenneet hankaluudet mainittiin 4 kertaa. Eniten ihmettelyä herätti palautemuotojen erilainen toiminta sali- ja verkkoharjoituksissa. Syyksi epäiltiin hakuohjelman asetuksia tai hakujärjestelmän erityisongelmia. Tiedonhakupelin eri palautemuotojen puutteellinen toiminta verkkoharjoituksissa tuotiin esiin 20 vastauksessa. Saanti- ja tarkkuusprosenttien, relevanssipalkin tai vertailukäyrän vuoroittainen näkyminen ja näkymättömyys hakupalautteessa oli tiedonhakupelin yleisimmin havaittu tekninen ongelma, joka vei pohjaa eri hakujärjestelmillä toteutettujen hakujen tuloksellisuuden vertailulta. TRIPissä oli sovelluksen suunnitteluvirhe, joka liittyi hakulausekkeessa esiintyvien yhdyssanojen pilkkomiseen ja yhteenkirjoittamiseen. Ongelmallista oli se, ettei molempia hakuohjelmia voinut pitää auki samanaikaisesti. Pohdittiin, miksi ohjelma poimii hakutuloksen kärkeen dokumentteja, jotka eivät liity haettuun aiheeseen. Osittaistäsmäyttävä InQuery-hakuohjelma pudotti kyselystä pois juuri lisättyjä hakusanoja, jolloin sama hakulauseke saattoi toistua useita kertoja peräkkäin.

7.1.4 *Parannusehdotuksia tiedonhakupeliin*

Tiedonhakupelissä on kiintoisaa se, että hakua voi rakentaa vähitellen. Samalla näkee, miten pienetkin muutokset vaikuttavat hakutulokseen. Tiedonhakupeliin esitettiin parannusehdotuksia kaikkiaan 17 esseevastauksessa. Peliin toivottiin 11 vastauksessa vihjejärjestelmää, joka edesauttaisi paremman tarkkuuden saavuttamista hauissa. Neljä opiskelijaa ehdotti, että hyviä esimerkkihakuja voisi lisätä kaikkien nähtäväksi kunniatauluun. Jos dokumenttien keskeiset hakuavaimet näkyisivät selkeästi hakutuloksissa, niin hakulausekkeiden ideointi helpottuisi. Pelin jatkokehittelyä ajatellen olisi hyödyllistä tehdä näkyväksi myös se, miten ja millaisilla hakulausekkeilla hyvät hakutulokset on saatu. Itsenäisissä hakuharjoituksissa olisi eniten hyötyä tehtäväkohtaisista ohjeista. Automaattiset lisävinkit ovat tarpeen erityisesti silloin, kun hakutulos useista yrityksistä huolimatta pysyy saanti- ja tarkkuusarvoiltaan suunnilleen samana. Mikäli haku ei tuota tulosta, tiedonhakupelin ohjelma voisi opastaa hakijaa mielekkääseen suuntaan. Eräs vastaaja ihmetteli, mihin tiedonhakupeliä koskevat omat pohdinnat olisi pitänyt kirjata. Tiedonhakupelin oheen ehdotettiinkin lisättäväksi palautelomake, joka olisi käytettävissä harjoitusten aikana. Näin eivät mieleen tulevat asiat pääsisi unohtumaan. Pelin ohjeisiin toivottiin listausta

hakuavainten katkaisumerkeistä ja käytössä olevista operaattoreista. TRIPiin ehdotettiin toimintoa, joka lisää hakulausekkeeseen automaattisesti sulkumerkit. Tiedonhakupelin aineistojen ja hakutehtävien monipuolistamista esitettiin myös, koska samojen aiheiden jatkuvaa työstämistä pidettiin uuvuttavana.

Tiedonhakupelin ja sen palautetoimintojen ohjeilta odotettiin ennen kaikkea selkeyttä. Moni vastaaja toivoi hakujen tuloksellisuustietojen tarkistusmahdollisuutta joko kunniataulun tai hakuhistorian yhteyteen. Tiedonhakupelin asetukset tulisi selostaa käyttäjälle erityisen tarkasti. Ohjeessa kerrotaan, että peli on rajattu näyttämään 100 ensimmäistä tulosdokumenttia. Yhteenvetopalkki ilmoittaa kuitenkin löytyneiden dokumenttien määräksi 200+. Kun rajattu hakutulos käsittää 200 dokumenttia ja haun saanti- ja tarkkuusarvot lasketaan vain rajatulta alueelta, niin ihmeteltiin, ovatko tulokset lainkaan vertailukelpoisia?

Hakuhistorian käyttömahdollisuuksien laajentamista esitettiin useissa vastauksissa. Pohdittiin, onko mahdollista nähdä mistään vanhojen hakujen saanti- ja tarkkuusprosentteja. Hakuhistoria kun ei niitä verkkoharjoituksissa näyttänyt. Saanti- ja tarkkuuslukuja toivottiin näkyviin juuri hakuhistoriaan. Olisi hyvä, jos hakuhistoriasta voisi poimia suoraan jonkun haun tuloksia katsottavaksi ja muokattavaksi uutta hakua varten. Osittais- ja täystäsmäyttävän haun hakuhistoriaa ei voinut katsoa erikseen eikä vertailla eri hakujen tuloksia, koska siirtyminen hakuohjelmasta toiseen edellytti käytössä olevan ohjelman sulkemista ensin.

7.1.5 *Hakuavainten valinta*

Tiedonhakupelin siirtojen vaikutus näkyy välittömästi hakutuloksessa. Hakuharjoitusten tehtävät on ohjeistettu, mikä auttaa hakijaa pääsemään alkuun hakulausekkeiden muotoilussa. Hakijalle on annettu pääkäsitteet ja alustava hakustrategia, mutta hakuavainten ja niiden yhdistelmien valinta jää hakijan itsensä harkittavaksi. Opiskelijat ovat pohtineet esseeaineistossa hakuavainten valintaan liittyviä kysymyksiä useista näkökulmista. Kerronta etenee moniäänisenä. Hyvään hakutulokseen pääsee käyttämällä monia erilaisia hakuavaimia, joten hakutulosta kannattaa yrittää parantaa useampaan kertaan. Useilla yrityksillä ja haun uudelleenmuotoilulla saattaa päästä optimaaliseen tulokseen. Onko niin, että hyvä tarkkuus tuottaa minimaalisen hakutuloksen? Pitäisikö täystäsmäyttävissä TRIP-

hauissa tietää, mitä hakuavaimia tietokannan relevanteissa artikkeleissa ylipäätään on käytetty. Mistä mahtaa johtua, ettei hyvä hakulauseke toimikaan odotetulla tavalla? Muita hakukieleen liittyviä kysymyksiä ovat synonyymien käyttö ja määrä hakulausekkeessa sekä synonyymien käytöstä koituva mahdollinen hyöty tai haitta hakutuloksen kannalta. Tällaiset hakukielen ongelmat liittynevät osaltaan tietokannan käänteistiedostoon tallennettuihin dokumenteista poimittuihin hakuavainsanastoihin. Synonyymien käytöstä saatava hyöty on näennäinen, koska hakuohjelma ei pysty tunnistamaan synonyymien käsitteellistä vastaavuutta. Kun haussa on saavutettu tietty ”kylläisyysaste”, ei hakutulos enää parane, vaikka hakijan sanavarasto olisi miten laaja tahansa ja monet mahdolliset hakuavaimet kokeiltu hakulausekkeessa. Hakustrategian muuttaminen saattaa olla ratkaisu tilanteessa, jossa saturaatio on tapahtunut.

Noudatettu hakustrategia vaikuttaa hakutuloksen laatuun. Hyvän hakustrategian voi oppia laatimaan vain harjaantumalla. Jotkut hakuharjoituksiin osallistuneet opiskelijat ovat eritelleet tehtäväsuoritustensa vaiheita tarkasti. Vastaajat ovat pohtineet hakukielen toimivuutta yleensä, hakulausekkeiden muotoilua, fasettien käyttöä, hakuavainten muotoa sekä erilaisia tapoja laajentaa tai rajata hakua. Vastaajien pohdinnoissa korostuvat hakulausekkeen syntaksiin liittyvät kysymykset. Hakulausekkeen muotoilu lähtee liikkeelle hakupyynnön käsitteiden ja fasettien tunnistamisesta. Lingvistisellä tasolla käsitteiden merkitys riippuu niiden semanttisesta merkityskentästä.

Hakuavaimena toimii yksittäinen sana tai fraasi. Kun hakusuunnitelma on laadittu, se käännetään konelukuiseen muotoon merkkijonoksi haun suorittamista varten. Merkkijonotasolla hakuavaimet yhdistetään hakuoperaattoreiden avulla. Opiskelijat ovat pohtineet esseissään, miksi huolella muotoillut haut eivät aina toimikaan ongelmitta. Haun laajentaminen synonyymeja, assosiaatiotermejä ja fasetteja lisäämällä voi huonontaa hakutulosta merkittävästi. Kun hakuavainten tarkkuusvaatimus on hakijalle selkiytymätön ja Boolen operaattoreiden käyttö täystäsmäyttävässä TRIP-hakujärjestelmässä hankalaa, ei muotoiluyrityksistä ole sanottavasti hyötyä. Epärelevanttien dokumenttien poistaminen tulosjoukosta on osaltaan aiheuttanut päänvaivaa hakuharjoituksissa. Hakuavaimet eivät saa olla liian laajoja, mutta toisaalta hakuja ei pidä rajata liikaa. Esimerkiksi jollekin hakuaiheelle

ominaisen ammattikielen hakuavainten vaikutus hakutulokseen selviää vain kokeilemalla. Opiskelijoiden laatimissa esseevastauksissa on useita mainintoja osittaistäsmäyttävän InQuery-hakujärjestelmän helppokäyttöisyydestä ja sillä saavutettujen hakutulosten paremmuudesta hakujärjestelmien keskinäisessä vertailussa.

Vaikka InQueryssa hakuavainten lisääminen johtaa tulosten tarkkuuden nousuun, TRIPissä tarkkuus laskee. Huolella muotoilluissa hauissa esiintyy usein ongelmia. InQuery-hakujärjestelmää pidetään yleisesti TRIPiin verrattuna ”pitkämielisenä”, mutta esseevastauksissa on mainintoja myös InQueryn käytön yhteydessä ilmenneistä hankaluuksista. Eräs InQuery-hakujärjestelmän yhteydessä esseissä mainittu ongelma liittyi hakuavainten katkaisuun. Lokitiedostossa on nimittäin esimerkkejä sellaisista osittaistäsmäyttävistä hakulausekkeista, joissa on käytetty joko katkaisumerkkiä (#) tai asteriskia (*) hakuavaimen katkaisuun. Katkaisumerkki esiintyy hakuavaimen alussa tai lopussa liitettynä hakulausekkeen sanalistan yksittäiseen sanaan tai sanavartaloon. Tiedonhakupelissä ei osittaistäsmäytyksen hakusyntaksissa katkaisumerkkiä käytetä.

Tarkkuuden lisääminen ja epärelevanttien dokumenttien karsiminen näyttää olevan vaikeaa myös osittaistäsmäyttävässä hakujärjestelmässä. InQueryssa kannattaa silti kokeilla erilaisia hakutaktiikoita ja lähteä vasta sen jälkeen karsimaan huonoja termejä. Täystäsmäyttävän TRIPin ja Boolen operaattoreiden vaikeaselkoisuus on mainittu esseevastauksissa erikseen useita kertoja. Tarkkuuden saavuttaminen on hankalaa molemmissa hakujärjestelmissä. Mikä aiheuttaa saannin ja tarkkuuden laskun? Eräs tiedonhakupelin hakuharjoitusten osallistuja kertoo yrittäneensä parantaa hakutuloksensa tarkkuutta valitsemalla taktiikan, jossa hän etsi hakutulosten epärelevanteista dokumenteista hakuavaimia, joita ei ehkä esiintyisi relevanteissa dokumenteissa. Taktiikan tarkoituksena oli karsia hakutuloksesta epärelevantit dokumentit, mutta karsintamenetelmä osoittautui vaikeaksi toteuttaa. Hakija mainitseekin hakutulosten olleen heikot ja epäilee käyttämänsä taktiikan toimivan huonosti käytännön hakutehtävissä.

7.2 Opastuksen tarve ja hakuharjoituksissa esiin tulleet ongelmat käytettäessä eri hakujärjestelmiä

Kerronnat opastuksen tarpeesta ja hakuharjoituksissa esiintyneistä ongelmista näyttävät limittyvän toisiinsa opiskelijoiden vastauksissa. Vaikka useat opiskelijat mainitsivat, etteivät olleet hakeneet eivätkä saaneet ulkopuolista apua itsenäisten hakutehtävien suorittamisessa, monet heistä kertoivat saliharjoitusten yhteydessä saaneensa opastusta ja vinkkejä ohjaajalta tai kurssikaverilta. Opastusta pidettiin tarpeellisena ja hyödyllisenä. Useat opiskelijat kertoivat myös, ettei heillä ollut varsinaisesti ollut ongelmia tiedonhakupelin kanssa. TRIP-hakujen teko saatettiin kokea hankalaksi täystäsmäytyvyyden vaatimuksen vuoksi. Vastauksissa oli joitain mainintoja InQuery-hakujärjestelmän paremmuudesta, koska sillä pääsi nopeammin hyviin hakutuloksiin. Eräs opiskelija toivoi, että TRIPin ja InQueryn hakujärjestelmäkohtaisista eroista kerrottaisiin selkeästi jo saliharjoitusten yhteydessä. Joissain vastauksissa ihmeteltiin, pääseekö TRIP-haulla kunniatauluun tai onko yleensä mahdollista tehdä täydellinen haku, jossa kaikki hakuaiheeseen liittyvät relevantit dokumentit olisivat mukana hakutuloksessa. Eräs vastaajista halusi muistuttaa, ettei kenellekään saa muodostua pakkomielteeksi optimaalisen hakutuloksen saavuttaminen tiedonhakupelillä. Edelleen kyseltiin, kuka auttaa hakutehtävän analysoinnissa ja onko onnistuminen kiinni sattumasta.

Omien virheiden paikantaminen saattaa olla vaikeaa tiedonhakupelin palautejärjestelmästä huolimatta. Mikäli ohjaajan henkilökohtaista opastusta ei ole saatavilla, hakijat kokeilevat hakulausekkeiden toimivuutta oman päättelynsä varassa. Jotkut hakijat ovat selonteoissaan esittäneet parannuksia tiedonhakupeliin. Peliin on toivottu lisättävän vihjejärjestelmä, joka auttaa näkemään, mikä aiheuttaa haun epäonnistumisen ja miten hakulauseketta pitää korjata, jotta tietokannasta saadaan esiin relevantteja dokumentteja. Hakuistuntoihin osallistuneet hakijat näyttävät tuijottavan pelkästään prosenttilukuihin puhuessaan hakujensa saannista ja tarkkuudesta. Nämä hakijat eivät ole suhteuttaneet päätelmiään hakujensa tuloksellisuudesta niiden käsitteelliseen, lingvistiseen tai merkkijonotasoon eivätkä tietokantaan sisältyvien relevanttien dokumenttien määrään. Vaikka tehtävänannossa hakijoita kehotettiinkin kiinnittämään huomio hakukokeilujen saannin ja tarkkuuden

kehittymiseen, ei kaikkien hakijoiden käsitys hyvästä hakutuloksesta ole näin yksioikoinen.

