Kalervo Järvelin
Jaana Kekäläinen

Kuinka evaluoida tiedonhakumenetelmiä parhaiden dokumenttien löytämisen kannalta?*

Kalervo Järvelin & Jaana Kekäläinen. Kuinka evaluoida tiedonhakumenetelmiä parhaiden dokumenttien löytämisen kannalta? [IR evaluation methods for retrieving highly relevant documents]. Informaatiotutkimus 19 (3): 63-73, 2000.

This paper proposes evaluation methods based on the use of non-dichotomous relevance judgements in IR experiments. It is argued that evaluation methods should credit IR methods for their ability to retrieve highly relevant documents. This is desirable from the user point-of- view in modern large IR environments. The proposed methods are (1) a novel application of P-R -curves and average precision computations based on separate recall bases for documents of different degrees of relevance, and (2) two novel measures computing the cumulative gain the user obtains by examining the retrieval result up to a given ranked position. We then demonstrate the use of these evaluation methods in a case study on the effectiveness of query types, based on combinations of query structures and expansion, in retrieving documents of various degrees of relevance. The test was run with a best match retrieval system in a text database consisting of newspaper articles. The results indicate that the tested strong query structures are most effective in retrieving highly relevant documents. The differences between the query types are practically essential and statistically significant. More generally, the novel evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable in IR experiments, may reveal interesting phenomena, and allow harder testing of IR methods.

Address: University of Tampere, Department of Information Studies, FIN-33014 University of Tampere, Finland. Email: [likaja, lijakr]@uta.fi

1. Johdanto

Relevanssin käsite ja relevanssin arviointi ovat keskeisiä ongelmia tiedonhaun laboratoriomalliin perustuvassa kokeellisessa tutkimuksessa. Useimmissa kokeissa dokumenttien relevanssi on binäärinen: dokumentit ovat joko relevantteja tai epärelevantteja suhteessa hakuaiheeseen. Binäärinen arviointi ei kuitenkaan pysty osoittamaan dokumenttien relevanssin astetta – jotkut relevanteista dokumenteista ovat erittäin relevantteja hakuaiheen kannalta, jotkut ovat osittain relevantteja, mutta erotettavissa edelleen täysin epärelevanteista dokumenteista. Moniportaista relevanssiasteikkoa on käytetty joissakin tutkimuksissa, mutta vain muutamissa relevanssin astetta on käytetty tulosten tulkinnassa hyväksi (e.g., Hersh & Hickam, 1995). Tavallisesti moniportainen relevanssi on jaettu kahteen luokkaan saannin ja tarkkuuden laskemiseksi (e.g., Blair & Maron, 1985; Saracevic, Kantor, Chamis & Trivison, 1988; Smithson, 1994).

Nykyisissä tiedonhakuympäristöissä tulosjoukon aiheenmukaisesti relevanttien dokumenttien määrä on usein huomattavasti suurempi kuin mitä käyttäjä on valmis selaamaan. Relevanteimpien dokumenttien

*Artikkeli perustuu tutkijoiden Ateenassa heinäkuussa 2000 pidetyn ACM Sigir-konferenssin esitelmään. Esitelmä arvioitiin konferenssin parhaimmaksi.

lajittelu tulosjoukon kärkeen olisi käyttäjän näkökulmasta sangen toivottavaa. Nykyinen, binääriseen relevanssiarviointiin perustuva evaluointi ei tee eroa osittain ja erittäin relevantteja dokumentteja tulosjoukkoon tuottavien hakumenetelmien välillä, mutta tiedonhakumenetelmien kehittämisessä ja evaluoinnissa relevanssin aste tulisi ottaa huomioon. Tässä artikkelissa tarkastelemme moniportaisen relevanssiasteikon käyttöä hakumenetelmien evaluoinnissa ja osoitamme esimerkkitapauksen avulla, että relevanssin asteen huomioon ottaminen paljastaa merkittäviä eroja hakumenetelmien välillä.

Relevanssin asteen vaikutusta voidaan arvioida käyttäen perinteisiä tiedonhaun evaluointimenetelmiä, kuten saanti-tarkkuus-käyriä. Tässä työssä olemme piirtäneet käyrät erikseen jokaiselle relevanssitasolle. Esittelemme myös käyttäjän näkökulmaa painottavan uuden evaluointimitan. Tämä menetelmä estimoi käyttäjän kumuloituvaa hyötyä, joka hänen voidaan olettaa saavan tietyn kokoista tulosjoukkoa selatessaan. Molempien evaluointimenetelmien avulla voidaan vertailla hakujärjestelmiä tai –menetelmiä niiden tulosjoukkojen dokumenttien relevanssin asteen perusteella. Uusi kumuloituvan hyödyn mitta muistuttaa eräiltä ominaisuuksiltaan haun keskipituutta (average search length, ASL, Losee, 1998), tulosjoukon relevanssikertymän puoliintumissijaa ja suhteellista relevanssia (ranked half-life, RHL; relative relevance, RR, Borlund ja Ingwersen, 1998). Kumuloituvan hyödyn etuna on sen kyky ottaa huomioon niin relevanssin aste kuin tulosjoukon järjestys (joka on relevanssin todennäköisyyden määräämä)[1].

Tapausesimerkissämme, joka testaa kyselyjen laajentamista todennäköisyyslaskentaan perustuvassa hakujärjestelmässä, käytämme moniportaista relevanssiasteikkoa ja sovellamme niin perinteisiä kuin uusia evaluointimenetelmiä. Olemme aikaisemmin osoittaneet, että kyselyjen rakenne vaikuttaa tiedonhaun tuloksellisuuteen kyselyjä laajennettaessa, toisin sanoen, kun hakuavainten määrä kyselyissä on suuri (Kekäläinen, 1999; Kekäläinen & Järvelin, 2000). Kyselyjen rakenne tarkoittaa hakulausekkeen syntaktista rakennetta, joka osoitetaan operaattoreilla ja sulkumerkeillä. Kyselyrakenteet jaamme heikkoihin ja vahvoihin rakenteisiin. Heikkorakenteisissa kyselyissä hakuavaimet muodostavat sanajoukon, jossa – periaatteessa – kaikki avaimet ovat samanveroisia. Vahvarakenteisissa kyselyissä hakuavaimet on ryhmitelty edustamansa käsitteen tai fasetin mukaisesti. Aikaisempien tulostemme perusteella oletamme, että hakujen tuloksellisuutta voidaan merkitsevästi parantaa yhdistämällä vahvarakenteiset kyselyt käsiteperusteiseen hakujen laajentamiseen. Koska käytimme binäärisiä relevanssiarvioita, emme tiedä miten erirakenteiset, laajentamattomat tai laajennetut kyselyt tuovat dokumentteja eri relevanssitasoilta tulosjoukkoon. Testaamme tätä esimerkissämme.

