FIRE@uta.fi: Tiedonhaun tutkimusta Tampereen yliopistossa – osa 1

Kalervo Järvelin (toim.)

(http://www.info.uta.fi/research/fire.html, http://www.info.uta.fi/tutkimus/labra.htm)

Kalervo Järvelin (ed.), FIRE@uta.fi : Information retrieval research at the University of Tampere – Part 1. Informaatiotutkimus 19(1): 21-29.

FIRE (the Finnish Information Retrieval Expert Group) is the research group on Information Retrieval active at the Department of Information Studies, University of Tampere, Finland. The first part of this article describes the IR Laboratory of the department and the FIRE research activities in text retrieval. In the second part, multimedia retrieval, IR learning environments and IR in the information seeking context will be presented.

Editor's address: Department of Information Studies, University of Tampere, FIN-33014 TAMPERE, Finland. E-mail: kalervo.jarvelin@uta.fi

1. Johdanto

FIRE (Finnish Information Retrieval Expert Group) on Tampereen yliopiston Informaatiotutkimuksen laitoksella toimiva tiedonhaun tutkimusryhmä. Tutkimusryhmä edistää tiedonhaun tutkimusta Suomessa yhdessä ja jäsentensä kautta toteuttamalla tutkimusprojekteja, tuottamalla tutkimusjulkaisuja, ohjaamalla opinnäytteitä maisteri- ja tohtoritasolla, osallistumalla kansainväliseen tutkimusyhteistyöhön ja konferenssitoimintaan, antamalla opetusta ja pitämällä säännöllistä seminaaria. FIREn toiminta alkoi 1990-luvun alussa, kun laitoksen tutkimustoimintaa ja muutakin kehittämistä linjattiin tiedonhaun, -hankinnan ja tietohallinnon tutkimukseen. Laitokselle perustettiin Tiedonhaun tutkimuslaboratorio ja ensimmäiset jatko-opiskelijat alkoivat ihmetellä laboratorion mahdollisuuksia. Koska FIRE toimii nimenomaan Tampereen yliopistossa, pääotsikko FIRE@uta.fi kertoo lyhyesti, mistä tässä artikkelissa on kyse: artikkelissa esitellään FIREn piirissä tapahtuvaa tutkimusta (sähköpostiosoitteena se ei kuitenkaan toimi). Ylläolevista verkko-osoitteista löytyy myöhemminkin ajankohtaista tietoa FIREn toiminnasta.

Tiedonhaun tutkimuksen ydinryhmä, tämän artikkelin kirjoittajat, koostuu seuraavista, pääasiassa tiedonhaun tutkimuslaboratorion yhteydessä toimivista tutkijoista: Riitta Alkula, Hannele Fabritius, Kai Halttunen, Turid Hedlund, Kalervo Järvelin, Jaana Kekäläinen, Heikki Keskustalo, Marjo Markkula, Ari Pirkola ja Eero Sormunen sekä Pertti Vakkari.

Tämä katsaus jakautuu kahteen osaan: ensimmäisessä esittelemme Tiedonhaun laboratorion ja FIREn tekstitiedonhaun tutkimushankkeita; toisessa osassa esittelemme multimediatiedonhaun tutkimusta, tiedonhaun oppimisympäristöjen tutkimusta sekä tiedonhaun tutkimusta osana tiedonhankintaa. Artikkelin tutkimusesittelyt ovat tutkijoiden itsensä kirjoittamia. Kunkin osuuden kirjoittaja näkyy esittelyn yhteydessä.

2. Tiedonhaun tutkimuslaboratorio

Heikki Keskustalo (ccheke@uta.fi)

Tiedonhaun tutkimuslaboratorio perustettiin vuonna 1991. Perusideana oli Unix-palvelimen ympärille rakentuva ympäristö, joka tukee tiedonhaun kokeellista tutkimusta. Alkuvaiheessa keskityttiin yksinomaan tekstitiedonhaun tutkimukseen. Viime vuosina toiminta on laajentunut kuvatiedonhakuun ja tulevaisuudessa multimediatiedonhaun eri alueille. Toiminta on myös

laajentunut tukemaan opetusta sekä tilaustutkimushankkeiden toteuttamista. Ensimmäinen laboratoriota hyödyntävä sivuaineen tutkielma sekä lisensiaattityö valmistuivat vuonna 1994 ja ensimmäiset väitöskirjat vuonna 1999.

Laboratoriossa on käytettävissä tekstitiedonhaun ja kuvatiedonhaun menetelmien ja järjestelmien testaukseen soveltuvat tutkimusympäristöt. Keskeisimmät näistä ovat suomenkielinen ja englanninkielinen tekstitiedonhaun tutkimustietokanta sekä kuvatiedonhaun tietokanta. Kukin tutkimustietokanta koostuu kolmesta osasta: artikkelitietokannasta, hakukysymyskokoelmasta sekä relevanssiarvioista.

Suomenkielisen tekstitiedonhaun tutkimustietokannan aineisto on saatu Aamulehdeltä, Keskisuomalaiselta ja Kauppalehdeltä. Artikkeliaineisto käsittää yli 50.000 sanomalehtijuttua, joita silmälläpitäen on laadittu 35 valmista tehtävää eli hakukysymystä. Tehtävien perusteella artikkeleita voidaan yrittää hakea tietokannasta. Hakujen onnistumista voidaan arvioida, sillä tietokannasta on etsitty tehtäviin sopivat relevantit dokumentit. Relevanssitietokanta käsittää noin 17.000 intellektuaalista relevanssiarviota.

Englanninkielinen TREC -tutkimustietokanta on saatu kansainvälisenä tutkimusyhteistyönä. Laboratoriossa käytössä oleva tietokannan osio sisältää yli puoli miljoonaa englanninkielistä dokumenttia sekä 350 testikysymystä relevanssitietoineen. Lisäksi laboratoriossa on käytettävissä useita pieniä englanninkielisiä viitekokoelmia.