Varsinkin täystäsmäyttävän TRIP-hakujärjestelmän käytön yhteydessä pidettiin kokeneemman hakijan antamia vinkkejä hyödyllisinä. Useat hakijat mainitsivat tiedonhakupelin käyttöä hankaloittavaksi tekijäksi hakujärjestelmässä ilmenevät tekniset ongelmat ja pohtivat niiden syitä. Tekniset ongelmat liittyivät yleisimmin tiedonhakupelin palautejärjestelmän toimintoihin. Esseevastauksissa pohdittiin myös sitä, kuka auttaa turhautuvaa hakijaa ideoiden loppuessa. Muutamat hakijat toivoivat hakujärjestelmältä automaattisia lisävinkkejä silloin, kun hakutehtävän suorittaminen juuttuu paikoilleen. Hakijat tunnistavat yleensä ohjaajan avun tarpeellisuuden arvioidessaan käyttämiensä hakuavainten hyvyyttä ja toimivuutta yksittäisissä hakulausekkeissa. Hyvän hakulausekkeen oppii muotoilemaan, mikäli siihen opastetaan perusteellisesti. Ohjausta tarvitaan, jotta hakujärjestelmän ydin ja hakuavainten käyttäytyminen hakujärjestelmässä voidaan ymmärtää paremmin. Mikäli oma ajattelu pyörii totuttua kehää, kannattaa saliharjoituksissa kysyä neuvoa toiselta, sillä näin voidaan saada hakutehtävän ratkaisemiseen mukaan uusia näkökulmia. Osa ongelmista liittyy joka tapauksessa hakijan ajatuskulkuihin. Hakijoiden pohdinnoissa mainittiin, että hakulauseketta muotoillessa on vaikeaa olla luova, koska hakijalle jäävä liikkumavara on tiedonhakupelissä rajallinen.

Hakijoiden tiedonhakupeliä kohtaan osoittama kritiikki on välillä voimakkaasti painotettua. Erään hakijan mukaan asiat tulisi vääntää noviisille rautalangasta. Sanomalehtikielen koetaan rajoittavan hakuavainten ideointia. Sanojen katkaisu ja ryhmittely aiheuttavat ongelmia hakutehtävässä kuten harvinaisemmat operaattoritkin. Ovatko kaikki hakuavaimet perusmuotoisia? Milloin hakuavaimet tulee katkaista? Miten fasetteja käytetään järkevästi? Kun hakijan toiminta tapahtuu kokeiluperiaatteella eikä tulosta synny, saattaisi jonkin yleispätevän vihjetoiminnon sisällyttäminen tiedonhakupeliin olla hakijalle hyvä kannustin. Hakuhistorian käyttöominaisuuksia toivottiin laajennettavan niin, että sen kautta voitaisiin vertailla tehtyjä hakuja. Tiedonhakupeliin toivottiin myös mahdollisuutta jo suoritetun hakulausekkeen joustavaan editointiin niin, ettei kaikkia haun vaiheita tarvitse käydä erikseen läpi toistamiseen. Palautejärjestelmän toimintojen tulkintakin tuotti ongelmia: hakija ihmetteli, ilmaisevatko näytön prosenttiluvut haun saantia vai tarkkuutta.

Kriittisimpiin tiedonhakupeliä koskeviin arvioihin liittyi kommentti, jonka mukaan tiedonhakupelin ainoa haaste hakijan näkökulmasta oli pysyä perillä siitä, mitkä haut olivat tuottaneet parhaat tulokset. Kysyttiin, eikö tiedonhaun menetelmien omaksuminen ja niiden soveltaminen ole sinänsä merkittävämpää kuin yksittäiseen hakuun käytetty aika tai hakujen tuloksellisuus. Mikä on aikaelementin merkitys hakuharjoituksissa? Onko aikaelementillä itseisarvo? Vaaditaanko tiedonhakupelissä omien hakujen vertaamista muiden hakijoiden tuloksiin? Mikä hakijalta jää tajuamatta, kun haku epäonnistuu? Erityisen vaikeilta tuntuivat hakijan mielestä harjoitukset, joissa piti parantaa sekä hakutuloksen saantia että tarkkuutta samanaikaisesti. Vaikka perusmuotoisessa hakemistossa hakuavainten katkaisulla ei olekaan kovin suurta merkitystä, voitaisiin tiedonhakupelin ohjeisiin listata sekä katkaisumerkit että käytössä olevat operaattorit. Eräs hakijoista ehdotti, että tiedonhakupeliin voisi sisällyttää malleja ”huippulausekkeista”. Pohdittiin, olisiko saliharjoituksissa käytävistä opiskelijoiden keskinäisistä keskusteluista ja parityöskentelystä apua hakulausekkeiden muotoilussa? Vaikka hakujen konkreettinen muoto ei olisikaan ongelma, saattaa hyvien hakuehtojen keksiminen olla hankalaa, koska tiedonhakupelin ja hakuaiheen taustalla olevan ajatusmaailman hahmottaminen on vaikeaa. Yleistiedosta on hyötyä hakutehtävissä, mutta kokeilu on aina työlästä.

7.3 Kuvaus erään hakuprosessin etenemisestä

Vaihtoehtoisia hakusuunnitelmia käyttämällä voidaan hakuaihekohtaisesti päästä lähes optimaaliseen hakutulokseen. Haun suunnittelussa on mahdollista yhdistellä useita erilaisia fasetteja, jolloin myös yksittäiseen fasettiin voidaan sisällyttää erilaisia toimivia hakuavainten yhdistelmiä. Hakusuunnitelma on lähes optimaalinen silloin, kun sen tuottaman hakutuloksen keskitarkkuus on korkeintaan 10 % parhaan tunnetun kyselyn keskitarkkuutta heikompi. Esseevastauksissa on hakijoiden yksityiskohtaisia kuvauksia siitä, miten heidän hakutuloksensa ovat kehittyneet, kun he ovat etsineet tiedonhakupelin avulla harjoitustietokannasta relevantteja aihedokumentteja. Seuraavassa esseevastauksista poimitussa hakijan omassa kalaruoka-aiheisessa esimerkissä on maininta hakutuloksen tarkkuudesta. Hakija on käyttänyt TRIPillä suorittamassaan haussa samoja termejä kuin InQuery-haussa. Hakutuloksessa on 77 dokumenttia, joista 67 on relevanttia. TRIP-haun saanti oli 0,96 ja tarkkuus 0,87, mikä

on tyypillisen hyvä harjoitustietokannan täsmähaun tulos. Rakenteinen TRIP-hakulauseke oli kirjoitettu muotoon

> (kala OR silli OR lohi OR silakka OR kuha OR ahven) AND (resepti OR ohje OR neuvo) AND (g OR dl OR rkl) AND (file OR annos OR pala) AND (suola OR pippuri OR tilli)

InQuery-hakulausekkeen rakenteeton sanalista

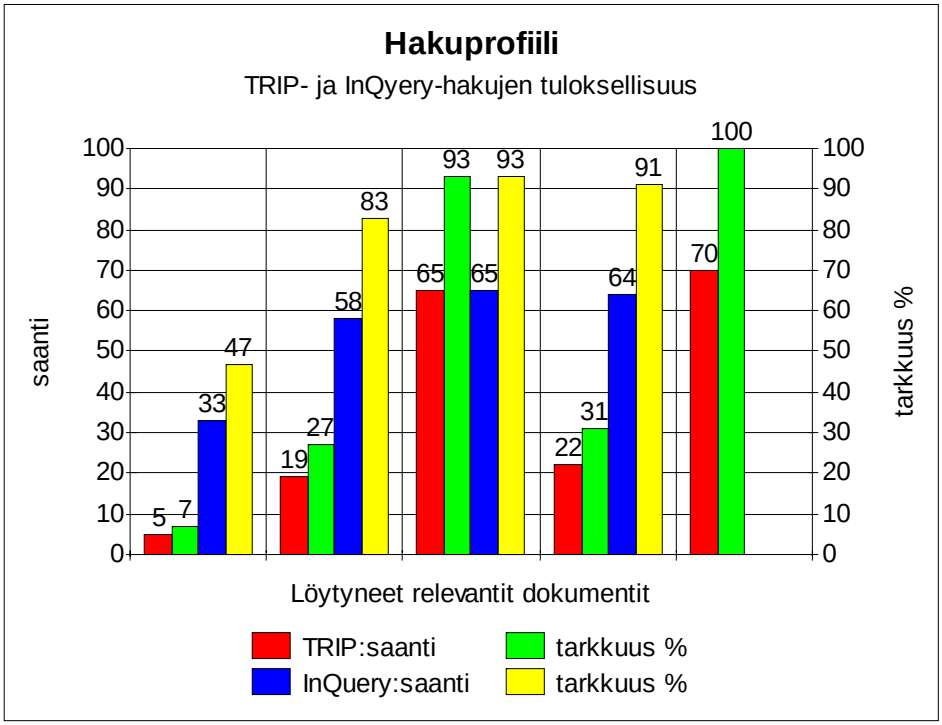
> #sum(kala lohi silakka kuha silli ahven resepti neuvo ohje g rkl dl pippuri suola tilli annos file pala)

tuotti tuloksen, jonka saanti oli 0,96 tarkkuuden ollessa 0,34. Molemmissa hakulausekkeissa on käytetty viittä ankkuritermiä, kuutta hyvää hakutermiä ja viittä kapea-alaista termiä. Yksi hakuavaimena käytetty sana kuuluu kaatoluokkaan ”muut”. Hakuavaintyyppien nimitykset perustuvat hakuaiheen hakuavaimista tehtyyn laatuluokitukseen, joka myös kuuluu tässä työssä käytettävään aineistoon. Pitäydyn tutkielmassani hakuavaintyyppien laatuluokituksessa käytetyissä nimityksissä. Mittayksiköiden lyhenteitä lukuun ottamatta kaikki hakulausekkeen hakuavaimet ovat perusmuotoisia. Ankkuritermejä ovat ’kala’, ’suola’, ’pippuri’ ja mittayksiköiden lyhenteet ’g’ ja ’rkl’. Hyviä termejä, joiden avulla tavoitetaan tietokannasta relevantteja dokumentteja ovat ’lohi’, ’silakka’, ’ohje’, ’annos’, ’tilli’ ja lyhenne ’dl’. Hakuavain ’pala’ edustaa monimerkityksisenä kaatoluokkaa.

Seuraavassa kuviossa on vertailtu yksittäisen hakijan tiedonhakupelin eri hakujärjestelmiä hyödyntämällä löytämien relevanttien dokumenttien lukumääriä ja hakujen tarkkuusprosentteja suhteessa saantikantaan. Esimerkkinä on hakutehtävä, jossa etsittiin Aamulehden ruokaosaston kalaruokareseptejä. Harjoitustietokantaan sisältyi kaikkiaan 70 reseptiä. Tässä esimerkissä ei ole mainintaa mahdollisista epärelevanteista tuloksista. Hakuyrityksiä on yhteensä yhdeksän. Niistä neljä tehtiin osittaistäsmäyttävällä InQuery-hakujärjestelmällä, kun taas viisi hakuyritystä toteutettiin käyttämällä täystäsmäyttävää TRIPiä. Hakuprofiili osoittaa useiden eri hakijoiden selonteoissaan mainitseman InQueryn helppokäyttöisyyden, selkeyden ja tuloksellisuuden TRIPiin verrattuna. Koska hakukielen syntaksi toimii erilailla osittais- ja täystäsmäyttävässä hakujärjestelmässä, tuottavat samat hakuelementtien yhdistelmät sekä saanniltaan että tarkkuudeltaan toisistaan poikkeavia tuloksia. Hakulausekkeesta puuttuva tai väärään kohtaan asetettu sulkumerkki vaikuttaa

heikentävästi hakutulokseen. Kun hakulausekkeeseen lisätään aiherelevantteja hakuavaimia, hakutulos saattaa kyllä parantua, muttei loputtomasti. Kun hakuavainten lisääminen ei enää vaikuta kohentavasti hakutulokseen, hakija esittää ratkaisuksi epätarkoituksenmukaisten hakuavainten karsimista hakulausekkeesta.

Sekä täys- että osittaistäsmäyttävää hakujärjestelmää käytettiin hakuprosessin edetessä hakijan harkinnan mukaan limittäin ja vuorotellen. Molemmissa vaihtoehdoissa lähdettiin liikkeelle samasta annetusta hakulausekkeesta. Hakuesimerkin avainfasetit ovat ’kala’ ja ’ruoka’. Täystäsmäytyksen esimerkkinä oli Boolen lauseke (kala OR lohi OR silakka OR kuha) AND ruoka. Osittaistäsmäytyksen aloitusesimerkkinä käytettiin hakulauseketta #sum(kala lohi silakka kuha ruoka).



Kuvio 1: Yksittäisen käyttäjän tiedonhakupelin eri hakujärjestelmillä suorittamien hakujen tuloksellisuuden vertailu

Kaikki huolimattomuusvirheet, kirjoitusvirheet ja hakulausekkeen syntaksissa esiintyvät virheet vaikuttavat hakutulokseen. Pylväskuviossa esitetyt täys- ja osittaistäsmäyttävällä järjestelmällä saadut tulokset ovat suuntaa antavia. Ne perustuvat yksittäisen hakijan esseevastauksessaan antamaan selontekoon. Esseevastaukset on kirjoitettu itsenäisesti suoritettujen verkkoharjoitusten yhteydessä. Verkkoharjoitusten keskeinen teema oli hakulausekkeen muotoileminen.

Opiskelijoiden tehtävä oli parantaa annettua hakulauseketta täys- ja osittaistäsmäyttävässä hakujärjestelmässä, verrata onnistuneiden hakujen tuloksellisuutta ja arvioida eri hakulauseketyyppien välisiä tuloksellisuuseroja. Opiskelijat pohtivat myös, mitä fasettien ja hakuavainten valinnassa kannattaa ottaa huomioon täys- ja osittaistäsmäyttävissä hauissa. Tiedonhakupeli on näyttänyt tässä vaiheessa vain 100 ensimmäistä tulosdokumenttia, joten visualisoinneissa näkyvät saanti- ja tarkkuusluvutkin on laskettu 100 dokumentin perusteella.

Ensimmäisessä täystäsmäyttävän haun pylväsparissa on nähtävissä esimerkki huolimattomuusvirheen vaikutuksesta hakutulokseen. Puuttuva sulkumerkki on aiheuttanut sen, että haun tuloksena löytyi vain 7 % tietokannan kaikista relevanteista dokumenteista. Kun puuttuva sulkumerkki lisätään esimerkkilausekkeeseen, niin oikeaksi saantiluvuksi saadaan TRIPillä 27 %, mikä on InQueryn tuottamaa tulosta heikompi. Lajinimiä, mittayksiköitä ja tarkoituksenmukaisia hakua rajaavia hakuavaimia lisäämällä täystäsmäytyksessä tulee saantiluvuksi 93 %. Kun yksityiskohtaisia hakuavaimia on liikaa, saanti laskee 31 %:iin. Tässä esimerkissä hakuavainten järkevä karsiminen on tuottanut yllättäen tulokseksi 100 % saannin. Kannattaa kuitenkin muistaa, että saannin ja tarkkuuden suhde pyrkii olemaan käänteinen (Järvelin, 2002). Kun saanti paranee, tarkkuus kääntyy yleisesti laskuun. Kun saantiluvut ovat erityisen korkeita, tarkkuus saattaa lähetä nollaa. Esimerkin saantikanta sisältää 70 dokumenttia, joista yhdellä jalostetulla hakulausekkeella on tavoitettu kaikki relevantit. Pieni saantikanta on toisin sanoen mahdollistanut täydellisen saannin. Yhdessä hakutehtävässä saatu täydellinen saanti ei silti riitä selittämään hakujen tuloksellisuutta.