Luvussa 2 esittelemme evaluointimetodologiamme: saanti-tarkkuus–käyrien käytön eri relevanssitasoilla ja kumulatiiviseen hyötyyn perustuvat mitat. Luvussa 3 raportoidaan tapaustutkimuksemme, sen testiympäristö, relevanssiarviointi, kyselyrakenteet ja kyselyjen laajentaminen. Johtopäätökset ovat luvussa 4.

2. Moniportainen relevanssiasteikko tiedonhaun evaluoinnissa

2.1 Tarkkuus saannin funktiona

Tiedonhaun tuloksellisuutta kuvataan useimmiten tarkkuuden keskiarvona vakioitujen saantitasojen yli ja saanti-tarkkuus-käyrinä. Tällöin tyypillisesti käytetään binäärisiä relevanssiarvioita. Jotta tiedonhakumenetelmiä voitaisiin evaluoida tulosjoukkojen sisältämien eri relevanssitasoilla olevien dokumenttien perusteella, täytyy tuloksellisuutta tarkastella erikseen kullakin relevanssitasolla. Jos esimerkiksi käytetään neliportaista relevanssiasteikkoa (olkoot relevanssitasot 0 - 3), tarvitaan erilliset saantikannat erittäin relevanteille dokumenteille (relevanssitaso 3), melko relevanteille dokumenteille (relevanssitaso 2) ja osittain relevanteille dokumenteille (relevanssitaso 1). Muita tietokannan dokumentteja pidetään epärelevantteina (relevanssitaso 0). Tutkimuksessamme saantikannat muodostettiin tällä tavalla saanti-tarkkuus-käyrien laskemiseksi.

2.2 Kumuloituvaan hyötyyn perustuvat mitat

Lajiteltua tulosjoukkoa tarkasteltaessa on ilmeistä, että

1. erittäin relevantit dokumentit ovat hyödyllisempiä kuin osittain relevantit,
2. mitä alempana lajitellussa tulosjoukossa relevantti – mitä tahansa relevanssitasoa edustava – dokumentti on tulosjoukossa, sitä vähemmän hyötyä siitä käyttäjälle on, koska käyttäjän halukkuus selata tuloslistaa vähenee dokumentin sijaluvun kasvaessa.

Ensimmäinen kohta johtaa tiedonhaun menetelmien vertailuun dokumentin sijalukuun perustuvan kumuloituvan hyödyn perusteella. Jokaisen dokumentin relevanssin astetta käytetään osoittamaan saavutetun hyödyn määrää dokumentin sijalla lajitellussa

tulosjoukossa. Hyöty lasketaan kumuloituvasti sijalta 1 sijalle *n*. Siten järjestetyt tulosjoukot muutetaan saavutetun hyödyn kuvaajiksi korvaamalla dokumenttien tunnukset niiden relevanssin arvoilla. Oletetaan, että relevanssiarvot 0 – 3 ovat käytössä (3 kuvaa erittäin relevanttia dokumenttia, 0 epärelevanttia). Muuttamalla 200 dokumentin tulosjoukko vastaavaksi relevanssiarvolistaksi saadaan 200-komponenttinen vektori, jonka jokaisen komponentin arvo joko 0, 1, 2 tai 3. Esimerkki:

$$G' = < 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \ldots >$$

Kumuloitunut hyöty *i*:nnellä järjestyssijalla saadaan laskemalla yhteen arvot sijalta 1 sijalle *i*, kun *i* saa arvot 1:stä 200:teen. Sijan *i* tuottamaa hyötyä hyötyvektorissa G merkitsemme G[*i*]. Määrittelemme kumuloituvan hyödyn vektorin rekursiivisesti vektoriksi CG, siten että

$$CG[i] = \begin{cases} G[1], \text{ jos } i=1 \\ CG[i-1]+G[i], \text{ muulloin} \end{cases} \quad (1)$$

kun $i = 1 \ldots n$. Vektorista G' saamme vektorin CG' = <3,5,8,8,8,9,11,13,16,16,...>. Kumuloitunut hyöty millä tahansa järjestyssijalla voidaan lukea suoraan vektorista, esimerkiksi sijalla 7 se on 11.

Kohta kaksi johtaa IR menetelmien vertailuun testikyselyjen tuoman kumuloituneen hyödyn perusteella siten, että dokumentin sijalukua käytetään hyödyn arvoa alentavana tekijänä: mitä suurempi on dokumentin sijaluku, sitä pienempi on dokumentin kumuloituun hyötyyn lisäämä arvo. Mitä suurempi on dokumentin sijaluku tulosjoukossa – millä tahansa relevanssitasolla – sitä vähempiarvoinen dokumentti on käyttäjälle, koska tulosjoukon kasvaessa käyttäjä yhä suuremmalla todennäköisyydellä luopuu sen selaamisesta ajan puutteen, liian suuren vaivan tai jo aikaisemmista dokumenteista tarpeeksi suureksi kumuloituneen hyödyn takia. Tarvitaan funktio, joka alentaa dokumenttien hyötyarvoa progressiivisesti (mutta ei liian jyrkästi, esim. jakamalla sijaluvulla) ja auttaa mallintamaan käyttäjien sitkeyttä uusien dokumenttien selaamisessa tulosjoukon kasvaessa. Yksinkertainen mutta vaatimukset täyttävä tapa arvon alentamiseen on dokumentin hyötyarvon jakaminen dokumentin sijaluvun logaritmilla. Esimerkiksi $^2\log 2 = 1$ ja $^2\log 1024 = 10$, joten tulosjoukon dokumentti sijaluvulla 1024 saisi yhä kymmenesosan alkuperäisestä hyötyarvostaan. Logaritmin kantaluvun valinta vaikuttaa siihen, mallinnetaanko käyttäjien sitkeyttä jyrkemmällä vai loivemmalla arvon alentamisella. Olkoot b logaritmin kantaluku. Määrittelemme alennetun kumuloituvan hyödyn vektorin rekursiivisesti vektoriksi DCG siten että:

$$DCG[i] = \begin{cases} G[1], \text{ jos } i=1 \\ DCG[i-1]+G[i]/{}^b\log i, \text{ muulloin} \end{cases} \quad (2)$$

kun $i = 1 \ldots n$. On huomattava, että logaritmiin perustuvaa hyödyn alennusta ei voida käyttää sijaluvulla 1, koska $^b\log 1 = 0$.