Kuvatietokantojen tutkimus- ja opetusympäristö kattaa yli 50.000 Aamulehden arkistosta saatua, pääasiassa kuvatoimisto Lehtikuvan, digitaalista uutiskuvaa. Aineiston pohjalta on kehitetty uusiin innovaatioihin perustuva hahmopohjaisten kuvahakualgoritmien testikokoelma. Lisäksi noin 2500 kuvan osajoukko on täydennysindeksoitu Informaatiotutkimuksen laitoksella hakuominaisuuksien monipuolistamiseksi.

Tekstitiedonhaun ohjelmistoina käytetään InQuery- ja TRIP -sovelluksia. Massachusettsin yliopistossa kehitetyssä InQueryssä hakulauseke voidaan täsmäyttää dokumenttiin joko Boolen logiikkaan tai Bayesin päättelyverkkomalliin perustuen (viimeksi mainitussa relevanssilajittelu). TietoEnatorin toimittama TRIP on perinteinen Boolen logiikkaan pohjautuva tekstitietokantojen hallintaohjelmisto, joka soveltuu erityisesti opetuskäyttöön. InQuerylle on asennettu suomen- ja englanninkieliset tekstihaun tutkimuskannat ja TRIPille suomenkielinen tekstitiedonhaun tutkimuskanta.

Kuvatiedonhaun ohjelmistona käytetään ruotsalaisen JOB Systemintegrationenin toimittamaa NEWSLINK -kuva-arkistojärjestelmää, joka on TRIP -sovellus. Kuvahaun tutkimuksessa kokeillaan lisäksi mm. hahmopohjaiseen hakuun perustuvaa CST -ohjelmistoa, jolle on asennettu 25.000 kuvan testitietokanta. Hakuohjelmistoja käytetään myös laitoksen opetusprojekteissa sekä tiedonhakuun että tietokantojen suunnitteluun ja toteutukseen.

Tiedonhaun tärkeitä apuohjelmistoja ovat morfologiset ohjelmat, sanavartalo-ohjelmat ja elektroniset sanakirjat. Käytettävissä on Lingsoft Oy:n suomen, englannin ja saksan morfologiset ohjelmat FINTWOL, ENGTWOL ja GERTWOL. Morfologisia ohjelmia ja sanavartalo-ohjelmia (mm. Porter, kstem) käytetään mm. tekstitietokantojen hakemistojen rakentamisessa InQueryssä. Kielikoneen toimittamia elektronisia sanakirjoja käytetään mm. käännettäessä hakulausekkeita automaattisesti kieleltä toiselle.

Tiedonhaun interaktiivisen tutkimus- ja oppimisympäristön (tiedonhakupeli, IR-Game) ensimmäinen versio valmistui vuonna 1998. Kyseessä on laitoksella kehitetty uudentyyppinen tiedonhaun oppimisympäristö, jossa tiedonhaun tutkija tai opiskelija saa järjestelmältä palautetta hakujen onnistumisesta havainnollisessa muodossa monin eri tavoin. Yhtenä esimerkkinä palautteesta voidaan mainita hakutulosten välitön ja automaattinen esittäminen saanti-tarkkuuskäyräparvina. Tiedonhakupelissä integroituvat laboratorion tutkimusaineistot ja tekstitiedonhaun hakuohjelmistot sekä tiedonhaun apuohjelmistot. Vuoden 2000 alkupuolella ohjelmistosta on laadittu täysin uudistettu versio (ks. artikkelin osa 2, jakso 3.2).

3. Tekstitiedonhaun tutkimus

3.1. Menetelmä Boolen kyselyjen evaluointiin tiedonhakujärjestelmän eri toimintatasoilla

Eero Sormunen (lieeso@uta.fi)

FREETEXT -hankkeessa on kehitetty laboratorioympäristöön soveltuva Boolen kyselyiden evaluointimenetelmä, jolla voidaan arvioida kyselyiden tehokkuutta järjestelmän toiminnan eri tasoilla. Perinteisesti Boolen hakujärjestelmien laboratoriotestaus on perustunut yhden kyselyn muodostamiseen kustakin hakutehtävästä. Kyselyn tehokkuutta on mitattu laskemalla yksittäisen kyselyn tulosjoukosta saanti ja tarkkuus sekä edelleen keskiarvot saannille ja tarkkuudelle yli kaikkien hakutehtävien. Käytännössä

kyselyt edustavat lähes satunnaisesti järjestelmän toiminnan eri tasoja (esim. saantiasteikolla tarkastellen). Näin saanti- ja tarkkuuskeskiarvojen perusteella on vaikea tehdä päteviä johtopäätöksiä eri hakumenetelmien välisistä tehokkuuseroista. Lisäksi ongelmana on ymmärtää, mitkä tehokkuuserot selittyvät hakijan, mitkä järjestelmän toiminnan perusteella.

Kehitetty menetelmä perustuu uuteen tapaan hyödyntää ammattitaitoista hakujen suunnittelijaa. Hakujen suunnitteluprosessissa tuotetaan *kattavat hakusuunnitelmat* (*Inclusive Query Plans*), joissa on pyritty luettelemaan kaikki hakutehtävän kannalta käyttökelpoiset kyselyfasetit ja edelleen kaikki vaihtoehtoiset hakusanat kustakin fasetista. Lohkostrategian muodossa esitetystä kattavasta hakusuunnitelmasta tuotetaan *alkeiskyselyt* (*Elementary Queries*), joista voidaan sopivasti yhdistelemällä rakentaa kullakin toiminnallisella tasolla *optimikyselyt* (*Optimal Queries*) erityistä optimointialgoritmia käyttäen. Optimoinnissa hyödynnetään ns. täydellisiä relevanssitietoja, joten menetelmän käyttö rajoittuu perinteisiin laboratoriokokoelmiin.