Annetulla hakulausekkeella suoritetun haun tuloksellisuus osittaistäsmäyttävässä järjestelmässä on nähtävissä kuvion toisessa pylväsparissa. Kuviossa esitetyt luvut on poimittu esseevastauksesta. Saanti on ilmaistu kappalemäärinä ja prosenttilukuina hakujen suoritusjärjestyksen mukaisesti. InQuery tuotti annetulla hakulausekkeella saanniksi 47 %. Kun hakuavaimina käytettiin myös kalalajien nimiä ja ruoanvalmistuksessa yleisiä mittayksiköitä, saanti nousi 83 %:iin. Spesifien hakuavainten lisääminen – mausteet, ateriatyyppi – nosti saannin 93 %:iin. Viimeisessä haussa sovellettiin TRIPissä jo käytettyjä ja siinä hyvin toimineita

hakuavaimia. Kokeilut, joissa hakuavaimia lisättiin ja toisaalta karsittiin hyödyttömiä hakuavaimia tuottivat saantiluvuksi 91 %.

Lokitiedostosta näkyy, että hakijan kalaruoka-aiheesta tekemien parhaiden hakujen tarkkuusarvot olivat TRIPillä 0,528 ja InQuerylla 0,614. Seuraavassa taulukossa esitän vertailun hakijan molemmista hakujärjestelmistä tekemien parhaiden hakujen tuloksellisuustiedoista. Haut on luokiteltu tarkkuusarvojensa mukaan ja suhteutettu suurempaan tulosjoukkoon. Vertailuarvona on keskitarkkuus, joka on haun saantiin ja tarkkuuteen perustuva tuloksellisuuden mittari. Keskitarkkuus, joka on yksittäisten relevanttien dokumenttien kohdalla saatujen tarkkuuslukujen keskiarvo, on tässä aineistossa kaikkien syksyn 2003 perusopintojen tiedonhaun harjoituskurssiin osallistuneiden opiskelijoiden parhaiden hakujen tarkkuuslukujen tehtäväkohtaisesti laskettu keskiarvo. Hakuesimerkin hakija on epäonnistunut TRIP-haussa, sillä haun tarkkuus on jäänyt hänellä huomattavasti harjoitusryhmän kaikkien hakujen tehtäväkohtaisesti lasketun keskitarkkuusarvon alapuolelle. InQuery-haku on puolestaan onnistunut keskimääräistä paremmin. Seuraavassa taulukossa näkyy saman hakijan kahden hakuaiheen hakujen tuloksellisuus, jota verrataan koko harjoitusryhmän hakujen tuloksellisuuteen. On hyvä kuitenkin muistaa, että Boolen hakujen tarkkuus laskee tulosjoukon koon kasvaessa.

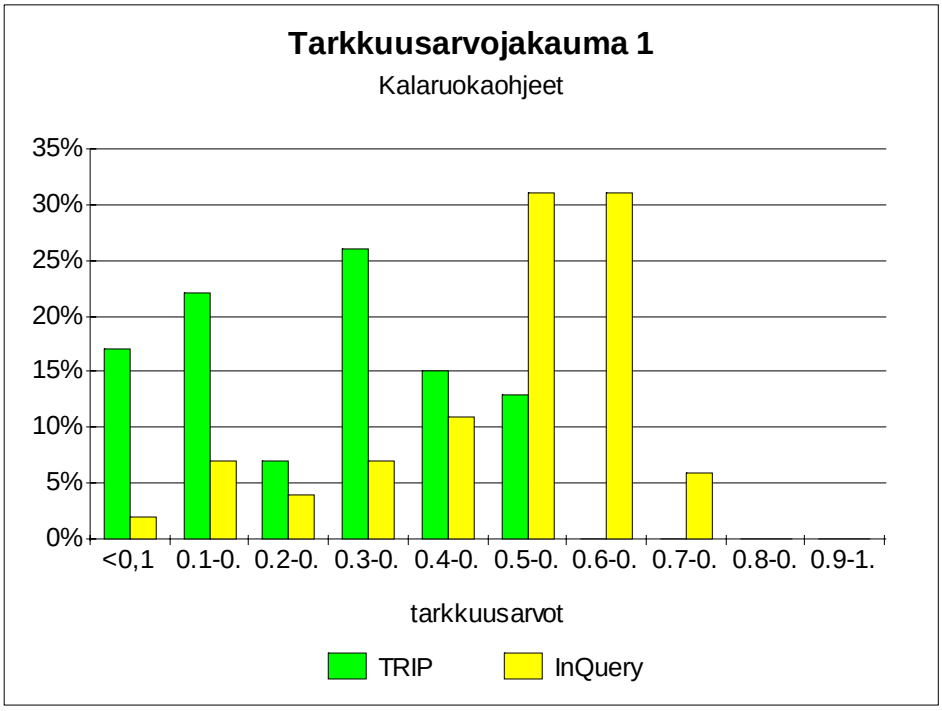
TRIP	Hakijan keskitarkkuus	Kaikkien keskitarkkuus	InQuery	Hakijan keskitarkkuus	Kaikkien keskitarkkuus
Kalaruoka	0,242	0,297		0,394	0,513
Etelä-Amerikka	0,025	0,151		0,073	0,12

Taulukko 5: Yhden hakijan ja harjoitusryhmän keskitarkkuudet kahdelle aiheelle täys- ja osittaistäsmäyttävässä hakujärjestelmässä.

Hakijan Etelä-Amerikan velkakriisi-aiheesta tekemien onnistuneimpien hakujen tarkkuus oli TRIPillä 0,043 ja InQuerylla 0,053. Vilkaisu lokitiedostoon osoittaa, että hakijalla oli tästä tehtävästä 4 hakuyritystä TRIPillä ja 5 InQuerylla. Hakuaiheen hankaluus käy ilmi paitsi hakijan parhaiden hakujen tarkkuusarvoista, myös kaikkien hakijan tekemien hakujen keskitarkkuudesta, joka on harjoitusryhmän tehtäväkohtaisten keskitarkkuusarvojen alapuolella. Tuloksellisuustietojen vertailu osoittaa, että vaikeasta hakuaiheesta tehdyt täystäsmäyttävät TRIP-haut ovat tarkkuudeltaan huonompia kuin osittaistäsmäyttävät InQuery-haut.

7.4 Onnistuneiden hakujen tuloksellisuus

Opiskelijoiden esseevastauksissaan yleisesti käsittelemästä hakuaiheesta on tallennettu kaikki hakulausekkeet ja tuloksellisuustiedot tiedonhakupelin lokitiedostoon. Esittelen yhteenvetona joitain seikkoja kalaruoka-aiheeseen liittyvien onnistuneiden hakujen tuloksellisuudesta. Samat tiedot näkyvät myös oheisesta pystypylväskuviosta. Yhteenvedon tarkoitus on edellä esitetyn, esseevastauksista satunnaisesti valitun yksittäisen hakuprofiilin sijoittaminen laajempaan viitekehykseensä. Syksyn 2003 tiedonhaun harjoituskurssin parhaat haut on eritelty lokitiedostosta hakuaiheittain omaksi tiedostokseen, jonka tehtävässä 2 on esitetty kalaruokaohjeisiin liittyvien hakujen tuloksellisuustiedot. Parhaiden hakujen tuloksellisuustiedoista näkyy, että kalaruoka-aiheesta on mukana kaikkiaan 54 opiskelijan onnistuneimmat TRIP- ja InQuery-haut. Tästä hakuaiheesta on tiedostossa yhteensä 268 haun tarkkuusarvot. TRIP-hakujen keskitarkkuus on tässä tehtävässä 0,297, kun taas InQuery-hakujen keskitarkkuus on 0,513. Kun verrataan aiheesta tehtyjen parhaiden hakujen tehtäväkohtaista keskitarkkuutta, niin TRIPillä saavutettu arvo oli 58 % InQueryn vastaavasta arvosta. Seitsemän TRIP-hakua saavutti tarkkuuden 0,5 - 0,6, mikä merkitsee 13 % parhaista TRIP-hauista. Kolme hakijaa ylsi InQuerylla 0,7 - 0,8 tarkkuuteen. Ainoastaan 6 % kaikista osittaistäsmäyttävistä hauista onnistui näin hyvin. Seuraavassa histogrammissa on nähtävissä harjoitusryhmässä TRIPillä ja InQuerylla tehtyjen hakujen jakautuminen, kun hakujen tuloksellisuuden mittarina on käytetty niiden tarkkuusarvoja. Kuviosta näkyy eri tarkkuusarvoja saavuttaneiden hakujen prosentuaalinen jakauma.



Kuvio 4: Harjoitusryhmän parhaiden hakujen tarkkuus

Olen analysoinut tutkielmani ensimmäisessä osuudessa syksyn 2003 tiedonhaun harjoituskurssin opiskelijoiden selontekoja kurssikokemuksistaan. Opiskelijat ovat pyrkineet hahmottamaan harjoitusten avulla TRIP- ja InQuery-hauissa käytettävien menetelmien eroja. Heidän lähtökohtanaan on ollut kulloiseenkin hakujärjestelmään soveltuvan hakustrategian löytäminen. Selonteot antavat hyvän ja monipuolisen kuvan tiedonhakupelin problematiikasta, hakustrategioiden suunnittelusta sekä niiden onnistuneesta tai epäonnistuneesta soveltamisesta tiedonhakupelissä. Samalla kun opiskelijat ovat kuvanneet verkossa suorittamiensa hakuharjoitusten kulkua ja toteuttamiensa hakujen tuloksellisuutta, he ovat esittäneet useita kommentteja tiedonhakupelistä ja tehneet pelin käytettävyyteen liittyviä parannusehdotuksia. Opiskelijoiden henkilökohtaiset selonteot sitovat tiedonhakupelin kontekstiin, jossa hakuharjoitusten lokitiedot on tallennettu.

8. Lokitiedostojen analyysi

Siirryn tarkastelemaan seuraavaksi tiedonhaun harjoituskurssin lokitiedostoja lähemmin. Lokitiedot mahdollistavat hakuprosessien kulun analysoinnin. Parhaiten onnistuneiden hakujen kriteerinä on tarkkuusarvo, kun hakulausekkeista poimitaan hakuavaimia fasettianalyysejä varten. Hakujen tarkkuuteen vaikuttavat yksittäisten

hakuavainten ominaisuudet näkyvät hauissa käytettyjen hakuavainten tehtäväkohtaisesti analysoiduista laatuluokituksista. Olen valinnut tarkasteltavakseni aineistosta Etelä-Amerikan velkakriisiin liittyvien hakutehtävien lokitiedot. Jotkut harjoituksiin osallistuneet opiskelijat pitivät tätä aihealuetta vaikeasti lähestyttävänä, joskin aiheesta on tehty monia hakuyrityksiä. Etelä-Amerikan velkakriisiin liittyvä hakuaihe korostui tiedonhakupelin harjoitusten vaikeimpana aiheena. Muutamissa opiskelijoiden esseissä leimattiin tämä tehtävä hakuharjoitusten negatiivisimpien kokemusten aiheuttajaksi. Hyvää hakutulosta oli erityisen vaikea saavuttaa, koska vajavainen aihetietämys hankaloitti toimivien hakulausekkeiden kehittelyä. Vaikka aikapula ei suosinutkaan aiheeseen tutustumista, dokumenttipalkki kuitenkin mahdollisti pääsyn artikkelitietokantaan. Tästä oli hyötyä hakulausekkeiden muotoilussa, sillä artikkeleista pyrittiin etsimään käyttökelpoisia, tehtävän ratkaisua helpottavia hakuavaimia. Käsittelen parhaiden täys- ja osittaistäsmäyttävien hakujen tehokkuudessa ja rakenteissa havaittavia eroja. Tarkoitukseni on tutkia lokitietojen pohjalta hakuyritysten tehokkuusjakautumaa. Kartoitan samalla parhaista hauista hakuavainten käyttöön liittyviä ja hakujen tuloksellisuuteen vaikuttavia piirteitä.

Syksyn 2003 tiedonhaun harjoituskurssilla oli valinnaisena hakuaiheena tiedon etsiminen Etelä-Amerikan maiden taloutta uhkaavista tekijöistä. Tiedonhakupeliin on sisällytetty 51 aiherelevanttia dokumenttia. Vihjeeksi saliharjoituksiin osallistuneille hakijoille oli tehtäväsuorituksen alussa annettu talouskriisin vaikutusten ja tilanteen parannuskeinojen selvittäminen. Hakuavainten ideointia varten oli lueteltu joitain velkakriisiin liittyviä oireita: velan lyhennysongelmat, korkojen nousu, inflaatio, devalvaatio ja valuuttapako ulkomaille. Tavoitteena oli ensinnäkin hakujen saannin ja tarkkuuden seuranta. Opiskelijat tekivät hakuaiheesta käsiteanalyysin. He muodostivat aiheesta mahdollisimman monta fasettia sekä lisäsivät jokaiseen fasettiin mahdollisimman paljon käsitteitä ja hakuavamia. Haut suoritettiin käyttämällä peräkkäisten fasettien strategiaa. Haku aloitettiin fasetilla, jonka avulla oletettiin tavoitettavan mahdollisimman suuri saanti. Hakutehtävän avainfasetit ovat ’Etelä-Amerikka’ ja ’ongelmat’. Fasetteja lisättiin yksi kerrallaan ja samalla seurattiin lisättyjen fasettien vaikutusta hakutulokseen. Saliharjoitusten puitteissa pyrittiin selvittämään saannin kasvua ja sen kääntymistä laskuun. Haut toteutettiin käyttämällä vuorotellen sekä osittaistäsmäyttävää InQueryä että täystäsmäyttävää TRIP-hakujärjestelmää. Hakujärjestelmiä vaihtamalla haluttiin selvittää, miten hakutulos

kehittyy eri järjestelmiä käytettäessä. Milloin hakulauseke on hyvä? Millainen hakulauseke on toimiva? Voidaanko todella hyvät hakuavaimet tunnistaa parhaiden hakujen tarkkuuden perusteella?

Käytän esimerkkianalyysissa muutamien hakijoiden parhaiden hakujen joukosta valitsemiani mahdollisimman lyhyitä hakulausekkeita tarkkuusarvoineen. Tarkastelen hakulausekkeiden toimivuutta täys- ja osittaistäsmäyttävässä hakujärjestelmässä, kun käytössä on peräkkäisten fasettien strategia. Näissä hauissa tulosjoukon koko vaihtelee 35:stä 200:een hakujen tarkkuuden vaihdellessa 0,004:stä 0,435:een. Mukana on kaksi TRIPillä ja neljä InQuerylla tehtyä hakua. Käsittelen ensin täystäsmäyttävän haun esimerkit.