Olkoot esimerkiksi b = 2. Vektorista G' saadaan vektori DCG' = <3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, ...>.

Kyselyn kyky (tai sen puute) lajitella erittäin relevantit dokumentit tulosjoukon kärkeen tulisi näkyä kumuloituvan hyödyn (CG) ja alennetun kumuloituvan hyödyn (DCG) vektoreissa. Tiedonhaun menetelmien keskimääräistä suoritustasoa voidaan analysoida laskemalla keskiarvovektorit hakuaiheita edustavien kyselyjen tuloksille. Keskiarvovektorit voidaan helposti kuvata hyöty-sijaluku-käyrinä.

CG-mitalla on monia etuja verrattuna haun keskipituuteen (ASL – Losee, 1998) tai RR- ja RHL – mittoihin (Borlund & Ingwersen, 1998):

1. Mitta yhdistää dokumentin relevanssin asteen ja sijaluvun (joka määräytyy relevanssin todennäköisyyden perusteella) johdonmukaisella tavalla. Suhteellinen relevanssi (RR) laskee korrelaation järjestelmäperusteisen relevanssin todennäköisyyden ja käyttäjien määrittelemän relevanssin asteen välille. Haun keskipituus perustuu binäärisiin relevanssiarvioihin.

2. Mitta antaa mille tahansa tulosjoukon sijaluvulle (tulosjoukon tarkasteltujen dokumenttien määrälle) arvion kumuloituneesta hyödystä riippumatta saantikannan koosta. Haun keskipituus kertoo vain relevantin dokumentin sijaluvun keskiarvon tietyn kokoiselle saantikannalle. RHL-mitta ilmaisee kumuloituneen relevanssin mediaanin annetussa tulosjoukossa. Mediaani voi olla täysin sama tuloksellisuudeltaan erilaisille kyselyille.

3. Mitta ei ole herkkä ääriarvoille (relevanteille dokumenteille, joiden sijaluku on suuri), koska siinä lasketaan vain tulosjoukon alusta kumuloituva hyöty. Haun keskipituus ja RHL ovat riippuvia ääriarvoista, joskin RHL vähemmän.

4. Mitta on helppo tulkita, se on välittömämpi kuin saanti-tarkkuus–käyrät eikä se peitä huonoa tulosta. RHL yksinään ei ole riittävä tuloksellisuuden mittaamiseen.

DCG-mitta tarjoaa lisäksi seuraavat edut, joita ei ASL- tai RHL-mitoissa ole:

1. DCG alentaa realistisella tavalla tulosjoukossa huonosti sijoittuneiden dokumenttien tuomaa hyötyä.
2. DCG:n avulla voidaan mallintaa erilaisten käyttäjien sitkeyttä suurten, järjestettyjen tulosjoukkojen selaamisessa muuttamalla alentavaa tekijää, eli logaritmin kantalukua.

3. Esimerkkitapaus: kyselyjen laajentamisen ja rakenteiden vaikutus hakujen tuloksellisuuteen eri relevanssitasoilla

Esittelemme ehdottamiemme mittojen käyttöä kokeellisessa tutkimuksessa, jossa testattiin kyselyrakenteiden ja kyselyjen laajentamisen vaikutuksia hakujen tuloksellisuuteen käyttäen moniportaisia relevanssiarvioita. Oletamme aikaisempien tulostemme perusteella, että yleensä heikkorakenteiset kyselyt eivät hyödy laajentamisesta, kun taas vahvarakenteisten kyselyjen tulos paranee (Kekäläinen ja Järvelin, 1998; Kekäläinen, 1999). Nyt tarkasteltavassa kokeessa haluttiin selvittää, eroavatko eri rakenteisten kyselyjen tulokset eri relevanssitasoilla. Esitämme tulokset saanti-tarkkuus-käyrinä eri relevanssitasoilla ja relevanssin astetta hyödyntävinä CG- ja DCG-käyrinä. Hypoteesimme on, että vahvarakenteisten laajennettujen kyselyjen tulosjoukoissa erittäin relevantit dokumentit sijoittuvat paremmin kuin laajentamattomien kyselyjen tai heikkorakenteisten – laajentamattomien ja laajennettujen – kyselyjen tulosjoukoissa. Tästä seuraa myös, että tuloksellisuuden erot kyselytyyppien välillä ovat vähäisiä tarkasteltaessa osittain relevantteja dokumentteja, mutta merkitseviä tarkasteltaessa erittäin relevantteja dokumentteja.

3.1 Testiaineisto

Testi tehtiin tekstitietokannassa, joka sisältää sanomalehtiartikkeleita ja jonka hakujärjestelmänä on todennäköisyyslaskentaan perustuva InQuery–hakujärjestelmä. Tietokannassa on 53893 artikkelia kolmesta eri sanomalehdestä. Tekstien hakuavaimet on tallennettu hakemistoon perusmuotoisina ja yhdyssanat on pilkottu yhdysosiinsa, niin että yhdyssana on tallennettu kokonaan perusmuotoisena ja perusmuotoistettuina osinaan. Testikokoelmaan[2] kuuluu joukko hakuaiheita, jotka ovat 1 - 2 virkkeen pituisia tiedontarpeiden kuvauksia. Hakuaiheisiin on 17337 artikkelin saantikanta, joka jakautuu neljään relevanssitasoon (ks. relevanssiarviot). Saantikanta on koottu yhdistämällä eri tutkimuksiin tehtyjen tuhansien eri kyselyjen tulosjoukot. Kyselyjä on tehty niin osittais- kuin täystäsmäytysjärjestelmiin. Koska kaikkien artikkeleiden relevanssia jokaisen hakuaiheen suhteen ei ole selvitetty, kokeidemme saantiarvot perustuvat *suhteellisen saannin* arvoihin. Otaksumme suhteellisen saannin arviomme luotettaviksi saantikantojen suuruuden ja niiden muodostamisessa käytettyjen testikyselyjen määrän ja monipuolisuuden takia. Kyselyrakenteiden testaukseen valittiin testikokoelman hakuaiheista 30 niiden laajennettavuuden perusteella, toisin sanoen niiden avulla haluttiin tarkastella laajentamisen ja kyselyrakenteiden vuorovaikutusta. (Kekäläinen & Järvelin, 1998, Kekäläinen & Järvelin, 2000.)