Menetelmällä on useita hyviä ominaisuuksia. Siinä pystytään integroimaan hakija evaluointiprosessiin kontrolloidulla tavalla, mikä ei ole onnistunut aiemmissa menetelmissä. Kattavien hakusuunnitelmien avulla pystytään esittämään edustavasti *kyselyjen säätelyavaruus* (*Query Tuning Space*), mikä mahdollistaa optimaalisesti toimivan kyselyn hakemisen yksilöllisesti kullakin vertailtavalla hakumenetelmällä. Optimoinnissa kyselyjen *tyhjentävyys* (*exhaustivity*) ja *kattavuus* (*extent*) voivat hakeutua optimaalisiin arvoihinsa eri hakutilanteissa.

Kyselyiden optimointialgoritmi kehitettiin Harterin (1990) esittämän idean pohjalta ja siinä sovellettiin periaatteita algoritmeista, joita on sovellettu lineaarisen ohjelmoinnin erääseen tyyppiongelmaan: *binääriseen repuntäyttöongelmaan* (*Zero-One Knapsack Problem*). Algoritmin avulla voidaan hakea optimaalinen alkeiskyselyiden kombinaatio halutuissa *vakiotoimintapisteissä* (*Standard Point of Operation*), kuten kiinteillä saantitasoilla $R_{0.1}$...$R_{1.0}$ tai valituilla tulosjoukon koon maksimiarvoilla (esimerkiksi 2, 5, 10,...500 dokumenttia). Näin Boolen kyselyistä saatavat testitulokset ovat yhteismitallisia osittaistäsmäyttävien järjestelmien testitulosten kanssa.

Uuden menetelmän käyttökelpoisuutta testattiin laajassa kokeessa, jossa selvitettiin vapaatekstihaun tehokkuuteen vaikuttavia tekijöitä suurissa tekstitietokannoissa. Tutkimuksessa tutkittiin sekä tietokannan koon että relevantin aineiston tiheyden vaikutusta kyselyiden tehokkuuteen. Suureen saantiin tähtäävissä hauissa keskeinen havainto oli, että kyselyjen tuloksellisuuteen vaikutti vahvasti pienehkö joukko dokumentteja, joissa osa hakusuunnitelmien faseteista oli ilmaistu implisiittisesti. Nämä *vaikeimmin löydettävät dokumentit* (*Least Searchable Documents*) pakottavat vähentämään kyselyiden tyhjentävyyttä yleensä tyhjentävyystasolle yksi (yhden fasetin kysely), jolloin tarkkuus romahtaa. Tulokset antoivat empiirisen vahvistuksen tunnetun STAIRS -tutkimuksen herättämille epäilyille vapaatekstihaun ongelmista suurissa tekstitietokannoissa (ks. Blair & Maron, 1985).

Suureen tarkkuuteen tähtäävissä hauissa tulokset paljastivat mm. kyselyjen tyhjentävyyden keskeisen roolin kaikkein relevanteimpien dokumenttien löytämisessä. Parhaat dokumentit rikastuvat tulosjoukkoon tyhjentäviä kyselyjä käytettäessä. Samalla saatiin näyttöä siitä, että aiemmat tutkimustulokset läheisyysoperaattorien tuomasta tarkkuusedusta AND -rajauksiin verrattuna ovat vahvasti liioiteltuja. Pienissä tulosjoukoissa AND -operaattoreita käyttäen päästään keskimäärin samaan tarkkuuteen kuin läheisyysoperaattoreilla.

Tutkimuksessa selvitettiin useilla eri testeillä menetelmän validiteettia, reliabiliteettia ja tehokkuutta. Tiedonhakupelillä (ks. artikkelin osa 2, jakso 3.2) suoritettavat vertailuoptimoinnit osoittivat että kehitetty optimointialgoritmi toimii luotettavasti.

Yksityiskohtaiset tulokset on julkaistu tekijän väitöskirjassa (Sormunen, 2000).

3.2. Kyselyjen rakenteet ja laajentaminen

Jaana Kekäläinen (lijakr@uta.fi)

Tekstitiedonhaussa hakuaiheen kuvaaminen kyselynä on pulmallista muun muassa seuraavien seikkojen takia: Ensinnäkin hakuaihe ilmaistaan tyypillisesti joukkona hakuavaimia, joiden välisiä suhteita ei kuitenkaan kyselyissä pystytä ilmaisemaan samalla tarkkuudella kuin luonnollisessa kielessä. Syy ja seuraus, subjekti ja objekti eivät erotu toisistaan kyselyissä. Toiseksi kyselyyn tulisi saada ne lukuisat eri ilmaisut, joita dokumenteissa voidaan hakuaiheesta käyttää. Kolmanneksi kyselyn avainten tulisi esiintyä vain toivotuissa, hakijan kannalta mielenkiintoisissa dokumenteissa, ei 'väärissä' teksteissä. Hakuavainten valinta on miltei mahdoton tehtävä, ainakin jos haetut dokumentit ovat ennalta tuntemattomia. Kyselyjen muotoilu, hakuavainten etsiminen ja valinta ovatkin, edellä mainituista syistä, suosittuja aiheita tiedonhauntutkimuksessa. Pyrkimyksenä on kehittää

menetelmiä, jotka pystyvät tarjoamaan hakuaihetta kuvaavia hakuavaimia ja muodostamaan kyselyn, joko automaattisesti tai hakijaa avustaen. Tutkimuksemme tällä saralla liittyy kyselyrakenteisiin ja kyselyjen laajentamiseen uusilla hakuavaimilla.