TRIP: täystäsmäyttävät hakulausekkeet	Tarkkuus
((Etelä-Amerik# OR latinalainen Amerik#) AND (velka# OR talous#))	0,083
(Etelä-Amerik# OR ((latinalai# AND Amerik#) AND velka#))	0,109

Taulukko 6a: Hakulausekkeiden toimivuus peräkkäisten fasettien strategialla

Edellä olevissa esimerkeissä on käytetty avainfasettia, jonka toimivuus hauissa on todennäköisesti hyvä. Avainfasetin odotetaan rajaavan hakua, muttei silti karsivan relevantteja dokumentteja hakutuloksesta. Ensimmäisessä hakulausekkeessa TRIPin hakuavaimena toimii vartalosta katkaistu pääkäsite ’Etelä-Amerik#’, joka on erisnimenä kapea hakuavain. Hakua on laajennettu katkaistulla rinnakkaiskäsitteestä muodostetulla hakuavaimella ’latinalainen Amerik#’. Tässä tapauksessa katkaisumerkin käyttö on turhaa tai vaihtoehtoisesti olisi ollut järkevää käyttää myös rinnakkaiskäsitteen ensimmäisestä osasta muodostetusta hakuavaimesta ’latinalainen’ katkaistua vartalomuotoa ’latinalai#’, kuten toisessa esimerkkilausekkeessa on tehty. Hakuavaimet on edelleen yhdistetty kahteen vaihtoehtoiseen perusmuotoiseen, mutta katkaisumerkillä varustettuun hakua laajentavaan hakuavaimeen ’velka#’ tai ’talous#’. Katkaisumerkin käytöllä on haluttu varmistaa, että hakujärjestelmä löytää myös dokumentit, joissa nuo hakuavaimet esiintyvät yhdyssanojen alkuosina. Hakulausekkeella saavutettu tarkkuusarvo on 0,083 tulosjoukon koon ollessa 35 dokumenttia. Toisella TRIP-hakujärjestelmässä käytetyllä hakulausekkeella on saatu tarkkuusarvoksi 0,109, kun tulosjoukon koko on 200 dokumenttia. Lausekkeen

hakuavaimena on hakuaiheen pääkäsite katkaistussa muodossa. Pääkäsitteelle rinnakkainen käsite on esitetty kahden hakuelementin yhdisteenä ’latinalai#’ ja ’Amerik#’, jossa molemmat osat on katkaistu sanavartalosta. Pääkäsite ja sen rinnakkaiskäsite on yhdistetty hakua laajentamaan ja-operaattorin avulla hyvään perusmuotoiseen hakuavaimeen ’velka#’, johon on liitetty katkaisumerkki. Näiden kahden hakutuloksen tarkkuusero on 0,026 suuremman hakutulosjoukon eduksi.

Seuraavassa on esimerkkejä InQuery-hakulausekkeiden syntaksin ominaisuuksista. Hakulausekkeissa on käytetty summaoperaattoria ja sulkumerkkien sisällä kahta hakuavainta, joista toinen on hakuaiheen pääkäsite tai sen osa ja toinen perusmuotoinen, hyvä, hakua rajaava käsite. Hakulausekkeet eivät ole mitenkään monimutkaisia. Niissä on käytetty samoja hakuavaimia kuin täystäsmäyttävissä fasettiperustaisissa hauissa. Ensimmäisessä hakulausekkeessa haun tarkkuus on 0,435 tulosjoukon koon ollessa 200 dokumenttia. Vaikka toista hakulauseketta on muunneltu niin, että hakua rajaavaan hakuavaimeen on liitetty katkaisumerkiksi asteriski, haun tarkkuusarvo pysyy ennallaan. Asteriski on web-tiedonhaun yleinen katkaisumerkki, mutta oheisessa osittaistäsmäyttävässä hakulausekkeessa sen käyttö on syntaksivirhe, joka tosin ei vaikuta hakutulokseen. Kahdessa jälkimmäisessä esimerkissä tarkkuusarvo 0,373 pysyy samana hakuavainten kirjoitusasusta riippumatta. Iso alkukirjain ja epälooginen pienen alkukirjaimen käyttö tai yhdysviiva maanosan nimessä eivät näytä vaikuttavan hakuohjelman kykyyn tunnistaa tietokannan relevantteja dokumentteja.

InQuery: osittaistäsmäyttävät hakulausekkeet	Tarkkuus
#sum(velka amerikka)	0,435
#sum(amerikka velka*)	0,435
#sum(etelä amerikka velka)	0,373
#sum(Etelä-amerikka velka)	0,373

Taulukko 6b: Hakulausekkeiden toimivuus peräkkäisten fasettien strategialla

Tarkkuusarvoista voi päätellä, missä määrin hakujen tuloksiin yleensä liittyy hälyä. Joka tapauksessa näistäkin esimerkeistä näkyy, että InQuerylla saadut hakutulokset ovat keskimäärin huomattavasti tarkempia kuin TRIP-hakujen tulokset.

Seuraavassa analyysissa on informaatiotutkimuksen perusopintojen syksyn 2003 tiedonhakukurssin harjoitusten lokitiedoista kummastakin hakujärjestelmästä 84 parasta hakulauseketta, jotka olen järjestänyt alenevan tarkkuuden mukaiseen järjestykseen. Osa-aineistossa on kaikkiaan 168 hakulauseketta. Yhteenvedossa näkyy täystäsmäyttävien hakujen tarkkuuden jakautuminen osa-aineistossa. Etelä-Amerikan velkakriisi -tehtävän onnistuneimpien täystäsmäyttävien hakujen keskitarkkuusarvo on osa-aineistossa 48 % onnistuneimpien osittaistäsmäyttävien hakujen keskitarkkuudesta. Keskitarkkuus on hakujen tuloksellisuuden mittari.

TRIP	Keski-tarkkuus	Tarkkuus	% tuloksista	Kyselyiden määrä
	0,151	< 0,2	75	65
		0,2 - 0,3	15	13
		0,3 - 0,4	6	5
		0,4 - 0,5	1	1
		0,5 <	0	0
Yhteensä			97	84

Taulukko 7a: Täystäsmäyttävien hakujen tarkkuuden jakautuminen harjoitusryhmän aineistossa verrattuna hauissa saavutettuun keskitarkkuuteen

Osa-aineiston hakujärjestelmäkohtaiset keskitarkkuusarvot ovat TRIPillä 0,151 ja InQuerylla 0,315. TRIP-hauissa tavoitettu keskitarkkuus 0,151 osoittaa, ettei täystäsmäyttävien hakujen tuloksellisuus ole harjoitusryhmässä kovin hyvä.

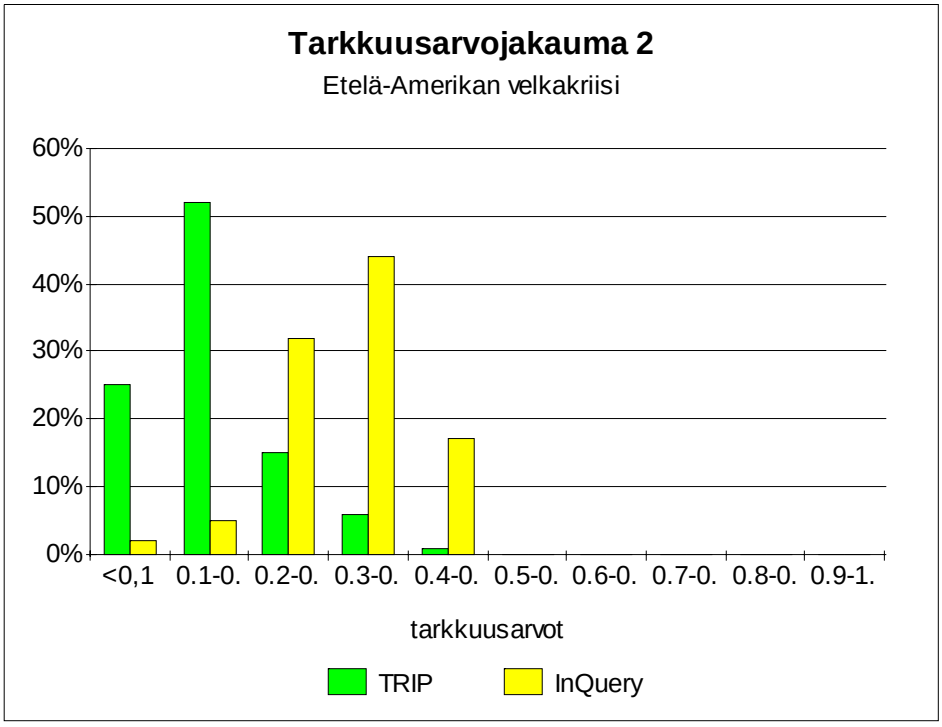
Täys- ja osittaistäsmäyttävien hakujen tarkkuustaso ei ole suoraan verrannollinen, koska hakutulosjoukon koko vaihtelee suuresti täystäsmättävien hakujen kohdalla. Tässä aineistossa aiheesta tehtyjen osittaistäsmäyttävien hakujen tarkkuus on suhteutettu hakutulosjoukkoon, jonka koko pysyy vakiona. Osittaistäsmäytyksessä tulosjoukon koko on 200. Osittaistäsmäyttävien hakujen keskitarkkuus on 0,315.

InQuery	Keskitarkkuus	Tarkkuus	% tuloksista	Kyselyiden määrä
		< 0,2	7	6
		0,2 – 0,3	32	27
	0,315	0,3 – 0,4	44	37
		0,4 – 0,5	17	14
		0,5 <	0	0
Yht.			100	84

Taulukko 7b: Osittaistäsmäyttävien hakujen tarkkuuden jakautuminen harjoitusryhmän aineistossa verrattuna hauissa saavutettuun keskitarkkuuteen

Hakulausekkeen pituus tai erilaisten hakuavainten valinta eivät sinänsä vaikuta kohentavasti hakutulokseen. Elleivät hakulausekkeen syntaksiin ja hakuavainten välisiin suhteisiin liittyvät asiat ole kohdallaan, ei haku toimi. Kun hakuaihe on vaikea, on täystäsmäyttävän haun toteuttaminen hankalaa. Osittaistäsmäyttävä hakumenetelmä on täystäsmäyttävää menetelmää joustavampi ja helpommin muunneltavissa, mikä vaikuttaa tuloksellisuuteen.

Opiskelijoiden esseevastauksissa oli mainittu Etelä-Amerikan velkakriisi -aiheisen hakutehtävän erityinen vaikeus. Lokitiedostoon tallentuneita osa-aineistosta tehtyjä hakuja on kaikkiaan 168. TRIP-hauista 52 % saavutti tarkkuuden 0,1 - 0,2 ja yhteensä 25 % jäi tarkkuusarvon 0,1 alapuolelle. InQuerylla paras saavutettu tarkkuus 44 %:lla hauista oli 0,3 - 0,4. Ainoastaan 17 % tehdyistä hauista ylsi 0,4 - 0,5 tarkkuuteen. Koko harjoitusryhmän parhaiden täys- ja osittaistäsmäyttävien hakujen tarkkuusarvojen prosentuaalinen jakauma näkyy seuraavassa histogrammissa.



Kuvio 5: Harjoitusryhmän parhaiden hakujen tarkkuus

8.1 Parhaiden hakulausekkeiden ominaisuudet

Kun hakua rajataan Boolen hakulausekkeissa ja-operaattorilla, niin kaikkien hakuavainten pitää sisältyä dokumenttiin, jotta tämä valikoituu hakutulokseen. Vähintään yhden tai-operaattorilla yhdistetyistä hakuavaimista on esiinnyttävä hakutulokseen valikoituvassa dokumentissa. Tai-operaattorilla yhdistetään saman fasetin eri käsitteet sekä yhden käsitteen erilaiset ilmiasut. Ei-operaattori sulkee pois kaikki dokumentit, joissa sen avulla yhdistetty hakuavain esiintyy. Sulkumerkit edesauttavat Boolen hakulausekkeen jäsentymistä ja ohjaavat hakulausekkeen prosessointia hakujärjestelmässä. Peräkkäisten fasettien strategiassa tuloksen tarkentuminen tapahtuu asteittain, kun hakulausekkeeseen lisätään rajaavia fasetteja. Tiedonhaun harjoituskurssilla tehdyt täystäsmäyttävät Boolen haut on toteutettu käyttämällä ja- sekä tai-operaattoreita, katkaisumerkkejä ja hakuavainten erilaisia kirjoitusasuja. Osittaistäsmäyttävissä InQueryn hakulausekkeissa näkyy sekä luonnollisen kielen ilmauksia että rakenteisia summaoperaattorilla yhdistettyjä hakuavainketjuja. Hakuavainten katkaisua näyttää esiintyvän joissain perusopintojen tiedonhakuharjoitusten osittaistäsmäyttävissä hakulausekkeissa. Kun molemmissa hakujärjestelmissä on käytetty samoja hakuavaimia, hakuavaimet on siirretty täystäsmäytyksen tarjoaman mallin mukaisesti katkaistuna osittaistäsmäyttäviin hakulausekkeisiin. Osittaistäsmäyttävissä hakulausekkeissa on käytetty useita todella pitkiä hakuavainketjuja. Hakuohjelma pystyy muokkaamaan luonnollisella kielellä ilmaistun hakulausekkeen osittaistäsmäyttävään hakujärjestelmään sopivaksi. InQueryn parhaiden hakulausekkeiden muotoilussa on turvauduttu ainoastaan summaoperaattoriin. Muita osittaistäsmäyttävän menetelmän operaattoreita ei ole käytetty hakuharjoitusten esimerkkilausekkeissa.

8.2 Hakulausekkeiden analyysi

Otan jatkossa tarkasteluun osa-aineiston 168 hakusuorituksesta 30 parhaaseen tarkkuuteen yltänyttä täystäsmäyttävää ja 30 parhaaseen tarkkuuteen yltänyttä osittaistäsmäyttävää hakua hakulausekkeineen ja tuloksellisuustietoineen. Hakulausekkeiden toimivuuden analysointi vaatii myös eri hakujärjestelmissä käytettyjen hakulausekkeiden sisällön purkamista auki. Tarkasteltavat 60 hakulauseketta ovat tutkielmani liitteissä 1 ja 2. Liitteessä 3 on esitetty mainittujen täys- ja osittaistäsmäyttävien hakujen tarkkuusarvojen vertailukäyrät ja

keskitarkkuusarvot. Olen luokitellut liitteissä olevat hakulausekkeet alenevan tarkkuuden mukaiseen järjestykseen. Tunnisteena on lokitiedoissa käytetty haun numero. Seuraavassa luokituksessa esitän 30 täystäsmäyttävän haun ankkuritermien analyysin. Ankkuritermien ominaisuuksiin kuuluu suuri esiintymistiheys, mikä samalla kohottaa niiden hakutehoa. Ankkuritermit ovat yleisiä käsitteitä, joita käyttämällä hauissa pyritään mahdollisimman suuriin saantilukuihin. Toisaalta ankkuritermeillä voidaan tavoittaa tekstitietokannan relevantteja dokumentteja, mutta toisaalta juuttuminen ankkuritermeihin saattaa estää vaikeasti tavoitettavien relevanttien dokumenttien löytymisen. Jotkut ankkuritermit toimivat täystäsmäytyksessä paremmin katkaistussa muodossa.