InQuery valittiin testin hakujärjestelmäksi, koska sen hakukielessä on runsaasti operaattoreita, muun muassa todennäköisyyteen perustuvat tulkinnat Boolen operaattoreille, ja hakuavainten painottaminen on mahdollista. InQuery perustuu Bayesin päättelyverkkoon. Järjestelmän seikkaperäinen kuvaus löytyy seuraavista lähteistä: Turtle (1990), Rajashekar ja Croft (1994), Allan et al. (1997) ja Kekäläinen (1999).

3.2 Relevanssiarviot

Testin relevanssiarviot teki neljä henkilöä, kaksi kokenutta toimittajaa ja kaksi informaatikkoa. Heille annettiin tiedontarpeiden kirjalliset kuvaukset (hakuaiheet) ja heitä pyydettiin arvioimaan artikkelien relevanssia suhteessa hakuaiheisiin käyttäen neliportaista asteikkoa: (0) epärelevantti, artikkeli ei käsittele hakuaihetta, (1) osittain relevantti, hakuaihe on mainittu artikkelissa, (2) melko relevantti, hakuaihetta käsitellään artikkelissa, mutta lyhyesti tai sivuteemana, (3) erittäin relevantti, hakuaihe on artikkelin pääteema. Kaksi henkilöä arvioi 20 hakuaiheen tulosjoukkojen relevanssin, lopuista hakuaiheista on yhden henkilön arviot. Relevanssiarviot olivat yhtenevät 73 %:ssa rinnakkaisista arvioinneista, 21 %:ssa tapauksista arviot erosivat yhden pisteen verran, 6 % kaksi tai kolme pistettä. Jos ero oli yhden pisteen verran, relevanssiarvio otettiin vuoroin kummaltakin arvioijalta. Jos ero oli kahdesta kolmeen pistettä, tutkija tarkisti artikkelista oliko eroon loogista syytä ja valitsi vaihtoehdoista uskottavimman. (Ks. Sormunen, 1994; 2000; Kekäläinen, 1999)

Tämän tutkimuksen 30 hakuaiheen saantikantaan kuuluu 366 erittäin relevanttia artikkelia (relevanssitaso 3), 700 melko relevanttia artikkelia (relevanssitaso 2), 857 osittain relevanttia artikkelia (relevanssitaso 1).

Loppuja tietokannan artikkelista, joita on 51970, pidetään epärelevantteina (relevanssitaso 0).

3.3 Kyselyrakenteet ja kyselyjen laajentaminen

Tekstitiedonhaussa tiedontarve ilmaistaan tyypillisesti joukkona hakuavaimia. Täystäsmäytyksessä – toisin sanoen Boolen haussa – hakuavainten väliset suhteet ilmaistaan kyselyissä tyypillisesti JA-operaattorilla, TAI-operaattorilla tai läheisyysoperaattoreilla, jotka oikeastaan ovat JA-operaattorin tiukempia muotoja. Kyselyllä on siten rakenne, joka perustuu hakuavainten konjunktioihin ja disjunktioihin. (Keen, 1991; Green, 1995.) Boolen lohkorakenteen mukainen kysely (joka on samalla konjunktiivisessa normaalimuodossa), on esimerkki fasettirakenteesta. Fasetissa kyselyn yhtä aspektia edustavat hakuavaimet on yhdistetty TAI-operaattorilla. Fasetit yhdistetään JA-operaattorilla. Fasetti sisältää yhden tai useampia käsitteitä.

Osittaistäsmäytyksessä dokumentit järjestetään tuloslistaksi relevanssin todennäköisyyttä kuvaavan lajitteluarvon mukaisesti. Lajitteluarvo lasketaan dokumentissa esiintyvien hakuavainten painojen perusteella. Avaimen paino perustuu tyypillisesti avaimen frekvenssiin dokumentissa ja avaimen käänteiseen frekvenssiin tietokannassa. (Ingwersen & Willett, 1995.) Osittaistäsmäytyksessä kyselyjen rakenne voi olla samankaltainen kuin Boolen kyselyjen tai kyselyt voivat olla 'rakenteettomia', toisin sanoen hakuavainten suhteita ei ole eroteltu.

Aikaisemmassa tutkimuksessamme testasimme kyselyjen rakenteen ja laajentamisen yhteisvaikutusta hakujen tuloksellisuuteen ja totesimme kyselyrakenteen merkityksen laajentamisen yhteydessä (Kekäläinen & Järvelin, 1998). Paras tulos saavutettiin laajennetuilla, fasettirakenteeseen perustuvilla kyselyillä. Käsillä olevaan tutkimukseen valitsimme parhaan heikkorakenteisen kyselyn (SUM) ja kaksi parasta vahvarakenteista kyselyä, joista toinen perustuu käsitteisiin (SSYN-C) ja toinen fasetteihin (WSYN). SUM-kyselyt edustavat tyypillisiä osittaistäsmäytyskyselyjä ja siksi niiden tulosta voidaan pitää vertailukohtana muiden kyselyrakenteiden tuottamille tuloksille.

Kyselyjen laajentaminen perustui käsitteellisiin hakusuunnitelmiin. Tutkijat tekivät kyselyt. Kyselyn muotoilu aloitettiin käsitetasolla[3] tunnistamalla hakuaiheen keskeiset käsitteet. Käsitteitä kuvaavat hakuavaimet valittiin hakutesauruksesta, jossa on yli tuhat testin hakuaiheisiin liittyvää käsitettä ja yli 1500 käsitteitä kuvaavaa ilmausta. Kyselyjä laajennettaessa niihin lisättiin hakuavaimet, jotka edustivat alkuperäisiin hakukäsitteisiin semanttisissa suhteissa olevia käsitteitä (synonyymit, hierarkkisia ja assosiatiivisia käsitteitä edustavat avaimet). Näin saatiin kahdenlaisia kyselyversioita, laajentamattomia (u) ja laajennettuja (e), jotka edelleen muotoiltiin eri kyselyrakenteiksi.