Rakenteisen kyselyn perusmalli on Boolen logiikan mukainen lohkokysely, joka perustuu aiheen fasettianalyysiin. Kyselyt voidaan rakentaa hakuavainten yhdisteiden ja leikkausten yhdistelminä, so. käyttäen OR -operaattoria yhdistämään kunkin fasetin vaihtoehtoisia hakuavaimia ja AND -operaattoria yhdistämään toisiaan täydentäviä hakuavainryhmiä. Täystäsmäytyksessä (Boolen logiikka) kyselyjen rakenteen ja hakuavainten määrän vaikutus hakutuloksiin tunnetaan melko hyvin. Osittaistäsmäytykseen, joka on muuttumassa vallitsevaksi hakumenetelmäksi, lohkorakenne ei kuitenkaan ole sellaisenaan siirrettävissä, koska täsmäytys perustuu dokumenttien lajitteluarvon laskemiseen niiden sisältämien kyselyn hakuavainten painojen perusteella. Tällöin ei ole itsestään selvää, millainen tulkinta lajitteluarvon laskennassa olisi annettava operaattoreille, jotka luovat kyselyn rakenteen. Suuri osa osittaistäsmäytysmenetelmistä ei tue tämän kaltaista kyselyrakennetta lainkaan, tai sallii korkeintaan hakuavainten painottamisen niiden keskinäisen merkityksen erottamiseksi. Nimitämme tällaisia kyselyjä *rakenteeltaan heikoiksi*. Operaatioiden käyttö on mahdollista osittaistäsmäytyksessäkin, niissä voidaan antaa tulkinta Boolen operaatioille tai käyttää muita operaatioita. Nämä operaatiot ohjaavat sitä, miten hakuavainten painot lasketaan dokumentin lajitteluarvoksi. Niiden avulla on mahdollista muodostaa rakenteisia kyselyjä. Mikäli kyselyssä voidaan tunnistaa fasettirakenne tai kyselyn käsitteet ovat tunnistettavissa, kutsumme kyselyä *rakenteeltaan vahvaksi*.

Kyselyjen laajentamisessa uusia hakuavaimia voidaan poimia eri lähteistä, kuten tietokannasta ensimmäisen hakutuloksen perusteella, tai etsimällä hakuavainten kanssa usein esiintyviä avaimia (tilastolliset menetelmät). Laajentaminen voi perustua myös sanakirjoihin tai tesauruksiin, jotka voivat olla sidoksissa tietokantaan tai riippumattomia siitä. Laajentaminen voidaan tehdä automaattisesti tai hakijan valintaan perustuen. Laajentamiskokeita leimaa pitkälti pyrkimys automaattisuuteen, siten hakijan valinnat tai inhimillistä ajattelua vaativa käsitteiden tulkinta eivät ole maailmalla kovin muodikkaita.

Tähän mennessä olemme testanneet kyselyjen rakenteen ja laajentamisen vaikutusta hakutulosten laatuun täystäsmäytyksessä (Kristensen, 1993; Järvelin & al., 1996) ja osittaistäsmäytyksessä (Kekäläinen & Järvelin, 1998; Kekäläinen, 1999; Kekäläinen & Järvelin, 2000). Viimeksi mainitussa tapauksessa testit tehtiin tiedonhaun tutkimuslaboratoriossa ja hakujärjestelmänä oli todennäköisyyslaskentaan perustuva InQuery, joka sallii niin heikkojen kuin vahvojen kyselyrakenteiden käytön. Testasimme erilaisten operaattoreiden vaikutusta laajentamattomissa ja laajennetuissa kyselyissä sekä heikko- ja vahvarakenteisten kyselyjen tuloksellisuuden eroja. Yleistäen voidaan sanoa, että rakenteella on väliä laajentamistavasta riippuen: laajennetut, vahvarakenteiset kyselyt vaikuttavat tehokkaimmilta. Tähän asti olemme käyttäneet tesaurusta kyselyjen laajentamiseen ja testitietokantamme on ollut suomenkielinen. Näyttää siltä, että niin kielen kuin laajentamistavan vaihtaminen tuo esiin uusia ongelmia: sekä sanaliittojen tunnistaminen ja painottaminen että käsite/fasettirakenteen säilyttäminen tilastollisia laajentamismenetelmiä käytettäessä vaativat selvittämistä.

3.3. Kieltenvälinen tiedonhaku

Ari Pirkola (pirkola@tukki.jyu.fi)

Kieltenvälinen tiedonhaku (Cross-Language Information Retrieval, CLIR) on tiedonhakua, jossa kysely esitetään eri kielellä kuin millä tietokannan dokumentit on kirjoitettu. Siten suomi-englanti CLIR -järjestelmässä palautteena suomenkieliseen kyselyyn järjestelmä antaa englanninkielisiä dokumentteja. Jotta kielirajat ylittävä tiedonhaku olisi mahdollista, kyselyt (dokumentit) on käännettävä dokumenttien (kyselyn) kielelle sanakirjojen, korpusten tai koneellisten käännösjärjestelmien avulla.

Pirkola tutki väitöskirjatyössään (1999) tekstihaun lingvistisiä ongelmia. Tärkeän osan työstä muodosti kieltenvälinen tiedonhaku. Hän tarkasteli työssä kieltenvälisen tekstihaun keskeisiä käsitteitä, menetelmiä ja ongelmia. Kääntämiseen liittyviä merkittäviä ongelmia ovat *käännöspolysemia* sekä *sanakirjojen rajallinen kattavuus*. Käännöspolysemiassa on kyse sanojen merkitysten määrän lisääntymisestä korvattaessa lähdekielen sana sanakirjaan sisältyvillä vastinsanoillaan. Koska yhdellä sanalla on usein monia vastinsanoja toisessa kielessä, joista useat ovat monimerkityksisiä, on tavallista että kohdekieliset kyselyt sisältävät runsaasti kyselyn aiheen kannalta huonoja hakuavaimia. Kattavuusongelma koskee varsinkin erikoisalojen termejä, joita yleissanakirjoissa on niukasti.