Hakutehtävän pääkäsitteet ovat ’Etelä-Amerikka’, ’ongelmat’ ja ’talous’. TRIP-hakulausekkeissa käytettiin kaikkiaan 133:a ankkuritermiä. Yleisin ankkuritermi oli ensimmäisen pääkäsitteen yhteydessä ’amerikka’, josta käytettiin hakulausekkeissa myös katkaistua vartalomuotoa ’amerik’. Katkaistu muoto laajentaa hakua täystäsmäytyksessä. Pääkäsitteen ’Etelä-Amerikka’ ohella ankkuritermi esiintyi yleisesti rinnakkaiskäsitteestä muodostetun hakuavaimen ’Latinalainen Amerikka’ yhteydessä. Hakulausekkeissa esiintyivät hakuavaimina myös ’Väli-Amerikka’ ja ’Keski-Amerikka’.

TRIP: 30 tarkinta hakua									
Haun nro	Käsitteet								
	Etelä-Amerikka		Ongelmat		Talous				
	Ankkuritermit								
	amerikka	brasilia	kriisi	#kriisi	kauppa	pankki	talous	#talous	Yhteensä
25673	2	1	2	0	0	0	2	0	7
25676	2	0	0	2	0	0	2	0	6
24909	2	1	0	2	0	0	1	0	6
26546	1	1	0	0	0	0	1	0	3
26550	1	1	0	0	0	0	0	0	2
26683	1	1	0	0	0	0	0	0	2
26150	1	1	0	1	0	0	0	1	4
25216	2	1	0	0	0	0	0	0	3
25567	1	1	0	0	0	0	0	1	3
25761	2	1	1	0	0	0	2	0	6
25941	2	1	0	0	1	0	1	1	6
26928	1	0	0	0	0	0	0	0	1
25406	2	1	0	0	0	0	1	0	4
25962	1	1	0	1	0	0	0	1	4
26944	2	1	0	0	0	0	1	0	4
26680	2	0	1	0	0	0	1	0	4
26667	2	1	0	0	0	0	0	0	3
24505	2	0	3	0	0	0	1	0	6
26682	2	1	0	0	0	0	0	0	3
25899	1	1	1	0	1	0	0	0	4
26168	1	1	1	0	0	2	1	0	6
24337	2	1	0	0	0	1	0	0	4
25054	1	1	1	0	0	0	1	0	4
26512	1	0	0	1	1	0	0	1	4
24420	2	1	1	0	0	1	2	0	7
25543	2	1	1	0	0	0	1	0	5
25836	2	1	2	0	0	0	0	0	5
24368	2	1	3	0	0	0	2	0	8
25694	2	1	2	0	0	0	3	0	8
26245	1	0	0	0	0	0	0	0	1
Yhteensä	48	24	19	7	3	4	23	5	133

Taulukko 8a: Ankkuritermien esiintyminen täystäsmäyttävissä hakulausekkeissa

Ankkuritermiä ’amerikka’ käytettiin täystäsmäyttävissä hakulausekkeissa kaikkiaan 48 kertaa. Yleisiä ankkuritermejä olivat ’brasilia’ ja ’talous’. Ankkuritermi ’brasilia’ oli käytössä 24:ssä ja ’talous’ 23:ssa lausekkeessa. Katkaistu ankkuritermi #talous esiintyi aineistossa 5 kertaa. Ankkuritermiä ’kriisi’ käytettiin perusmuodossa 19 kertaa. Katkaistuna #kriisi esiintyi 7 hakulausekkeessa. Täystäsmäytyksessä yleisimmät ankkuritermit ’amerikka’, ’brasilia’, ’kriisi’ ja ’talous’ esiintyivät hakulausekkeissa yhteensä 114 kertaa.

Hakuavaimista on käytettävissäni laajan hakuaineiston pohjalta laadittu valmis laadullinen luokitusanalyysi, jonka mukaisesti olen erotellut hakuavainten analyysia varten ankkuritermit. Pääkäsitteen ’Etelä-Amerikka’ alle kuuluvia ankkuritermejä ovat ’amerikka’ ja ’brasilia’. Pääkäsitteen ’ongelma’ alle kuuluu ankkuritermi ’kriisi’. Kolmanteen pääkäsitteeseen ’talous’ sisältyvät ankkuritermit ’kauppa’, ’korko’, ’pankki’, ’talous’ ja ’valuutta’. Kaikkia mainittuja ankkuritermejä on käytetty yleisesti tarkasteltavan osa-aineiston täys- ja osittaistäsmäyttävissä hakulausekkeissa.

Osittaistäsmäytyksessä ei sallita hakuavainten katkaisua samalla tavalla kuin täystäsmäytyksessä. InQueryn hakusyntaksi käsittää monipuolisen valikoiman hakuoperaattoreita, joita voidaan käyttää vaativassa tiedonhaussa. Tiedonhaun harjoituskurssilaiset käyttivät osittaistäsmäyttävissä hakulausekkeissa ainoastaan summaoperaattoria. Osittaistäsmäytyksessä hyödynnettiin samoja hakuavaimia kuin fasettiperustaisessa täystäsmäyttävässä haussa.

Seuraavasta taulukosta käy ilmi ankkuritermien jakautuminen InQueryn 30:ssä tarkimmassa hakulausekkeessa. Osa-aineiston InQuery-hakulausekkeissa käytettiin 117:ää ankkuritermiä. Pääkäsitteen ’Etelä-Amerikka’ ankkuritermeistä ’amerikka’ esiintyi hakulausekkeissa 36 kertaa, ’brasilia’ 14 kertaa. Pääkäsitteen ’ongelma’ yhteydessä käytetty ankkuritermi ’kriisi’ esiintyi tässä aineistossa 29 kertaa. Pääkäsitteen ’talous’ ankkuritermit ’kauppa’, ’korko’, ’pankki’, ’talous’ ja ’valuutta’ olivat käytössä myös osittaistäsmäyttävissä hakulausekkeissa. Ankkuritermin ’kauppa’ esiintymistiheys oli tutkitussa aineistossa 3, ankkuritermiä ’korko’ käytettiin 2 hakulausekkeessa. ’Pankki’ esiintyi 6 kertaa, ’talous’ 26 kertaa ja ’valuutta’ kerran osittaistäsmäyttävissä hakulausekkeissa. Yleisimmät osittaistäsmäyttävissä hakulausekkeissa esiintyneet ankkuritermit olivat ’amerikka’, ’kriisi’ ja ’talous’: ankkuritermejä käytettiin yhteensä 91 kertaa.

InQuery: 30 tarkinta hakua									
Haun nro	Käsitteet								
	Etelä-Amerikka		Ongelmat	Talous					
	Ankkuritermit								
	amerikka	brasilia	kriisi	kauppa	korko	pankki	talous	valuutta	Yhteensä
25328	2	1	1	0	1	3	1	0	9
26823	1	0	1	1	0	0	0	0	3
26829	1	1	1	0	0	0	1	0	4
25105	2	0	1	0	0	0	1	0	4
26648	1	0	0	0	0	0	0	0	1
25507	1	0	0	0	0	0	0	0	1
25821	2	0	2	1	0	0	4	0	9
26752	1	1	1	0	0	0	1	0	4
25230	1	1	1	0	0	0	1	0	4
26461	1	1	1	0	0	0	2	0	5
25905	1	1	1	0	0	1	2	0	6
24652	2	1	1	0	0	0	1	0	5
26967	1	0	0	0	0	0	1	0	2
26779	1	1	1	1	0	0	1	0	5
24934	1	1	1	0	0	0	1	0	4
24886	2	1	1	0	0	0	0	0	4
25003	1	0	1	0	0	0	0	0	2
26757	1	0	1	0	0	0	1	1	4
26024	1	0	1	0	0	0	0	0	2
26818	1	0	0	0	0	0	0	0	1
24259	1	0	0	0	0	0	4	0	5
26207	2	1	4	0	0	0	0	0	7
25076	1	0	1	0	0	0	1	0	3
26391	1	0	1	0	0	0	0	0	2
26824	1	0	1	0	0	0	1	0	3
25840	1	0	1	0	0	0	1	0	3
26627	1	0	2	0	0	0	1	0	4
24659	1	1	1	0	0	1	0	0	4
26367	1	1	0	0	0	0	0	0	2
26390	1	1	1	0	1	1	0	0	5
Yhteensä	36	14	29	3	2	6	26	1	117

Taulukko 8b: Ankkuritermien esiintyminen osittaistäsmäyttävissä hakulausekkeissa

8.2.1 *Avainfasetit*

Pääkäsite ’Etelä-Amerikka’ on hakutehtävän avainfasetti peräkkäisten fasettien strategiassa. Kun hakuharjoituksissa on lähdetty laajentamaan hakua, mainitun avainfasetin edustuksia on etsitty tekstitietokannasta mm. Etelä-Amerikan eri valtioiden nimillä. Olen poiminut seuraaviin taulukoihin harjoitusten osa-aineistosta täys- ja osittaistäsmäyttävien hakulausekkeiden hakuavaimina käytetyt valtioiden nimet sekä niiden hakuavainkohtaisen esiintymistiheyden. Hakuavaimet on järjestetty taulukoon laatuluokituksensa mukaan. Niiden esiintyminen hakulausekkeissa

perusmuodossa näkyy taulukossa laskevassa järjestyksessä. Hakuavainten esiintymät katkaistussa muodossa on lajiteltu erilliseen sarakkeeseen. Taulukoissa noudatetaan yleisenä järjestysperiaatteena hakuavainten laatuluokitusta ja esiintymistiheyttä hakulausekkeissa perusmuodossa. Kahdessa ensimmäisessä taulukossa analysoin täystäsmäyttävien hakujen laajentamista, kun uusina hakuavaimina ovat valtioiden nimet.

TRIP: 30 tarkinta hakua					
Haun laajentaminen valtioiden nimillä					
Avainfasetti: Etelä-Amerikka			Laatuluokitus	Perusmuoto	Katkaisu
Termi	brasilia		ankkuritermi	20	3
	argentiina		hyvä	17	2
	venezuela		kapea	21	1
	chile		kapea	18	2
	columbia/kolumbia		kapea	16	0
	peru		kapea	16	0
	bolivia		kapea	10	1
	uruguay		kapea	6	1
	paraguay		kapea	5	1
	ecuador/equador		kapea	3	0
Yhteensä				133	11

Taulukko 9a: Täystäsmäyttävän haun hakuavaimina valtioiden nimet

Taulukkoon 9a on koottu avainfasettia ’Etelä-Amerikka’ edustavien, hakuavaimina käytettyjen kymmenen valtion nimet. Valtioiden nimistä on käytetty hakulausekkeissa erilaisia kirjoitusasuja, jotka näkyvät hakuavainlistalla. Ankkuriterminä on ’brasilia’, jota on käytetty perusmuodossa 20 kertaa ja katkaistuna 3 kertaa täystäsmäyttävissä hakulausekkeissa. Hyvä hakuavain on perusmuodossa 17 ja katkaistuna 2 kertaa hakulausekkeissa esiintyvä ’argentiina’. Muut kahdeksan hakuavainta ovat hakua rajaavia kapeita termejä. Nämä valtioiden nimet näyttäytyvät täystäsmäyttävien hakujen hakuavaimina kaikkiaan 133 kertaa perusmuotoisina ja 11 kertaa katkaistussa muodossa.

TRIP: 30 tarkinta hakua					
Haun laajentaminen valtioiden nimillä					
Avainfasetti: Etelä-Amerikka					
Rinnakkaistermi: Latinalainen Amerikka					
RT: Keski-Amerikka /					
Väli-Amerikka			Laatuluokitus	Perusmuoto	Katkaisu
Termi		meksiko	kapea	18	2
		kuuba	kapea	4	0
		costa rica	kapea	1	0
		nicaragua	kapea	1	0
Yhteensä				24	2

Taulukko 9b: Täystäsmäyttävän haun hakuavaimina valtioiden nimet

Taulukosta 9b näkyy, miten täystäsmäytyksessä on käytetty valtioiden nimiä haun laajentamiseen lisäämällä hakulausekkeeseen rinnakkaistermejä sekä niitä edustavia vaihtoehtoisia käsitteitä. Avainfasetin ’Etelä-Amerikka’ ohella hakuavaimina on käytetty rinnakkaistermiä ’Latinalainen Amerikka’, jolle rinnakkaisia käsitteitä ovat ’Keski-Amerikka’ ja ’Väli-Amerikka’. Rinnakkaisilla käsitteillä hakualue on saatu laajennettua. Hakulausekkeisiin on edelleen lisätty rinnakkaistermeihin liittyviä vaihtoehtoisia käsitteitä, jolloin mukaan on saatu myös kuusi Keski- ja Väli-Amerikan valtioiden nimeä. Termien laatuluokituksessa lisätyt hakuavaimet edustavat kapeita termejä. Valtioiden nimet ’Meksiko’, ’Kuuba’, ’Costa Rica’ ja ’Nicaragua’ esiintyvät perusmuotoisina täystäsmäyttävissä hakulausekkeissa 24 kertaa ja katkaistuna 2 kertaa. Yleisimmin hakulausekkeissa käytetty hakuavain oli ’Meksiko’, joka esiintyi perusmuodossa 18 ja katkaistussa muodossa 2 kertaa. Maiden nimet on kirjoitettu hakulausekkeissa milloin pienillä, milloin suurilla alkukirjaimilla. Mainitulla seikalla ei tosin ole haun kannalta merkitystä. Täystäsmäytyksessä hakulausekkeen kaventaminen rajaavia käsitteitä lisäämällä parantaa haun tyhjentävyyttä. Haun kohdentaminen kapeiden termien avulla saa hakutuloksen tarkkuuden kohentumaan.

Taulukossa 9c esitetään osittaistäsmäyttävien hakujen laajentamisesta samoilla hakuavaimilla tehty jaottelu. Osittaistäsmäyttävien hakulausekkeiden muotoilussa käytetyt hakuavaimet perustuvat täystäsmäytyksen fasetteihin. Vaikka termien väliset suhteet eivät näy hakulausekkeissa osittaistäsmäytyksessä, niin peruskäsite ja sille rinnakkainen käsite vastaavat täystäsmäytyksessä käytettyjä käsitteitä. Avainfasetti on

’Etelä-Amerikka’ ja sen rinnakkaistermi ’Latinalainen Amerikka’. Hakulausekkeissa esiintyy kahdeksan eri valtion nimet, jotka jakautuvat hakuavainten laatuluokituksessa ankkuritermiksi, hyväksi hakutermiksi ja kuudeksi kapeaksi termiksi.

InQuery: 30 tarkinta hakua				
Kyselyn laajentaminen valtioiden nimillä				
Avainfasetti: Etelä-Amerikka			Laatuluokitus	Perusmuoto
RT: Latinalainen Amerikka				
Termi	brasilia		ankkuritermi	13
	argentiina		hyvä	7
	venezuela		kapea	8
		meksiko	kapea	8
	chile		kapea	5
	columbia/kolumbia		kapea	4
	peru		kapea	4
	bolivia		kapea	1
Yhteensä				50

Taulukko 9c: Osittaistäsmäyttävän haun hakuavaimina valtioiden nimet

Ankkuritermi ’brasilia’ esiintyy hakulausekkeissa 13 kertaa. Avainfasetin rinnakkaistermi ’Latinalainen Amerikka’ on antanut perusteen kapean hakuavaimen ’meksiko’ käytölle osittaistäsmäyksessä. ’Meksiko’ näyttäytyy hakuavaimena 8 kertaa. Perusmuotoisia valtioiden nimiä on käytetty osittaistäsmäyttävissä hakulausekkeissa kaikkiaan 50 kertaa. Haun laajentaminen tuottaa suuren tulosjoukon, mutta alentaa haun tyhjentävyyttä, kun taas kapeiden rajaavien termien avulla hakua voidaan kohdentaa ja tarkentaa.