Testissä käytetyt kyselyrakenteet kuvataan seuraavien esimerkkien avulla. Esimerkit perustuvat hakuaiheeseen *Ydinvoimalaitosten tuottamien radioaktiivisten jätteiden varastointi*[4]. Esimerkeissä kyselyt ovat laajennettuja, laajentamattomien kyselyjen avaimet on merkitty kursiivilla.

SUM-kyselyt edustavat heikkoja rakenteita. Näissä kyselyissä hakuavaimet olivat yksittäisinä sanoina[5], sanaliittoja ei käytetty.

#sum(*#0(ydin voima laitos)* #0(ydin voimala) #0(ydin reaktori) reaktori #0(ydin voima) #0(ydin energia) *radioaktiivinen jäte* #0(ydinjäte) #0(ydin voima jäte) #0(ydin voimala jäte) käyttää #0(poltto aine) #0(ydin hauta) #0(jäte hauta) *varastointi* varastoiminen varastoida säilytys säilyttäminen säilyttää taltiointi taltioiminen taltioida varasto)

SSYN-C–kyselyissä jokainen hakukäsite muodostaa SYN-operaattoilla yhdistetyn lausekkeen (so. käsitettä edustavat avaimet on yhdistetty SYN-operaattorilla). SYN-lausekkeet yhdistettiin SUM-operaattorilla. Sanaliitot tunnistettiin (merkitty #3-operaattorilla). SYN-operaatiossa kaikkia avaimia käsitellään saman avaimen ilmentyminä (Rajashekar & Croft, 1995).

#sum(**#syn**(*#0(ydin voima laitos)* #0(ydin voimala) #0(ydin reaktori) reaktori #0(ydin voima) #0(ydin energia)) **#syn**(*#1(radioaktiivinen jäte)* #0(ydinjäte) #3(radioaktiivinen jäte) #0(ydin voima jäte) #0(ydin voimala jäte) #3(käyttää #0(poltto aine)) #0(ydin hauta) #0(jäte hauta)) **#syn**(*varastointi* varastoiminen varastoida säilytys säilyttäminen säilyttää taltiointi taltioiminen taltioida varasto))

WSYN-kyselyt olivat SSYN-kyselyjen kaltaisia muuten, mutta perustuivat käsitteiden sijasta fasetteihin. Fasetit jaettiin pää- ja sivufasetteihin sen mukaan, miten tärkeitä ne olivat hakuaiheen kannalta. WSYN-kyselyissä pääfaseteille annettin paino 10 ja sivufaseteille 7.

#wsum(1 10 #syn(*#0(ydin voima laitos)* #0(ydin voimala) #0(ydin reaktori) reaktori #0(ydin voima) #0(ydin energia)) **10 #syn**(*#1(radioaktiivinen jäte)* #0(ydinjäte) #3(radioaktiivinen jäte) #0(ydin voima jäte) #0(ydin voimala jäte) #3(käyttää #0(poltto aine)) #0(ydin hauta) #0(jäte hauta)) **7 #syn**(*varastointi* varastoiminen varastoida säilytys säilyttäminen säilyttää taltiointi taltioiminen taltioida varasto))

3.4 Testikyselyt ja evaluointimittojen soveltaminen

Kyselyjen kompleksisuus (fasettien määrä) vaihteli testin 30 hakuaiheessa 3:sta 5:een ja keskiarvo oli 3,7. SUM-kyselyjen tyhjentävyys (hakuavainten määrä kun sanaliittoja ei käytetty) oli laajentamattomissa versioissa keskimäärin 6,1, ja laajennetuissa 62,3. Tyhjentävyys sanaliittoja käytettäessä (SSYN-C ja WSYN) oli laajentamattomissa kyselyissä keskimäärin 5,4 ja laajennetuissa 52,4.

Esitämme tulokset kahdella tavalla: Ensiksi sovellamme perinteisiä saanti-tarkkuus-käyriä ja esitämme interpoloimattoman tarkkuuden keskiarvot. Kyselyversioiden välisten erojen tilastollinen merkitsevyys on testattu Friedmanin kaksisuuntaisella varianssianalyysilla (ks. Conover, 1980). Toiseksi esitämme CG- ja DCG-käyrät. Kumuloidun hyödyn evaluointia varten testasimme samat kyselyversiot käyttäen relevanssitasoja hyödyn arvioimiseen ja logarimin eri kantalukuja hyödyn alentamiseen seuraavasti:

1. Logaritmin kantalukuja 2, ja 10 testattiin DCG-vektoreiden muodostamisessa. Kantaluku 2 mallintaa kärsimätöntä käyttäjää, kantaluku 10 kärsivällistä.

2. Käytimme relevanssitasoja 0 – 3 suoraan hyödyn mittana. Tätä valintaa voidaan arvostella esimerkiksi siten, että erittäin relevantit artikkelit saattavat olla enemmän kuin kolme kertaa arvokkaampia kuin osittain relevantit artikkelit. Joka tapauksessa tämäkin menettely tekee selvän eron artikkeleiden hyötyarvoon.

3. Otimme ensin huomioon kaikki artikkelit relevanssitasoilla 1 – 3, sitten nollasimme relevanssitasolla 1 olevien dokumenttien hyötyarvon (koska niiden merkitys on käytännössä mitätön), ja lopuksi nollasimme relevanssitasolla 2 olevien artikkeleiden hyötyarvon, jotta saisimme pelkästään erittäin relevantit artikkelit esiin.

4. Todellisia CG- ja DCG-vektoreita verrattiin teoreettisesti parhaaseen mahdolliseen vektoriin, joka muodostettiin seuraavasti. Olkoot annetun kyselyn tulosjoukossa *k*, *l*, and *m* relevanttia artikkelia relevanssitasoilla 1, 2 ja 3. Ensin täytetään vektorin sijat 1 ... *m* arvolla 3, sitten sijat *m*+1 ... *m*+*l* arvolla 2, sitten sijat *m*+*l*+1 ... *m*+*l* +*k* arvolla 1, ja loput sijat arvolla 0. Lopuksi lasketaan CG- ja DCG–vektorit samoin kuin CG- ja DCG-keskiarvovektorit kuten yllä.