Pirkola tarkasteli väitöskirjassaan erityisesti

kyselyrakenteiden sekä erilaisten käännösmenetelmien vaikutusta kyselyjen tehokkuuteen kieltenvälisessä tekstihaussa, jossa lähdekielenä oli suomi ja kohdekielenä englanti. Tutkimus osoitti, että sanakirjakäännökseen perustuva kieltenvälinen tekstihaku voi olla lähes yhtä tehokasta kuin yksikielinen tekstihaku, jos kääntämisessä käytetään erikois- ja yleissanakirjaa sekä jos kyselyt ovat rakenteisia. Rakenteisilla kyselyillä tarkoitetaan tässä hakuavainten ryhmittämistä sanakirjojen antamien tulostietueiden tai käsiteanalyysin perusteella ja avainten liittämistä toisiinsa sopivilla hakuoperaattoreilla. Näillä menetelmillä käännöspolysemiaongelma sekä sanakirjojen kattavuusongelma voidaan ratkaista hyvällä menestyksellä.

Kieltenvälinen tiedonhaku on osa Suomen Akatemian rahoittamaa Informaatiotutkimuksen laitoksen tutkimusprojektia *Kyselyrakenteet ja sanakirjat käsiteperusteisen ja kieltenvälisen tiedonhaun välineinä*. Tutkimme (Ari Pirkola, Heikki Keskustalo, Kalervo Järvelin) mm. n-gram -tekniikkaan perustuvaa erisnimien täsmäyttämistä. Monissa hakuaiheissa erisnimet ovat keskeisiä, mutta niiden kirjoitusasu vaihtelee kielten välillä. On kuitenkin tavallista, että erisnimiä ei pystytä kääntämään, sillä niiden kattava esittäminen käännössanakirjoissa ei ole mahdollista. Erisnimet voidaan kuitenkin kääntämisen sijasta pilkkoa pienemmiksi osiksi (n-grammeiksi), jolloin kokonaisten sanojen sijasta täsmäytetään sanojen osia. N-gram -menetelmään perustuva erisnimien täsmäyttäminen voi olla hyvin tehokasta, sillä eri kieliset samaa merkitsevät erisnimet ovat usein kirjoitusasuiltaan samankaltaisia sisältäen monia yhteisiä n-grammeja.

FIREn piirissä kieltenvälistä tiedonhakua ovat tutkineet myös Deniz Puolamäki sekä Turid Hedlund. Deniz Puolamäki teki aiheesta pro gradu -tutkielman (Puolamäki, 1999). Tutkimuskohteena oli englanti - suomi -tiedonhaku.

Turid Hedlund tutkii väitöskirjatyössään ruotsin kielen ominaisuuksia tekstihaun kannalta sekä kieltenvälistä tiedonhakua, jossa ruotsi on lähde- tai kohdekielenä. Ruotsinkielisen tekstihaun tutkimus on tarpeellista, koska ruotsin morfologisten ja semanttisten ominaispiirteiden vaikutukset tekstihakuun tunnetaan huonosti. Ruotsille tunnusomaista on mm. homonyymisten sanojen ja yhdyssanojen yleisyys sekä yhdyssanojen muodostaminen yhdysmorfeemien avulla. Turid Hedlundin tutkimuksen avulla on mahdollista kehittää ruotsin kielen kieliteknologisia ohjelmia tai niiden sovelluksia entistä paremmin tekstihakuun ja kieltenväliseen tiedonhakuun soveltuviksi.

3.4. Merkkijonoista suomen kielen sanoiksi: Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja vaikutukset tiedonhakuun

Riitta Alkula (riitta.alkula@tieto.com)

Suomenkielisten tekstitietokantojen tallennus- ja hakutekniikat (FULLTEXT) -projektin tarkoituksena oli tuottaa perusselvitys siitä, miten suomen kielen morfologisten tulkintaohjelmien avulla voidaan ratkaista sellaisia tiedon tallennuksen ja haun ongelmia, jotka johtuvat suomen kielen erityispiirteistä.

Tutkimusta varten rakennettiin kaupallisen BASIS -hakujärjestelmän avulla testausympäristöjä. Samasta tekstiaineistosta tuotettiin joukko erilaisia tietokantoja soveltamalla suomen kielen morfologisia tulkintaohjelmia eri tavoin aineiston tallennuksessa ja tiedonhaussa. Näitä tietokantoja sekä niistä tehtyjen tiedonhakujen tuloksia vertailtiin toisiinsa.

Projektin tutkimusympäristöt olivat seuraavat:

T1) **Perinteinen** hakeminen: Hakijan katkaisemat hakusanat haettiin taivutusmuodot sisältävästä hakemistosta

T2) **Automaattinen katkaisu**: hakusanat katkaistiin automaattisesti sananvartaloita tuottavalla ohjelmalla; kysely tehtiin taivutusmuotohakemistosta

T3) **Seulonta**: Edellisellä tavalla tuotetut hakusanat tarkistettiin perusmuoto-ohjelmalla, kysely tehtiin taivutusmuotohakemistosta

T4) **Perusmuotojen** ja yhdyssanojen alkuosien hakeminen perusmuotohakemistosta perusmuotoisilla hakusanoilla

T5) **Perusmuotojen ja** yhdyssanan kaikkien **osien** hakeminen perusmuotohakemistosta perusmuotoisilla hakusanoilla

Kun hakemistoon tallennettavat sanat palautetaan perusmuotoon, hakemistoon tulee vähemmän sanoja kuin perinteiseen hakemistoon, johon sanat tallennetaan taivutusmuodossaan. Perusmuotohakemisto

vie vähemmän muistitilaa kuin taivutusmuotohakemisto - myös siinä tapauksessa, kun perusmuotohakemisto sisältää perusmuotojen lisäksi myös yhdyssanojen osat.