Toinen hakulausekkeissa esiintyvistä avainfaseteista on ’ongelmat’. Haun laajentamiseen on käytetty avainfasettiin liittyviä assosiaatiotermejä – alakäsitteitä, jotka on laatuluokituksessa määritelty joko hyviksi tai kapeiksi hakuavaimiksi. Tarkimpien täystäsmäyttävien hakulausekkeiden hyvät hakuavaimet ovat ’velka’, ’ongelma’, ’talouskriisi’, ’inflaatio’ ja ’lasku’. Hakuavaimena on käytetty edellisten lisäksi kapeaa termiä ’laina’, joka esiintyy perusmuodossa 8 kertaa ja katkaistussa muodossa 2 kertaa. Avainfasettia edustavien hakuavainten käyttö on jakautunut täystäsmäyttävissä hakulausekkeissa taulukon 10a osoittamalla tavalla. Hakuavaimet näyttäytyvät hakulausekkeissa perusmuodossa yhteensä 73 kertaa ja katkaistuna 13

kertaa. Rajaavia hakuavaimia ei ole käytetty täystäsmäytyksessä tämän avainfasetin yhteydessä, mikä vähentää haun tyhjentävyyttä.

TRIP: 30 tarkinta hakua				
Haun laajentaminen				
Avainfasetti:	Ongelmat	Laatuluokitus	Perusmuoto	Katkaisu
Termi:	velka	hyvä	51	7
	ongelma	hyvä	6	2
	talouskriisi	hyvä	5	1
	inflaatio	hyvä	2	1
	lasku	hyvä	1	0
	laina	kapea	8	2
Yhteensä:			73	13

Taulukko 10a: Täystäsmäyttävän haun muita hakuavaimia

Samat laatuluokitukseltaan hyvät hakuavaimet kuuluvat avainfasetin ’ongelmat’ piiriin sekä täys- että osittaistäsmäyttävissä hakulausekkeissa. Osittaistäsmäyttävät haut perustuvat täystäsmäyttävien hakujen hakuavaimiin, mutta ovat silti rakenteeltaan heikkoja. Taulukkoon 10b on koottu näiden hakuavainten esiintyminen osittaistäsmäyttävissä hakulausekkeissa. Hakuavainten laskenta on suoritettu siten, että hakuavaimen on katsottu esiintyvän osittaistäsmäyttävässä hakulausekkeessa, mikäli yhdyssanan molemmat osat esiintyvät sen hakuavainketjussa.

InQuery: 30 tarkinta hakua			
Haun laajentaminen			
Avainfasetti:	Ongelmat	Laatuluokitus	Perusmuoto
Termi:	velka	hyvä	47
	ongelma	hyvä	21
	talouskriisi	hyvä	17
	inflaatio	hyvä	2
	laina	kapea	10
	luotto	kapea	1
	romahdus	kapea	1
Yhteensä:			99

Taulukko 10b: Osittaistäsmäyttävän haun muita hakuavaimia

Hakulausekkeiden rajaavat kapeat termit ovat ’laina’, ’luotto’ ja ’romahdus’. Tarkimpiin osittaistäsmäyttäviin hakulausekkeisiin sisältyy yhteensä 99 tarkasteltua avainfasettia edustavaa hakuavainta. Hakuavainten assosiatiivisen käytön avulla on pyritty kohentamaan hakujen tuloksellisuutta. Yleisin täys- ja osittaistäsmäyttävissä

hakulausekkeissa käytetty hakuavain tarkastellun avainfasetin yhteydessä on ’velka’. Sana esiintyy hakuavaimena täystäsmäytyksessä 58 kertaa ja osittaistäsmäytyksessä 47 kertaa.

8.3 Täys- ja osittaistäsmäyttävien hakulausekkeiden hyvin toimivat hakuavaimet

Millaiselta tutkitun osa-aineiston toimiva ja tehokkuutta tavoitteleva hakulauseke vaikuttaa? Miten hakuavaimia käytettiin hakulausekkeissa? Täystäsmäyttävät haut olivat kattavuudeltaan heikompia kuin osittaistäsmäyttävät haut. Molempien hakujärjestelmien 30:een tarkimpaan hakuun sisältyi ankkuritermejä 1 - 9 kappaletta hakulauseketta kohti. Hakuavainten määrä hakulausekkeessa vaihteli täystäsmäytyksessä kuudesta kahteenkymmeneen kappaleeseen. Osittaistäsmäytyksen 30 hakulausekkeessa käytettiin 2 - 39 hakuavainta. Käytettyjen hakuavainten määrästä näkee osittaistäsmäyttävien hakulausekkeiden olleen täystäsmäyttäviä pidempiä. Osittaistäsmäytyksessä pyritään tarkempiin hakutuloksiin käyttämällä pitkiä hakulausekkeita. Täystäsmäytyksen keskeiset hakutaktiikat ovat hakulausekkeen laajentaminen tai kaventaminen hakuoperaattoreiden avulla. Nämä hakutaktiikat vaikuttavat haun tarkkuuteen ja tyhjentävyyteen.

Haun laajuuden mittarina käytetään hakuavainten keskimääräistä lukumäärää fasettia kohden. Haun kattavuus tarkoittaa fasetointiin perustuvan hakulausekkeen muotoilussa käytettyjen fasettien lukumäärää. Aineiston parhaissa täystäsmäyttävissä hakulausekkeissa käytettyjen hakuavainten keskimääräinen lukumäärä oli 14,3 hakua kohti mediaanin ollessa 15. Yleisin hakuavainten määrä aineiston täystäsmäyttävissä hakulausekkeissa oli 18, mikä on jakauman moodi. Hakuaiheen pääkäsitteistä ’Etelä-Amerikka’ esiintyi hakuavaimena keskimäärin 1,07 kertaa, ’velka’ 2,4 kertaa osa-aineiston jokaisessa täystäsmäyttävässä hakulausekkeessa. Hakuavaimen ’kriisi’ hakulausekekohtainen frekvenssi oli osa-aineistossa 0,77. Ankkuritermejä esiintyi 30 täystäsmäyttävässä haussa 4,43 yksittäistä hakulauseketta kohti mediaanin ollessa 4. Hauissa painottuu ankkuritermien jokseenkin suuri määrä. Tulosjoukkoon valikoituneita hakuyrityksiä oli täystäsmäytyksessä keskimäärin 46,43. Tarkimpien täystäsmäyttävien hakujen keskitarkkuudeksi saatiin laskennassa 0,236. Hakujen tuloksellisuutta voidaan pitää kohtalaisena.

TRIP: 30 tarkinta hakua		Pääkäsitteet					
Haun nro	Hakuavainten					Tulosjoukon	Tarkkuus
	kokonaismäärä					koko	
		Etelä-Amerikka	velka	kriisi	Ankkuritermit		
25673	19	10	9	0	7	67	0,429
25676	12	1	5	2	6	30	0,343
24909	18	1	2	2	6	26	0,311
26546	18	1	3	0	3	35	0,311
26550	10	0	1	0	2	31	0,308
26683	6	0	1	0	2	27	0,3
26150	13	1	2	1	4	196	0,282
25216	9	1	1	0	3	38	0,253
25567	9	1	2	0	3	200	0,253
25761	16	1	3	1	6	41	0,244
25941	18	1	1	0	6	34	0,241
26928	7	0	1	0	1	26	0,241
25406	17	1	3	0	4	49	0,238
25962	10	0	1	1	4	22	0,236
26944	13	1	4	0	4	32	0,225
26680	16	1	4	1	4	37	0,223
26667	11	1	1	0	3	42	0,218
24505	16	1	5	3	6	23	0,201
26682	15	1	2	0	3	12	0,2
25899	19	0	1	1	4	104	0,198
26168	20	1	1	1	6	32	0,197
24337	16	1	1	0	4	19	0,197
25054	13	0	2	1	4	45	0,192
26512	15	0	1	1	4	27	0,187
24420	17	1	1	1	7	32	0,183
25543	12	1	5	1	5	12	0,181
25836	18	1	4	2	5	45	0,177
24368	17	1	2	2	8	24	0,173
25694	18	1	3	2	8	24	0,172
26245	11	1	1	0	1	61	0,168
Keskiarvo	14,3	1,07	2,4	0,77	4,43	46,43	0,236

Taulukko 11a: Hakuavainten keskimääräinen esiintymistiheys parhaissa täystäsmäyttävissä hakulausekkeissa ja hakujen keskitarkkuus

Seuraavassa taulukossa esitetään 30 tarkimman osittaistäsmäyttävän haun laajuutta ja tuloksellisuutta mittaavia arvoja. Käytettyjen hakuavainten keskimäärä hakua kohti oli 13,83 mediaanin ollessa 10, mikä luku jäi jonkin verran täystäsmäytyksen vertailuarvoa pienemmäksi. Myös aineiston osittaistäsmäyttävissä hakulausekkeissa käytettyjen hakuavainten yleisin määrä eli jakauman moodi oli 18. Avainfasetti ’Etelä-Amerikka’ oli osittaistäsmäytyksen tarkimpien hakujen hakuavaimena keskimäärin 0,67 tapauksessa. Fasettien ’velka’ ja ’kriisi’ hakuavainkohtainen

esiintymistiheys hakulausekkeissa oli keskimäärin 2,2 ja 1. Ankkuritermejä esiintyi kutakin hakulauseketta kohti keskimäärin 3,9 kappaletta mediaanin ollessa 4. Tarkimpien osittaistäsmäyttävien hakujen keskitarkkuus suhteessa tulosjoukon kokoon oli 0,396, mikä on selkeästi täystäsmäyttävien hakujen vertailuarvoa parempi. Osittaistäsmäytyksessä hakutulosjoukon vakioinen koko oli 200. Hakuharjoitusten parhaiden osittaistäsmäyttävien hakujen keskimääräinen tuloksellisuus oli näytön perusteella suhteellisen hyvä.

InQuery: 30 tarkinta hakua							
Haun nro	Hakuavainten	Pääkäsitteet				Tulosjoukon	Tarkkuus
	kokonaismäärä					koko	
		Etelä-Amerikka	velka	kriisi	Ankkuritermit		
25328	39	1	8	1	9	200	0,459
26823	11	1	1	1	3	200	0,452
26829	23	0	4	1	4	200	0,444
25105	17	1	1	1	4	200	0,437
26648	2	0	1	0	1	200	0,435
25507	2	0	1	0	1	200	0,435
25821	18	1	6	2	9	200	0,418
26752	8	0	1	1	4	200	0,417
25230	19	1	2	1	4	200	0,416
26461	34	1	2	1	5	200	0,411
25905	18	1	7	1	6	200	0,41
24652	27	1	4	1	5	200	0,41
26967	3	1	1	0	2	200	0,406
26779	18	0	1	2	5	200	0,404
24934	12	1	3	1	4	200	0,395
24886	14	1	1	1	4	200	0,387
25003	7	1	1	1	2	200	0,386
26757	11	0	3	1	4	200	0,382
26024	6	1	1	1	2	200	0,379
26818	3	1	1	0	1	200	0,373
24259	2	1	1	0	5	200	0,373
26207	25	1	4	4	7	200	0,369
25076	4	1	1	1	3	200	0,363
26391	9	0	1	1	2	200	0,363
26824	6	1	1	1	3	200	0,363
25840	8	0	1	1	3	200	0,362
26627	5	1	1	1	4	200	0,361
24659	37	0	2	1	4	200	0,36
26367	18	1	2	1	2	200	0,358
26390	9	0	2	1	5	200	0,355
Keskiarvo	13,83	0,67	2,2	1	3,9	200	0,396

Taulukko 11b: Hakuavainten keskimääräinen esiintyminen parhaissa osittaistäsmäyttävissä hakulausekkeissa ja hakujen keskitarkkuus

9. Keskustelu

Tässä työssä on keskitytty tiedonhaun harjoituskurssiin osallistuneiden opiskelijoiden muotoilemien tarkimpien hakulausekkeiden tuloksellisuuden ja hauissa käytettyjen hakuavainten analysointiin. Hakulausekkeiden yksityiskohtaiseen muotoanalyysiin ei ole puututtu, ei liioin käyttäjävirheiden analyysiin. Täystäsmäyttävien hakujen tarkkuus vaihtelee tutkitussa osa-aineistossa 0,168:sta 0,429:ään. Kahden kärjessä olevan hakutuloksen tarkkuusero on 0,086. Muiden aineiston hakujen väliset tarkkuuserot ovat pieniä – enimmäkseen ero jää alle 0,01. Täystäsmäytyksessä 30 parhaan haun tarkkuuksien vaihteluväli on 0,261. Osittaistäsmäyttävien hakujen tarkkuusarvot vaihtelevat 0,355 ja 0,459 välillä. Tutkitussa osa-aineistossa 30 osittaistäsmäyttävän haun tarkkuusarvojen vaihteluväli 0,104 on täystäsmäytykseen verrattuna huomattavasti kapeampi. Osittaistäsmäyttävät haut ovat yleisesti täystäsmäyttäviä tarkempia. Pääsääntöisesti hakujen tarkkuuserot vaihtelevat 0,001:stä 0,003:een. Ero kahden kärjessä olevan haun tarkkuudessa on 0,007. Parhaan täys- ja osittaistäsmäyttävän haun välinen tarkkuusero on 0,030. Yhteinen piirre molemmille hakutyypeille on se, että toisistaan poikkeavilla hakulausekkeilla on saatu harjoituksissa tuloksia, joiden tarkkuusarvot ryhmässä ovat täsmälleen samoja. Tutkittu osa-aineisto on suhteellisen pieni, joten näiden tulosten perusteella ei voida tehdä pitkälle vietyjä johtopäätöksiä.

Täystäsmäyttävissä TRIP-hauissa käytettiin hakutaktiikkana peräkkäisten fasettien strategiaa, jossa hakujen tarkentuminen tapahtuu asteittain hakuprosessin edetessä. Ensimmäisen fasetin valinta perustuu peräkkäisten fasettien strategiassa mahdollisimman suuren saannin odotukseen. Hakulausekkeen muotoilulla pyritään vaikuttamaan hakutuloksen tarkkuuteen. Dokumentti valikoituu hakutulokseen vain, mikäli kaikki täystäsmäyttävässä hakulausekkeessa ilmaistut ehdot täyttyvät. Vaikka InQueryn osittaistäsmäyttävissä hakulausekkeissa hakuavaimina käytettyjen käsitteiden väliset suhteet eivät näy, voidaan niissä kuitenkin säilyttää TRIPin alkuperäiset hakuavaimet. Haun laajentamiseen voidaan käyttää monia hakuavaimia. Osittaistäsmäytyksessä hakutulokseen valikoituvat todennäköisen relevanssinsa perusteella myös ne dokumentit, jotka vastaavat hakulausekkeessa ilmaistuja ehtoja osittain. Fasetointiin perustuva vahva rakenne on keino parantaa haun keskimääräistä

tarkkuutta osittaistäsmäytyksessä. Tässä aineistossa ei kuitenkaan ole esimerkkejä vahvaa fasettipohjaista rakennetta edustavasta osittaistäsmäyttävästä hakutyypistä.