3.5 Saanti-tarkkuus –käyrät ja tarkkuuden keskiarvo

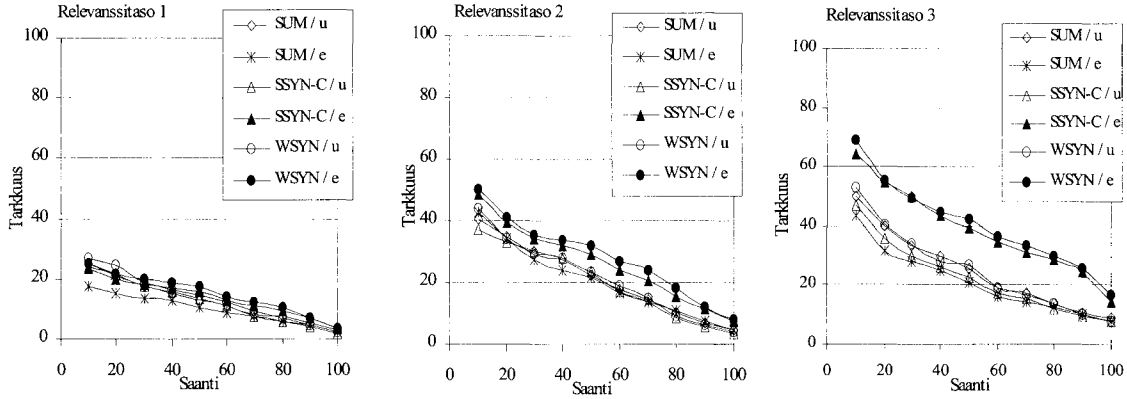
Kuvio 1 esittää kuuden kyselytyypin saanti-tarkkuus-käyrät eri saantitasoilla. Relevanssitasolla 1 käyriä on miltei mahdotonta erottaa toisistaan. Relevanssitasolla 2 laajennettujen SSYN-C ja WSYN –kyselyjen käyrät nousevat tehokkaampina jonkin verran muiden kyselyjen käyrien yläpuolelle. Relevanssitasolla 3 erot kasvavat yhä. Mitä korkeampi relevanssitaso on, sitä suurempi on ero parhaiden ja huonoimpien kyselyjen välillä.

Taulukossa 1 ovat interpoloimattoman tarkkuuden keskiarvot. Kyselyjen laajentaminen ei koskaan paranna SUM-kyselyjen tarkkuuden keskiarvoa. Sitä vastoin vahvarakenteisten kyselyjen tarkkuuden keskiarvo paranee aina kyselyjä laajennettaessa. Kun kyselyt ovat laajentamattomia, erot kyselyjen tarkkuudessa ovat vähäisiä kaikilla relevanssitasoilla. Paras tulos saavutettiin laajennetuilla WSYN-kyselyillä. Parhaimmillaan ero laajentamattomien SUM-kyselyjen ja laajennettujen WSYN-kyselyjen välillä on relevanssitasolla 3 (tarkkuuden keskiarvon muutos 15,1 prosenttiyksikköä tai 58,3 % parannus). Toisin sanoen laajennetut vahvarakenteiset kyselyt ovat tehokkaimpia löytämään erittäin relevantteja artikkeleja.

Friedmanin testi osoitti, että tarkkuuden erot kyselytyyppien välillä ovat merkitsevämpiä relevanssitasolla 3 kuin muilla relevanssitasoilla. Laajennetut vahvarakenteiset kyselyt ovat ylivoimaisia verrattuna laajennettuihin heikkorakenteisiin kyselyihin, mutta myös laajentamattomiin heikko- ja vahvarakenteisiin kyselyihin nähden.

3.6 Kumuloitu hyöty

Kuvio 2 kuvaa CG-vektoreita käyrinä tulosjoukkojen sijaluvuilla 1 – 100. Kuviosta ilmenevät kunkin 6 kyselytyypin suorituskäyrät sekä teoreettisesti paras saavutettavissa oleva keskimääräinen tulos. Kuviossa 2a esitetään käyrät, kun artikkelit relevanssitasoilla 2 ja 3 otetaan huomioon hyötyä laskettaessa. Teoreettisesti paras käyrä muuttuu lähes horisontaaliseksi sijalla 100, mikä osoittaa, että sijalla 100 voidaan periaatteessa löytää



Kuvio1. SUM, SSYN-C ja WSYN –kyselyjen saanti-tarkkuus –käyrät relevanssitasoilla 1, 2 ja 3.

Rel.taso	Laajennos-tapa	Kyselyn rakenne		
		SUM	SSYN-C	WSYN
1	u	12,8	12,4	13,8
	e	10,1	13,3	14,3
2	u	22,4	21,5	22,9
	e	21,1	27,4	29,3
3	u	25,9	23,5	25,7
	e	22,2	39,1	41,0

Taulukko 1. Interpoloimattoman tarkkuuden keskiarvot eri kyselytyypeille.

jo lähes kaikki relevantit artikkelit. Kaksi parasta (rakenteista) kyselytyyppiä sijoittuvat 18 - 27 pistettä (35 - 39 %) teoreettisesti parhaan alapuolelle sijoilla 20 - 100. Ero on suurimmillaan välin 20 – 100 keskivaiheilla. Muut kyselytyypit sijoittuvat edelleen 5 - 15 pistettä (noin 16 - 24 %) alemmaksi sijoilla 20 - 100. Ero teoreettisesti parhaisiin on 23 - 38 pistettä (50 %). Sijan 100 jälkeen erot teoreettisesti parhaan ja muiden käyrien välillä pienenevät, kun lisää relevantteja artikkeleja löytyy. Kuva 2b esittää tuloksen, kun vain relevanssitason 3 artikkelit otetaan huomioon. Lukuarvot ovat erilaisia ja absoluuttiset erot eri kyselytyyppien välillä ovat pienempiä, mutta suhteelliset erot kuitenkin suurempia.