Tiedonhaun onnistumista mitataan yleisesti saannilla ja tarkkuudella. Näiden arvot ovat yleensä käänteisiä: kun haun saanti paranee, tarkkuus huononee ja päinvastoin. FULLTEXT-projektissa kuitenkin todettiin, että tietyissä tapauksissa sekä hakujen tarkkuus että saanti paranevat verrattuna hakijan katkaisemilla hakusanoilla toteutettuun hakuun. Näin käy silloin, kun morfologisilla ohjelmilla käsitellyt hakusanat tuottavat tulosjoukkoon enimmäkseen relevantteja dokumentteja.

Tutkimuksessa seulonta (T3) osoittautui saanniltaan selvästi huonoimmaksi, koska seulontatapa tutkimuksessa oli säädetty tiukaksi ja näin ollen rajasi relevanttejakin dokumentteja pois tulosjoukosta. Mutta vaikka seulonnan tarkkuusarvot olivat vertailujoukon parhaat, ero toisiin oli suhteellisen pieni. Perusmuotohakemistosta (T4) ja ositetusta perusmuotohakemistosta (T5) saatujen tulosjoukkojen tarkkuus oli lähellä seulonnan tarkkuusarvoja.

Korkeimmat saantiarvot saatiin ositetusta perusmuotohakemistosta (T5). Tässä tapauksessa haku kattoi myös ne yhdyssanat, joissa hakusana ei ollut yhdyssanan alussa; muista hakemistoista haettaessa tällaiset yhdyssanat jäivät löytymättä.

Jos perusmuoto- tai ositetusta perusmuotohakemistoista hakee vain hakusanan täsmällisiä perusmuotoja, tulosjoukkojen saanti jää kehnoksi. Käytännössä kyselyt kannattaa näissä ympäristöissä laajentaa suoraan hakusanan johdoksiin tai yhdyssanoihin, jotka sisältävät hakusanan. Hakutulok-sen tarkkuus ei huonone samassa suhteessa kuin saanti kasvaa. Sen sijaan taivutusmuotohakemistossa vastaava laajentaminen ei toimi yhtä täsmällisesti, vaan katkaistut hakusanat tuottavat tulosjoukkoihin suhteessa enemmän epärelevantteja dokumentteja.

3.5. Ohjelmistot kyselyjen rakentamiseen ja laajentamiseen

Kalervo Järvelin (likaja@uta.fi)

Edellä todettiin, että tiedonhakijan alkuperäiset kyselymuotoilut eivät läheskään aina ole sisällöllisesti riittäviä, joten onnistuneen tuloksen saaminen edellyttää kyselyn laajentamista uusilla hakuavaimilla. Toinen tiedonhakijan ongelma monen eri tietokannan ja hakuohjelman ympäristössä on hakuohjelman kyselykieli. Hakijan pitää muotoilla kyselynsä käyttäen hakuohjelman syntaksia: operaattoreita, merkkijonon korvaussymboleja, sulkumerkkejä, jne. Sisällöllisesti sama kysely eri hakuympäristöissä voi olla esitystavaltaan varsin erilainen. Monen kyselykielen hallinta on hakijalle lähinnä rasite. Kolmas kyselyjen muotoiluongelma liittyy tutkimustarpeisiin: rakenteell-isesti erilaisten kyselyjen suorituskyky vaihtelee eri hakuympäristöissä (ks. jakso 3.2). Näin ollen on tarpeen luoda väline, jonka avulla kyselyjen laajentaminen, rakenteistaminen ja varustaminen asianmukaisella syntaksilla voidaan automatisoida. ExpansionTool -projekti pyrkii näihin päämääriin. (Järvelin & *al.*, 1996).

ExpansionTool -projektissa on kehitetty samanniminen ohjelmisto, jonka avulla kyselyjen laajentaminen, rakenteistaminen ja varustaminen asianmukaisella syntaksilla voidaan automatisoida. Lähtökohtana on ajatus, että tiedonhakija tunnistaa tiedontarvettaan kuvaavat käsitteet ohjelmiston käsitemallista ja kertoo parametrien avulla minkä tyyppisen kyselyn haluaa rakennettavaksi. Tämän jälkeen ohjelmisto tuottaa kyselyn automaattisesti.

ExpansionTool-ohjelmisto on jäsennetty tiedonhaun tasoperiaatteen (Järvelin, 1995) mukaisesti kolmeen tasoon: käsitetaso, ilmaisutaso ja merkkijonotaso. Tämä jäsennys näkyy sekä tietokannan rakenteessa että kyselyn rakentamisen jäsennyksessä. *Käsitetasolla* esitetään valitun aihealueen käsitemalli, joka koostuu nimetyistä käsitteistä ja niiden välisistä suhteista. Käsitteiden välisten suhteiden tyypit voidaan valita tarpeen mukaan, esimerkiksi voidaan esittää hierarkkiset ja assosiaatiosuhteet tai nämä voidaan esitellä hienojakoisemmin esim. laji-alalajisuhteisiin, osa-kokonaisuussuhteisiin, jne. Kullekin suhteelle annetaan sen luotettavuutta kuvaava tunnusluku. Käsitteet muodostavat syklisen käsiteverkon, josta voidaan etsiä valituille käsitteille lähikäsitteitä valitun etäisyyden tai yhteyden luotettavuuden mukaan.