Onko pienillä tarkkuuseroilla merkitystä keskusteltaessa hakujen tuloksellisuudesta yleensä? Jotta täystäsmäyttävillä hakulausekkeilla voitaisiin päästä samoihin tarkkuuksiin kuin osittaistäsmäytyksessä, pitäisi täystäsmäyttäviä TRIP-hakuja tehdä huomattavasti enemmän. Tarkkuuserot johtuvat paitsi hakujärjestelmän rakenteesta, myös TRIPin hakulausekkeille sallitusta rajatusta pituudesta. Yhteen TRIP-hakulausekkeeseen ei voida sisällyttää yhtä paljon hakuaiheen eri aspekteja ilmaisevia hakuavaimia kuin osittaistäsmäytyksessä. On vaikea kuvitella, että perusopintojen tiedonhakukurssin harjoituksissa tehdyt täystäsmäyttävät haut tuottaisivat harjoitustietokannoista tarkimpia mahdollisia tuloksia hetkessä. Pienet tarkkuuserot eivät ole sinänsä merkitseviä, kun hakujen määrä on suuri. Niiden vaikutus hakujen keskitarkkuuteen on vähäinen. Pienillä tarkkuuseroilla voi pikemminkin olla vertailuarvoa hakulausekkeen muotoiluun liittyvien monien mahdollisuuksien hahmottamisessa.

Tiedonhaun harjoituskurssista pyydettiin palautetta, jossa opiskelijat kertovat kokemuksiaan ja näkemyksiään kurssille osallistumisesta. Esseeaineiston analyysi on takautuvaa. Perinteinen haastattelumenetelmä pyrkii luotaamaan informantin käsityksiä tutkitusta aiheesta. Ääneen ajatteleminen on kirjallisten merkintöjen ohella yleinen tiedonhakuprosessin tutkimuksessa käytetty metodi. Tiedonhaun harjoituskurssin esseeaineistossa yhdistyvät haastattelumenetelmä, kurssipäiväkirja sekä ääneen ajatteleminen. Tiedonhaun tutkimuksen kysely- ja käyttäjäkeskeiset lähtökohdat tulevat selkeästi esiin harjoituskurssin hakulokien sisällöstä ja esseeaineiston kerronnasta. Aineistojen yhdistämisessä korostuu myös tiedonhaun harjoituskurssin prosessiluonne.

10. Loppulause

Tiedonhakupelin metaforat ovat peli, helppokäyttöisyys, selkeys, visuaalisuus ja havainnollisuus. Näistä elementeistä muodostuu se heijastuspinta, johon tiedonhaun harjoituskurssin esseeaineisto peilautuu. Tiedonhaun taustalla on ajatus kielipelistä,

joka toimii itselleen ominaisen logiikan mukaan. Erikoiskieli perustuu konventioihin ja noudattaa sääntöihin sidottua syntaksia. Tiedonhaun harjoituskurssin moniäänisestä esseeaineistosta voi löytää useita viittauksia täystäsmäyttävien hakulausekkeiden ongelmallisuuteen. Jotkut opiskelijat ovat kokeneet Boolen hakuoperaattoreiden käytön erityisen hankalaksi. Täystäsmäyttävän hakulausekkeen muotoilu itsessään edellyttää tarkkuutta ja Boolen logiikan ymmärtämistä. Kirjoitus- tai syntaksivirheet vaikuttavat välittömästi hakutulokseen. Osittaistäsmäyttävän hakulausekkeen erillisistä hakuavaimista koostuva lista poimii hakutulokseen parhaiten täsmäävät dokumentit vaatimatta kaikkien hakuehtojen täyttymistä. Opiskelijat ovat kuvanneet esseissään annetussa ympäristössä tapahtuvan ja annettua hakustrategiaa noudattavan hakuprosessin vaiheita monipuolisesti.

Tiedonhakupelin käsitteellä on sekä tekninen että sisällöllinen ulottuvuus. Tiedonhakupeliä on paljolti tutkittu opetussovelluksena ja oppimisympäristönä. Eräs tiedonhakupelin käyttötarkoituksista on tiedonhaun itsenäisen verkko-opiskelun mahdollistaminen. Tiedonhakupelin tutkimukseen on liittynyt paitsi järjestelmälähtöinen, myös vahva käyttäjälähtöinen painotus. Tiedonhakupelin käyttäjäystävällisyyttä on pyritty parantamaan käyttäjäpalautteen pohjalta. Esimerkiksi tiedonhakupeliin kehitetyn vihjetoiminnon prototyyppi on testattavissa verkossa käyttäjätunnuksin. Tiedonhaun harjoituskurssin opiskelijapalaute on osaltaan edesauttanut käyttäjänäkökulman hahmottamista.

Lokianalyysi sinänsä mahdollistaa hakuprosessin tutkimisen monista eri näkökulmista. Tutkijan käytettävissä on hakuprosessia koskevaa tilastollista tietoa, hakulausekkeet kirjautuvat lokeihin mahdollistaen hakusyntaksin tai siinä esiintyvien virheiden analysoinnin. Käyttötapahtumat ja käyttökonventiot voidaan ryhmitellä ja tulkita hakulokien perusteella. Hakujärjestelmän käytettävyyteen liittyviä asioita saadaan puolestaan esiin käyttäjäpalautteesta. Kysymyksenasettelusta riippuen käyttäjäpalautteesta saattaa välittyä tietoa myös hakijan henkilökohtaisista mentaalisista prosesseista.

Mikäli tiedonhakupelin käyttö opetus- ja oppimissovelluksena yleistyy, niin saatavilla on monenlaista hakulokeihin kirjautunutta tietoa eri tyyppisistä käyttöyhteyksistä ja -tilanteista. Tämä mahdollistaa sekä järjestelmä- että käyttäjälähtöiselle tutkimukselle

eri tyyppisten tutkimusasetelmien suunnittelun ja toteutuksen. Tutkimus voi olla kontekstisidonnaista toimintatutkimusta, hakukieleen ja hakusyntaksiin liittyvää tutkimusta tai hakujärjestelmän kehittämistä ja sen toimintojen testaamista laboratorio-olosuhteissa. Tiedonhakupeli on web-pohjainen sovellus tutkimus- ja opetuskäyttöön, mikä linkittää tiedonhakupelin web-tiedonhaun tutkimukseen.

Lähdeluettelo

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

Bates, Marcia J. (1987). How to use information search tactics online. Online (May): 47-54.

Blair, David C. & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3): 289 - 299.

Blair, David C. & Maron, M. E. (1990). Full-text information retrieval: further analysis and clarification. Information Processing & Management, 26(3): 437 - 447.

Blair, David C. (1996). STAIRS Redux: thoughts on the STAIRS evaluation, ten years after. Journal of the American Society for Information Science, 47(1): 4 - 22.

Borlund, Pia (2000). Evaluation of interactive information retrieval systems. Turku, Åbo Akademi University Press.

Hakala, J. (2002). METADATA: käyttökohteet ja formaatit. Teoksessa: Tietoverkot & kirjastot. Helsinki, BTJ Kirjastopalvelu.

Halttunen, K. (2000). Tiedonhaun peruskurssin opiskelijoiden kokemuksia IR-Game tiedonhakupelistä oppimisympäristön osana. Aikuiskasvatus 20(3): 201 - 214. Saatavilla www-muodossa:
<URL: http://www.info.uta.fi/tutkimus/fire/archive/ak2020003.pdf>.
(Luettu 06.03.2005).

Halttunen, K. (2004). Two information retrieval learning environments : their design and evaluation. Ph.D. dissertation. Acta Universitatis Tamperensis, vol. 1020. Saatavilla Pdf-muodossa.

Harter, S.P. (1990). Search term combination and retrieval overlap. A proposed methodology and case study. Journal of the American Society for Information Science, 41(2): 132 - 146.

InQuery -hakulausekkeiden syntaksi. Saatavilla www-muodossa:
<URL: http://ciir.cs.umass.edu/irdemo/inqinfo/inqueryhelp.html#general>.
(Luettu 06.03.2005).

Järvelin, K. & Kekäläinen, J. (2002). Tiedonhaun menetelmät -opintoaineisto. 4.2 Tiedontarpeen käsiteanalyysi. Hakusuunnitelman tyhjentävyys, tarkkuus ja kattavuus. Saatavilla www-muodossa: <URL: http://www.internetix.fi/opinnot/opintojaksot/0viestinta/informaatiotutkimus/po4>. (Luettu 06.03.2005).

Järvelin, K. & Sormunen, E. (1999). Dokumentit kateissa? Tiedon tallennus ja haku avuksi. Teoksessa: Tiedon tie: johdatusta informaatiotutkimukseen. Helsinki, BTJ Kirjastopalvelu.

Kekäläinen, J. (1999). The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Ph.D. dissertation. University of Tampere. Department of Information Studies. Saatavilla www-muodossa: <URL: http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf>. (Luettu 06.03.2005).

Kekäläinen, J. & Järvelin, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. Teoksessa: Proceedings of the CoLIS Conference, July 2002, Seattle, USA, 253 - 270. Saatavilla Pdf-muodossa.

Laakkonen, Mikko (2003). Tiedonhakupeli tietopalveluammattilaisille suunnatussa tiedonhaun täydennyskoulutuksessa. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu -tutkielma.

Lancaster, F.W., Rapport, Richard L. & Penry, J. Kiffin (1972). Evaluating the effectiveness of an on-line, natural language retrieval system. Inform. Stor. Retr., 8: 223 - 245.

Makkonen, Perttu (2002). Tiedonhakupeli ja tiedonhaun oppimisen arvioinnin ongelmat. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu -tutkielma.

Pennanen, Sami (2003). Tiedonhakupelin kontekstisidonnainen opetustoiminto – käyttäjävirheiden analyysi ja vihjetoiminnon alustava hahmotelma. Tampereen yliopisto. Informaatiotutkimuksen laitos. Pro gradu -tutkielma.

Pirkola, A. (1998). The effects of query structure and dictionary set-ups in dictionary-based cross-language information retrieval. Teoksessa: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR ´98), Melbourne, Australia, August 23 - 28, 1998. ACM Press, New York. Pp. 55 - 63.

Salton, G. (1970). Automatic text analysis. Science, 168, 3929, 335 - 343.

Salton, G. (1973). Recent studies in automatic text analysis and document retrieval. Journal of the ACM, 20(2), 258 - 278.

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

Saracevic, T. (1996). Relevance reconsidered ’96. Teoksessa: Proceedings of CoLIS 2. Copenhagen, Royal School of librarianship, 1996, 201 - 218.

Sormunen, E. (2001). Extensions to the STAIRS study – empirical evidence for the hypothesised ineffectiveness of Boolean queries in large full-text databases. Information Retrieval 4(3/4): 257 - 274.

Sormunen, E. (2002). A retrospective evaluation method for exact-match and best-match queries applying an interactive Query Performance Analyser. 24th BCS - IRSG European Colloquium on IR Research, March 25 - 27, 2002, Glasgow, Scotland, UK. Saatavilla www-muodossa:
<URL: http://www.info.uta.fi/tutkimus/fire/archive/BoolProb.pdf>.
(Luettu 06.03.2005).

Sormunen, E., Halttunen, K. & Keskustalo, H. (2002). Query Performance Analyser – a tool for bridging information retrieval research and instruction. University of Tampere. Department of Information Studies, RN 1(2002).

Sormunen, E. & Pennanen, S. (2004). The Challenge of automated tutoring in web-based learning environments for IR instruction. The 2003 Conference on Users in the Electronic Information Environments, September 8 - 9, Espoo, Finland. Verkkojulkaisussa: Information Research, 9(2) paper 169. Saatavilla www-muodossa: <URL: http://InformationR.net/ir/9-2/paper169.html>. (Luettu 06.03.2005).

Swanson, D.R. (1960). Searching natural language text by computer. Science, 132, 3434, 1099 - 1104.

Tenopir, C. (1985). Full Text Database Retrieval Performance. Online Review, 9:149 - 164.

Tenopir, C. & Shu, M.E. (1989). Magazines in full text: uses and search strategies. Online Review 13 (2), 107 - 118.

von Wright, G.H. (1982). Wittgensteinin myöhäisvaihe. Teoksessa: Logiikka, filosofia ja kieli. Ajattelijoita ja ajatussuuntia nykyajan filosofiassa, 228 - 244. Helsinki, Otava.

Liite 1
Informaatiotutkimuksen perusopintojen tiedonhaun harjoituskurssi/syksy 2003

TRIP: 30 tarkinta täystäsmäyttävää hakulauseketta. Jokaisen hakulausekkeen kohdalla on merkitty näkyviin hakulausekkeen tunniste, keskitarkkuus, tulosjoukon koko ja hakujärjestelmä.

25673 0,429 67 TRIP
((velka OR vela# OR velko# OR velkakriisi OR ulkomaanvelka OR ulkomaanvela# OR laina OR talouskriisi OR talouspolitiik#) AND (etelä-amerik# OR latinalai# amerik# OR brasilia OR argentiina OR chile OR bolivia OR peru OR venezuela OR columbia))

25676 0,343 30 TRIP
((Etelä-Amerik# OR Latinalai# Amerik#) AND (velkakriis# OR velkaantumi# OR velat OR velkataak# OR velka OR talousvaikeu# OR laina OR lainat OR talouskriis#))

24909 0,311 26 TRIP
((etelä-amerik# OR latinalai# amerik# OR chile OR brasilia OR peru OR uruguay OR paraguay OR bolivia OR kolumbia OR venezuela OR argentiina OR meksiko) AND (ulkomaanvel# OR laina#) AND (talouskriisi# OR velka# OR kriisi#))

26546 0,311 35 TRIP
((velka# OR ulkomaanvelka OR valtionvelka OR laina OR lainanotto OR maksukyky OR talousvaikeudet) AND (((etelä OR latinalainen) AND amerikka) OR brasilia OR argentiina OR bolivia OR equador OR chile OR kolumbia OR peru OR meksiko))

26550 0,308 31 TRIP
((latinalainen amerikka OR Brasilia OR Argentiina OR Meksiko OR Kolumbia OR Uruguay OR Paraguy OR Venezuela) AND velka#)

26683 0,300 27 TRIP
((latinalainen amerikka OR brasilia OR venezuela OR meksiko) AND velka#)

26150 0,282 196 TRIP
((Etelä-Ameri# OR latinalai# OR Brasil# OR Meksik# OR Argenti# OR Venez# OR Paragu# OR Urugu#) AND (velka# OR velan# OR laina# OR talousk# OR infl#))

25216 0,253 38 TRIP
((etelä-amerikka OR väli-amerikka OR latinalainen OR amerikka OR meksiko OR brasilia OR chile OR kuuba) AND #velka#)

25567 0,253 200 TRIP
((etelä-amer# OR kolumbia OR brasil# OR latinal# OR peru OR chil#) AND (velka# OR velan# OR talous#))

25761 0,244 41 TRIP
((etelä-amerikka OR (latinalainen AND amerikka) OR meksiko OR brasilia OR kolumbia OR chile OR peru) AND (velkakriisi OR talousongelma OR velka# OR talous OR laina OR velanhoito OR lyhentää OR takaisinmaksu))