Tulokset voidaan tulkita myös toisin. Relevanssitasolla 3 hakijan täytyy parhaita kyselytyyppejä käyttäessään tutkia 34 artikkelia, ja muita menetelmiä käyttäessään yli 60 artikkelia, saadakseen sen hyödyn, joka teoreettisesti on saatavissa tutkimalla vain 10 hyvin valittua artikkelia. Tässä suhteessa parhaat kyselytyypit ovat lähes kaksi kertaa parempia kuin muut. Relevanssitasoilla 2 - 3 vastaavat luvut ovat 20 ja 26 artikkelia. Suurimmillaan ero parhaiden ja muiden kyselytyyppien välillä on 6 – 8 pistettä (tai 2 artikkelia, relevanssitaso 3) sijoilla 40 – 60. Relevanssitasoilla 2 - 3 suurimmat erot 5 - 15 pistettä (tai 2 - 7 artikkelia) sijoilla 40 - 100.

3.7 Alennettu kumuloitu hyöty

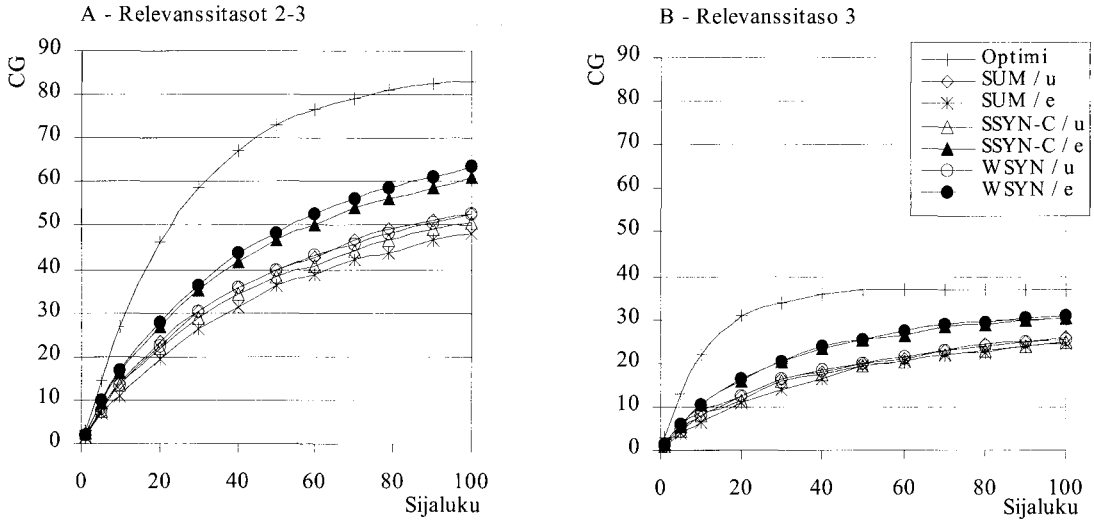
Kuvio 3 kuvaa DCG-vektoreita käyrinä tulosjoukkojen sijaluvuilla 1 – 50. Kuviosta ilmenevät kunkin 6 kyselytyypin suorituskäyrät sekä teoreettisesti paras saavutettavissa

oleva keskimääräinen tulos. Hyödyn alennukseen käytetään sijaluvun logaritmia, jonka kantalukuna on 2 ($\log_2$). Kuviossa 3a esitetään käyrät, kun artikkelit relevanssitasoilla 2 ja 3 otetaan huomioon hyötyä laskettaessa. Teoreettisesti paras käyrä kasvaa edelleen sijalla 50 (se tasoittuu horisontaaliseksi sijalla 90). Kaksi parasta (rakenteista) kyselytyyppiä sijoittuvat 5 - 9 pistettä (35 - 36 %) teoreettisesti parhaan alapuolelle sijoilla 10 – 50 ja ero kasvaa. Muut kyselytyypit sijoittuvat edelleen 2 - 4 pistettä (noin 15 - 27 %) alemmaksi sijoilla 10 - 50. Ero teoreettisesti parhaisiin on 7 - 13 pistettä (47 - 50 %). Sijan 50 jälkeen erot teoreettisesti parhaan ja muiden käyrien välillä vakiintuvat, kun myöhemmin löytyvien relevanttien artikkeleiden vaikutus käyriin on vähäinen. Kuva 3b esittää tuloksen, kun vain relevanssitason 3 artikkelit otetaan huomioon. Lukuarvot ovat erilaisia ja absoluuttiset erot eri kyselytyyppien välillä ovat pienempiä, mutta suhteelliset erot kuitenkin suurempia. Suurimmillaan erot parhaiden ja muiden kyselytyyppien välillä ovat 3 pistettä (tai yksi 3-tason artikkeli) sijalla 40 ja sen jälkeen. Se on pysyvä ja tilastollisesti merkitsevä ero, mutta huomaavatko käyttäjät sitä lainkaan?

Myös nämä tulokset voidaan tulkita toisin. Relevanssitasoilla 2-3 hakijan täytyy parhaita kyselytyyppejä käyttäessään tutkia 35 artikkelia, ja muita menetelmiä käyttäessään yli 60 artikkelia, saadakseen sen alennetun hyödyn, joka teoreettisesti on saatavissa tutkimalla vain 10 hyvin valittua artikkelia. Ei ole epärealistista odottaa hakijan tutkivan jopa 35 artikkelia, mutta 70 artikkelin tutkimisen täytyy olla jo harvinaista. Kyselytyyppien väliset erot ovat olennaisia. Relevanssitasolla 3 parhaiden kyselytyyppien tuottamat alennetut hyödyt eivät koskaan saavuta sitä hyötyä, joka on saavutettavissa tutkimalla 10 hyvin valittua artikkelia. Viiden hyvin valitun artikkelin tuottama hyöty saavutetaan vain parhailla menetelmillä ja vasta 50 artikkelia tutkimalla.

Epäilevä saattaisi väittää, että jos tiedonhakija tutkii jopa 70 dokumenttia, hän saa niiden todellisen arvon hyväkseen, eikä vain alennettua arvoa – ja tämän takia DCG-tuloksia ei voitaisi käyttää tiedonhakumenetelmien tuloksellisuuden vertailuun. Käyttäjän kannalta näin onkin, mutta silti DCG-vertailu on mielekäs tiedonhakumenetelmien suunnittelijalle. Mitä kauempana tuloslistalla relevantit artikkelit ovat, sitä todennäköisempää on, ettei tiedonhakija niitä koskaan katso. Siksi niille ei pidä antaa vertailussa niiden todellista arvoa, vaan alennettu arvo. Hakujärjestelmää tai – menetelmää pitää palkita artikkeleiden hyvästä sijoittamisesta paremmin kuin huonosta.