Ilmaisutasolla esitetään käsitemallin käsitteisiin liittyvät ilmaisut ja ilmaisujen väliset suhteet (synonyymisuhde). Kullakin käsitteellä voi olla useita, luotettavuudeltaan erilaisia ilmaisuja, joiden esiintyminen tekstissä antaa vihjeen ao. käsitteen esiintymisestä tekstissä. Yksi ilmaisu voi liittyä moneen eri käsitteeseen (esim. luonto). *Merkkijonotasolla* esitetään kunkin ilmaisun *täsmäytysmallit*, jotka kertovat kyselykielestä riippumattomalla tavalla, miten tarkasteltava ilmaisu voidaan tunnistaa tekstistä tai tietokannan hakemistosta. Edellä kuvattiin, että tietokannan hakemisto voidaan luoda taivutus-

muotoisista sanoista, perusmuotoisista sanoista kokonaisina tai yhdyssanat osittaen. Esimerkiksi taivutusmuotomalleissa annetaan sanavartalo, sanaliittojen tapauksessa annetaan sen osien sallittu etäisyys. Myös täsmäytysmalleihin liittyy luotettavuus kuvatun ilmaisun suhteen. Siten mm. valtion "Peru" sanavartalo "peru-" kuvataan epäluotettavaksi täsmäytysmalliksi. Toisin kuin englannissa, taivutusmuodot ja yhdyssanat ovat olennaisia piirteitä monissa muissa eurooppalaisissa kielissä.

Kyselyn laajentaminen tapahtuu kaikilla kolmella tasolla. Kullekin hakijan valitsemalle käsitteelle (esim. "radioaktiivinen jäte" ja "varastointi") etsitään käsitemallista lähikäsitteitä valittujen suhdetyyppien joukossa ja valitulla luotettavuustasolla. Esimerkiksi löydetään käsitteet: {radioaktiivinen jäte, korkea-aktiivinen jäte, matala-aktiivinen jäte, käytetty polttoaine, varastointi, varasto}. Tämän jälkeen kullekin muodostetulle käsitejoukolle valitaan riittävän luotettavat ja oikeantyyppiset ilmaisut (esim. termit, synonyymit) ja lopuksi ilmaisuille kaikki riittävän luotettavat ja oikeantyyppiset täsmäytysmallit (esim. taivutusmuotomallit). Esimerkiksi käsite "radioaktiivinen jäte" saa ilmaisut {radioaktiivinen jäte, ydinjäte} ja täsmäytysmallit {#1(radioaktiivinen jäte), ydinjäte}, jotka ilmaisevat peräkkäisten sanojen sanaliittoa ja yhdyssanaa perusmuotoisina.

Kyselyn rakenteistamista varten ExpansionTool -ohjelmisto tuntee joukon heikkoja ja vahvoja kyselyrakenteita, joista hakija ilmoittaa haluamansa. Heikon rakenteen tapauksessa hakutehtävän eri fasettien kaikkien käsitteiden kaikki täsmäytysmallit voidaan yhdistää yhdeksi joukoksi (tässä perusmuotomalleja ilman sanaliittojen merkintää):

{radioaktiivinen jäte ydinjäte korkea-aktiivinen jäte matala-aktiivinen jäte käyttää polttoaine varastointi varastoida varasto säiliö}.

Vahvan rakenteen tapauksessa hakutehtävän eri fasettien käsitteiden kaikki täsmäytysmallit voidaan jäsentää omiksi joukoikseen (sanaliitot ja yhdyssanat merkittyinä kuten edellä):

{{#1(radioaktiivinen jäte), ydinjäte, #1(korkea-aktiivinen jäte), #1(matala-aktiivinen jäte), #1(käyttää polttoaine)}, {varastointi varastoida varasto säiliö}}.

Kyselyn kääntämistä varten halutun hakuohjelman kyselykielelle ExpansionTool -ohjelmisto tuntee joukon kyselykieliä, joista hakija ilmoittaa haluamansa, esim. InQuery tai Trip. Kääntäminen tapahtuu logiikka-kielioppien avulla. Ylläolevista esimerkeistä saadaan mm. heikkorakenteinen InQuery -kysely (ilman painotusta ja sanaliittoja):

#sum(radioaktiivinen jäte ydinjäte korkea-aktiivinen jäte matala-aktiivinen jäte käyttää polttoaine varastointi varastoida varasto säiliö),

tai vahvarakenteinen Trip -kysely (fasettirakenne ja sanaliitot):

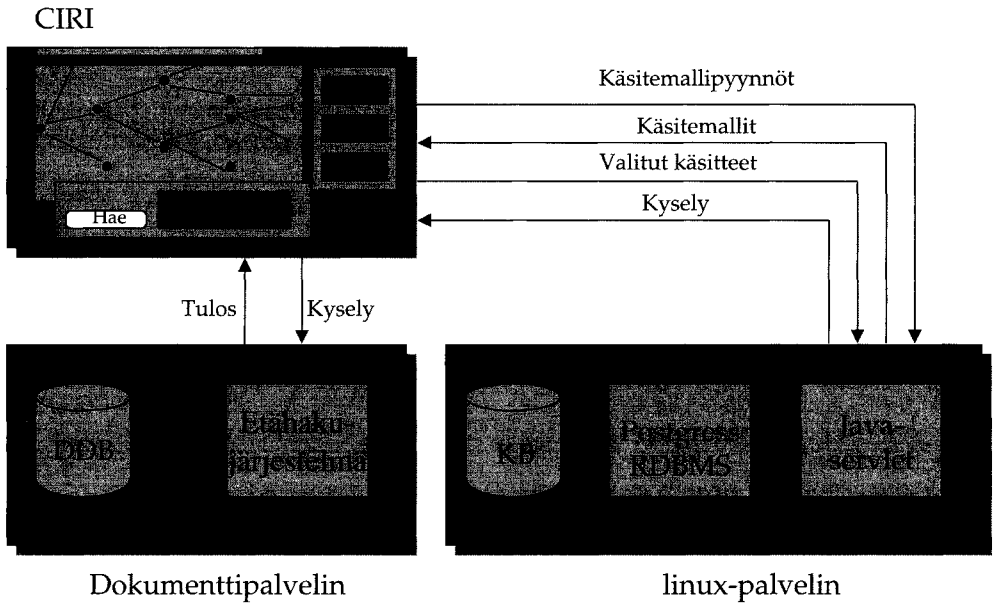
(radioaktiivinen . . jäte or ydinjäte or korkea-aktiivinen . . jäte or matala-aktiivinen . . jäte or käyttää . . polttoaine) and (varastointi or varastoida or varasto or säiliö).