25941 0,241 34 TRIP
((etelä-amerikka OR latinalainen amerikka OR brasilia OR venezuela OR peru OR argentiina OR meksiko OR chile OR uruguay OR kehitysmaat) AND velka# AND (talous# OR kansantalous OR takaisinmaksu OR öljy OR kauppa# OR teollisuus))

26928 0,241 26 TRIP
((etelä OR latinalainen OR amerikka OR meksiko OR chile OR venezuela) AND (velka#))

25406 0,238 49 TRIP
((Etelä-Amerikka OR chile OR brasilia OR kuuba OR argentiina OR meksiko OR väli-amerikka OR banaanivaltio OR tulimaa OR venezuela) AND (velkaantumisongelma# OR velka# OR velkataakka# OR ympäristöongelma# OR öljy OR köyhyys OR talousahdinko))

25962 0,236 22 TRIP
((velka# AND (ongelm# OR talous# OR kriisi#)) AND (Amerikka OR etelä OR latinalainen OR Brasilia OR Venezuela OR Meksiko))

26944 0,225 32 TRIP
((Etelä-Amerikka OR Latinalainen Amerikka OR Argentiina OR Meksiko OR Brasilia OR Peru) AND (velka# OR velkataakka OR velkaongelma OR talousongelma OR anteeksianto OR ulkomaanvelka))

26680 0,223 37 TRIP
((Etelä-Amerikka OR Latinalainen Amerikka OR Peru OR Bolivia OR Chile OR Guatemala OR Argentiina OR Venezuela OR Meksiko) AND (velkakriisi OR velka# OR velkaongelma OR ylivelkaantuminen OR talousvaikeu# OR vararikko))

26667 0,218 42 TRIP
((Etelä-amerik# OR latinalainen amerik# OR bolivia# OR chile# OR meksiko# OR brasilia# OR argentiina#) AND velka# AND (kehit# OR ratkais#))

24505 0,201 23 TRIP
((etelä-amerikka OR latinalainen amerikka OR meksiko OR chile OR kolumbia OR venezuela OR argentiina) AND (velkakriisi OR velkaantumisongelma OR talouskriisi OR velkataakka OR ulkomaanvelka OR öljykriisi OR inflaatio OR velanmaksu))

26682 0,200 12 TRIP
((Latinalai#en Amerik# OR Etelä-Amerik# OR Brasilia OR Chile OR Peru OR Guatemala OR Kolumbia OR Argentiina OR Venezuela OR Bolivia) AND (velka# OR budjettivaje) AND (velan# OR lainan#))

25899 0,198 104 TRIP
((maksuhäiriöt OR velka# OR kriisi OR maailmankauppa#) AND (Brasilia OR Ecuador OR Chile OR Peru OR Urugyau OR Argentiina OR Bolivia OR Kolumbia OR Meksiko OR Paraguay OR Venezuela OR etelä OR latinalainen OR amerikka OR kehitysma#))

26168 0,197 32 TRIP
((etelä-amerikka OR latinalainen OR meksiko OR peru OR bolivia OR brasilia OR argentiina OR venezuela OR kolumbia) AND (velka# OR talous OR kriisi OR avustus OR työttömyys OR pankki) AND (Usa OR japani OR yhdysvallat OR kehityspankki OR IMF))

24337 0,197 19 TRIP
((etelä amerikka OR latinalainen amerikka OR brasilia OR argentiina OR uruguay OR paraguay OR chile OR peru OR ecuador OR venezuela OR kolumbia) AND (maailmanpankki OR IMF OR velka))

25054 0,192 45 TRIP
((talous OR laina OR velka OR kriisi OR ulkomaanvelka OR öljy) AND (etelä OR latinalainen OR amerikka OR chile OR meksiko OR brasilia OR venezuela))

26512 0,187 27 TRIP
(((Etelä OR Amerikka OR latinalainen OR Meksiko OR Argentiina OR Kolumbia OR Chile OR Peru OR Bolivia) AND (talou# OR kriisi# OR kaup# OR ongelma# OR vapaa)) AND (velk#))

24420 0,183 32 TRIP
((etelä-amerikka OR latinalainen amerikka OR lattarimaat OR peru OR brasilia OR bolivia OR meksiko OR venezuela OR kolumbia) AND (velka OR laina OR talousongelmat OR maailmanpankki OR IMF OR kriisi OR talousvaikeudet))

25543 0,181 12 TRIP
((Etelä-Amerikka OR Latinalainen Amerikka OR Argentiina OR Venezuela OR Brasilia) AND (velkakriisi OR velkaantuminen OR velka OR velkaantumisongelma OR talousongelma OR ulkomaanvelka))

25836 0,177 45 TRIP
((etelä-Amerikka OR latinalainen amerikka OR brasilia OR kolumbia OR argentiina OR chile OR venezuela OR paraguay OR peru OR uruguay) AND (velka# OR velkakriisi OR velkaantumisongelma OR velkataakka OR kriisi OR ongelma OR köyhyys))

24368 0,173 24 TRIP
((etelä-amerikka OR latinalainen amerikka OR bolivia OR brasilia OR columbia OR chile OR Venezuela OR Costa Rica) AND (velkakriisi OR talousvaikeud# OR ulkomaanvelka OR mellakka OR inflaatio OR valtionyhti# OR öljykriisi OR talouskriisi))

25694 0,172 24 TRIP
((etelä-amerikka OR latinalainen amerikka OR lattarimaat OR latinomaat OR kolumbia OR argentiina OR peru OR kuuba OR brasilia OR chile) AND (velkakriisi OR velka# OR ylivelkaantuminen OR takaisinmaksu OR talousongelma OR talouskriisi OR talousvaikeus))

26245 0,168 61 TRIP
((Etelä-Amerikka OR Latinalainen OR Kuuba OR Nicaragua OR Venezuela OR Peru OR Meksiko) AND (velka# OR köyhyy# OR laina# OR maksu#))

Liite 2
Informaatiotutkimuksen perusopintojen tiedonhaun harjoituskurssi/syksy 2003

InQuery: 30 tarkinta osittaistäsmäyttävää hakulauseketta. Jokaisen hakulausekkeen kohdalla on merkitty näkyviin hakulausekkeen tunniste, keskitarkkuus, tulosjoukon koko ja hakujärjestelmä.

25328 0,459 200 InQuery
#sum(Etelä-Amerikka Latinalainen Amerikka Meksiko Venezuela Brasilia Chile Peru Kolumbia talous velkaantuminen velka ongelma kriisi ratkaisu ratkaista velkataakka velkaongelma myönnytys helpottaa helpotus kehityspankki maailmanpankki IMF velallinen velkainen hätäapulaina ulkomaanvelka velkoja korko öljyntuottaja vaikeus takaisinmaksu talousohjelma pankkipiiri kolmas maailma kehitysmaa laina)

26823 0,452 200 InQuery
#sum(etelä-amerikka velka ongelma kriisi ratkaisu köyhä latinalainen miljardi taakka laina kauppa)

26829 0,444 200 InQuery
#sum(etelä Amerikka latinalainen velkakriisi velka taakka loukko ongelma maksukyky laina kehitysmaa kolmas maailma rahoitussopeutusohjelma köyhä maa talousvaikeus ulkomaanvelka velkakakku bruttokansantuote Brasilia Argentiina Peru Kolumbia Meksiko)

25105 0,437 200 InQuery
#sum(Etelä-Amerikka velka kriisi ongelma kehitysmaa laina takaisinmaksu Latinalainen Amerikka öljy Venezuela talous vaikeudet taakka ympäristö suojelu köyhyys)

26648 0,435 200 InQuery
#sum(velka amerikka)

25507 0,435 200 InQuery
#sum(amerikka velka*)

25821 0,418 200 InQuery
#sum(etelä-amerikka latinalainen amerikka kehitysmaat velkakriisi velkaantumisongelma velkaantuminen velat velka ulkomaanvelka takaisinmaksu laina dollari talous talouskriisi kansantalous talousvaikeudet kauppa)

26752 0,417 200 InQuery
#sum(Etelä latinalainen Amerikka velka kriisi talous ongelma Brasilia laina)

25230 0,416 200 InQuery
#sum(Etelä-Amerikka Bolivia Argentiina Meksiko Brasilia Venezuela velka kriisi ongelma talous lama taakka ulkomaanvelka kehitys kehittyminen tausta syy ratkaisu ratkaiseminen)

26461 0,411 200 InQuery
#sum(siirtomaat maailmanpankki rahasto inflaatio laina ulkomainen velka ratkaisu ongelma vaalit kriisi Meksiko Venezuela Argentiina Brasilia Peru Latinalainen Etelä Amerikka ulkomainen velka ratkaisu ongelma hoito taakka talous kehitys maksukyky köyhyys tutkija tiede asiantuntija takaisinmaksu entiset)

25905 0,410 200 InQuery
#sum(etelä amerikka latinalainen velka kriisi velkaantuminen talous öljy velkaantuminen ongelma ulkomaanvelka brasilia venezuela velkojenhoito velanmaksu talousohjelma leikkaaminen velkakysymys)

24652 0,410 200 InQuery
#sum(Etelä-Amerikka Brasilia Argentiina Kolumbia Venezuela Chile velka kriisi velkaantuminen ongelma ratkaisu talouselämä latinalainen amerikka velanhoito talous romahdus öljy tulo maailmanpankki ulkomaanvelka laina kiista avustus hinnankorotus köyhä kehitysmaa taakka)

26967 0,406 200 InQuery
#sum(Etelä-Amerikka talous velka)

26779 0,404 200 InQuery
#sum(etelä latinalainen amerikka brasilia venezuela meksiko velka kriisi ongelma taakka velkaantuminen anteeksianto maksu kehitysmaa raha talous kauppa apu)

24934 0,395 200 InQuery
#sum(etelä amerikka latinalainen ongelma velka talous# kriisi köyhyys kehitysma# ulkomaanvel# ylivelkaantuminen brasilia)

24886 0,387 200 InQuery
#sum(#sum(Etelä-Amerikka or brasilia or venezuela or meksiko or chile) velka talou# kriisi ongelma ahdinko latinalainen amerikka kehitysma# laina)

25003 0,386 200 InQuery
#sum(etelä-amerikka latinalainen velka kriisi ongelma laina imf)

26757 0,382 200 InQuery
#sum(etelä latinalainen amerikka velka kriisi velan ongelma ylivelka kehitys ratkaisu hoito)

26024 0,379 200 InQuery
#sum(etelä amerikka latinalainen velka kriisi talous)

26818 0,373 200 InQuery
#sum(etelä amerikka velka)

24259 0,373 200 InQuery
#sum(Etelä-amerikka velka)

26207 0,369 200 InQuery
#sum(etelä-amerikka latinalainen amerikka argentiina brasilia chile meksiko peru velkakriisi talouskriisi velka talous valuuttakurssi öljykriisi lama talousvaikeudet velkataakka inflaatio ulkomaanvelka talousnäkymät kehitys avustus seuraukset finanssikriisi bruttokansantuote)

25076 0,363 200 InQuery
#sum(etelä-amerikka velka kriisi ongelma)

26391 0,363 200 InQuery
#sum(Amerikka etelä latinalainen velka talous politiikka näkymä ongelma kriisi)

26824 0,363 200 InQuery
#sum(velka Etelä Amerikka Latinalain kriisi ongelma)

26391 0,363 200 InQuery
#sum(Amerikka etelä latinalainen velka talous politiikka näkymä ongelma kriisi)

26824 0,363 200 InQuery
#sum(velka Etelä Amerikka Latinalain kriisi ongelma)

25840 0,362 200 InQuery
#sum(Etelä Latinalainen Amerikka velka kriisi talous finanssi ongelma)

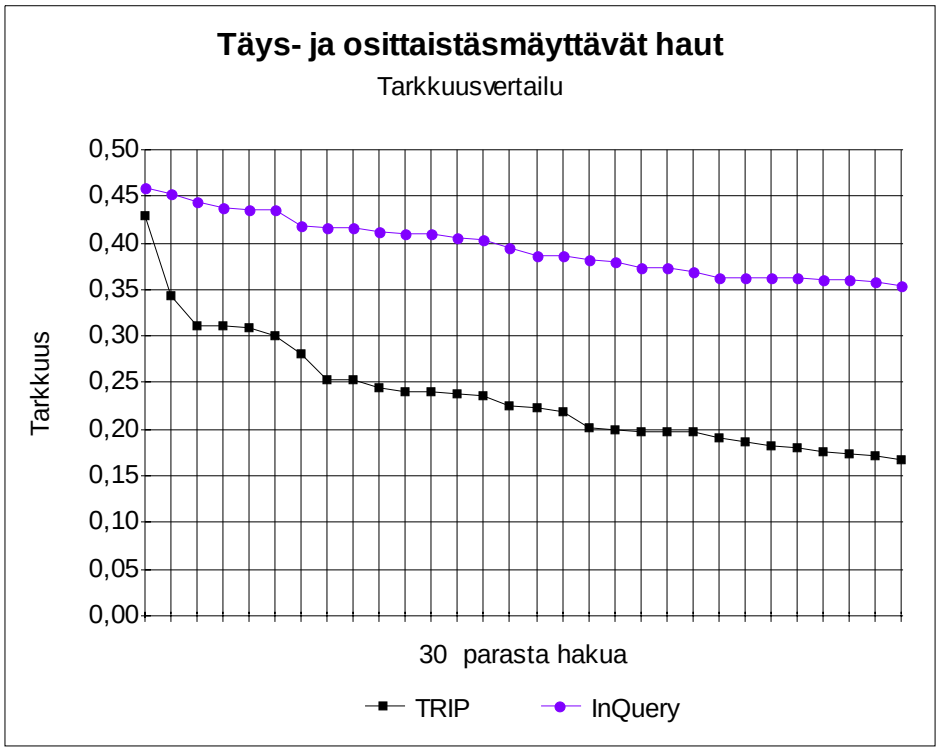
26627 0,361 200 InQuery
#sum(etelä-amerikka velka talous kriisi ongelma)

24659 0,360 200 InQuery
#sum(etelä keski latinalainen väli amerikka argentiina brasilia chile kolumbia meksiko latino lattari maat ulkomaan valtion velka taakka talous luotto kriisi ongelmat vaikeudet velkaantuminen maksu häiriöt kyky kehittyminen kehitys synty ratkaisut toimenpiteet toimet ongelmat dollari kolmas maailma köyhät maat öljy)

26367 0,358 200 InQuery
#sum(etelä-amerikka brasilia argentiina velkakriisi velka maailmanpankki latinalainen meksiko hoito taakka köyhyys bruttokansantuote ympäristönsuojelu yk wto öljykriisi hoito anteeksi)

26390 0,355 200 InQuery
#sum(Amerikka Etelä latinalainen velka kriisi kehitysmaaluotto bruttokansantuote CEPAL velanpienennysohjelma)

Liite 3
Informaatiotutkimuksen perusopintojen syksyn 2003 tiedonhaun harjoituskurssi: 30 parhaan täys- ja osittaistäsmäyttävän haun tarkkuus



Liitetaulukko 1: Täys- ja osittaistäsmäyttävien hakujen tarkkuusarvojen vertailu

Täystäsmäyttävien hakujen keskitarkkuudeksi saatiin 0,236. Osittaistäsmäyttävien hakujen keskitarkkuus oli tutkitussa osa-aineistossa 0,396.