Kootusti voidaan todeta, että ExpansionTool -ohjelmisto on tarkoitettu kyselyjen rakentamiseen seuraavalla tavalla:

- Kyselyn muotoilu perustuu tiedontarvetta kuvaaviin käsitteisiin, jotka valitaan käsitemallista.
- Kysely voidaan muodostaa automaattisesti halutulla rakenteella ja laajennoksella halutulle kyselykielelle asettamalla näitä ilmaisevat parametrit.
- Kyselyn laajennos etenee halutuissa käsitteellisessä ja ilmausten välisissä suhteissa sekä halutulla luotettavuustasolla asetettujen parametrien mukaisesti.
- Hakuavaimet (tai hakuavainkaaviot) muotoillaan automaattisesti tietokannan indeksoinnin mukaisella tavalla, siis ottaen huomioon se, käytetäänkö taivutusmuoto-, perusmuoto-, vai ositettua perusmuotohakemistoa.
- Kyselykielten syntaktiset erot operaattoreiden nimennässä tai sallituissa ilmaisurakenteissa otetaan automaattisesti huomioon eivätkä ne näy käyttäjälle.
- ExpansionTool on tarkoitettu kyselyn rakentamiseen ja laajentamiseen ennen hakijan aloituskyselyä (initial search). Jotkun toiset tekniikat, kuten relevanssipalaute edellyttävät aloituskyselyn, jota sitten parannetaan.

ExpansionTool -ohjelmiston pääkehittäjät ovat olleet Kalervo Järvelin ja Timo Niemi. Jaana Kekäläinen on osallistunut ohjelmiston testaa-miseen ja hän on käyttänyt ohjelmistoa omassa tutkimuksessaan. Monet muutkin FIREn jäsenet ovat osallistuneet varsinkin alku-vaiheessa ExpansionTool -ohjelmiston kehittämiseen.

ExpansionTool -ohjelmiston periaatteita on sovellettu CIRI -ohjelmistossa (Concept-based Information Retrieval Interface), joka www-selaimessa ja -palvelimella toimiva Java-sovellus, jonka tehtävänä on tarjota käsiteverkko tiedonhakijan selattavaksi (Kuva 1). Hakijan tehtyä käsitevalintansa palvelin muodostaa sitä vastaavan kyselyn, jonka käyttöliittymä lähettää tiedonhakujärjestelmälle suoritettavaksi. Hakutulos palautetaan selaimelle. Harri Kemppainen on toteuttanut CIRI-ohjelmiston.



Kuva 1. CIRI -ohjelmiston rakenne

Viiteluettelo

Blair, D.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.

Harter, S.P. (1990). Search Term Comb-inations and Retrieval Overlap: A Proposed Methodology and Case Study. *Journal of the American Society for Information Science 41*(2), 132-146.

Järvelin, K. (1995). *Tekstitiedonhaku tietokannoista.* Espoo, Finland: Suomen ATK-kustannus. (Asiantuntija-sarja: Tiedon haku.)

Järvelin, K. & Kristensen, J. & Niemi, T. & Sormunen, E. & Keskustalo, H. (1996). A Deductive Data Model for Query Expansion. Teoksessa: H.P. Frei, D. Harman, P. Schäuble & R. Wilkinson (toim.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM, 235-249.

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

Kekäläinen, J. & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. Teoksessa: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (toim.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM, 130–137.

Kekäläinen, J. & Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval 1*(4), 329-344.

Kristensen, J. (1993). Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing & Management 29*(6), 733–744.

Pirkola, A. (1998). The Effects of Query Structure and Dictionary Setups in Dictionary-based Cross-language Information Retrieval. Teoksessa: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (toim.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM, 55-63.

Pirkola, A. (1998). Kyselyrakenteiden ja erikoissanakirjan vaikutus sanakirjakäännökseen perustuvassa kieltenvälisessä tiedonhaussa. *Informaatiotutkimus 17*(3), 48-58.

Pirkola, A. (1999). *Studies on Linguistic Problems and Methods in Text Retrieval: The Effects of Anaphor and Ellipsis Resolution in Proximity Searching, and Translation and Query Structuring Methods in Cross-Language Retrieval.* Väitöskirja. Tampere: Tampereen yliopisto. Acta Universitatis Tamperensis 672.

Pirkola, A. & Keskustalo, H. (1999). *The Effects of Translation Method, Conjunction, and Facet Structure on Concept-based Cross-language Queries.* Finnish Information Studies, 13.

Pirkola, A., Keskustalo, H. & Järvelin, K. (1999). The Effects of Conjunction, Facet Structure, and Dictionary Combinations in Concept-based Cross-language Retrieval. *Information Retrieval 1*(3), 217-250.

Puolamäki, D. (1999). *Kielten välinen tiedon-haku: Käännöskyselyjen evaluointi englanti-suomi.* Tampe-reen yliopisto, Informaatiotutkimuksen laitos. Pro gradu - tutkielma.

Sormunen, E. (2000). *A Method for measuring Wide Range Performance of Boolean Queries in Full-Text Databases.* Väitöskirja. Tampere: Tampereen yliopisto. Acta Electronica Universitatis Tamperensis.
<URL: http://granum.uta.fi/pdf/951-44-4732-8.pdf>.