Päätulokset pysyvät samoina, vaikka DCG-käyrien laskennassa käytettyä logaritmia muutettaisiin. Parhaiden ja huonompien kyselytyyppien erojen suuruusluokka kasvaa kuitenkin 4 pisteestä logaritmilla $\log_2$ 13 pisteeseen logaritmilla $\log_{10}$ sijalla 50 (tämä seuraa suoraan logaritmien suhteesta). Voimme



Kuvio 2. Kumuloituvan hyödyn käyrät tulosjoukon sijoilla 1-100, relevanssitasoilla 2 - 3 ja 3.

kuitenkin tulkita eroja siten, että sitkeälle käyttäjälle parhaat kyselytyypit ovat 13 pistettä (tai 27%) parempia kuin jäljellejäävät. Kärsimättömälle hakijalle ne ovat vain 4 pistettä parempia.

4 Yhteenveto

Tiedonhakumenetelmien kehittäminen suuria nykyaikaisia tietokantaympäristöjä varten vaatii, että menetelmien kyky löytää erittäin relevantteja dokumentteja voidaan arvioida. Tämä on käyttäjien kannalta keskeistä eikä anna "synninpäästöä" tiedonhakumenetelmille liian helposti. Olemme tässä artikkelissa esittäneet kaksi uutta evaluointimenetelmää, joiden avulla voidaan ottaa tiedonhaussa löytyvien artikkelien relevanssin aste huomioon. Ensimmäinen menetelmä perustuu uuteen tapaan käyttää perinteisiä saanti-tarkkuus-käyriä ja erillisiä saantikantoja kullakin artikkeleiden relevanssitasolla. Toinen menetelmä perustuu kahteen uuteen mittariin, CG ja DCG, jotka antavat hakutuloksen (alennetun) kumuloidun hyödyn valittuun tuloksen sijalukuun saakka. Molemmat mittarit yhdistävät dokumentin sijan (sen relevanssin todennäköisyyden perusteella) ja relevanssin asteen systemaattisella ja järkevällä tavalla.

Tapausesimerkissämme sovelsimme evaluointimenetelmiämme kuuden kyselytyypin tuloksellisuuden evaluointiin todennäköisyyslaskentaan perustuvassa tiedonhaussa. Kyselytyypit erosivat toisistaan kyselyjen rakenteen ja laajennuksen asteen suhteen. Tutkimushypoteesejamme olivat:

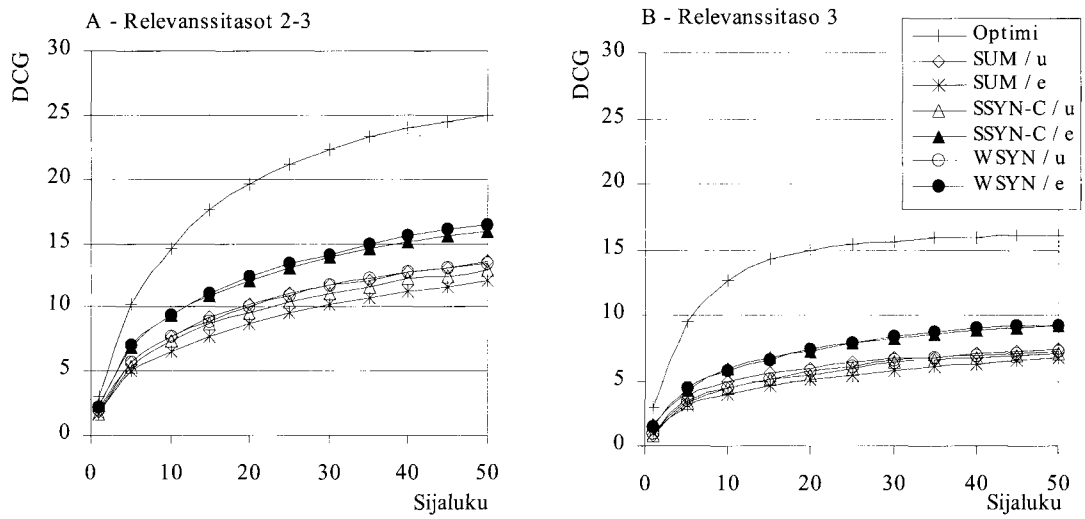
(1) Suorituskykyerot kyselytyyppien välillä haettaessa osittain relevantteja artikkeleita ovat vähäisiä, ja haettaessa erittäin relevantteja dokumentteja olennaisia, kun eroja tarkastellaan saanti-tarkkuus-tuloksilla ja -käyrillä.

(2) Laajennetut vahvarakenteiset kyselyt tuottavat paremman tuloksen kuin laajentamattomat tai muun rakenteiset kyselytyypit laajentamisesta riippumatta.

(3) Laajennetut vahvarakenteiset kyselyt tuottavat korkeammat CG- ja DCG-arvot kuin laajentamattomat tai muun rakenteiset kyselytyypit laajentamisesta riippumatta.

Tutkimustulokset osoittivat hypoteesit tosiksi. Suorituskykyerot parhaiden ja huonompien kyselytyyppien välillä ovat systemaattisia ja tilastollisesti merkitseviä. Esimerkissämme kuitenkin annoimme eri relevanssin astetta oleville artikkeleille varsin lähellä toisiaan olevat hyötyarvot (0 – 3 pistettä). Todellisuudessa tiedonhakijan mielestä erittäin relevanttien artikkeleiden hyötypistemäärä saattaisi olla selvästi korkeampi kuin muilla relevanssitasoilla. Vaikka tuloksemme ovatkin merkitseviä, ne saattavat kuitenkin olla varsin varovaisia.

Esimerkkitapauksemme saanti-tarkkuuskäyrät osoittavat, että laajennettujen vahvarakenteisten



Kuvio 3. Alennetun (log_2) kumuloituvan hyödyn käyrät tulosjoukon sijoilla 1-50, relevanssitasoilla 2 -3 ja 3.

kyselyjen hyvä suorituskyky johtuu ennen kaikkea niiden kyvystä sijoittaa erittäin relevantteja artikkeleita hakutuloslistan kärkipäähän. Kumuloituvan hyödyn käyrät osoittavat käyttäjän saamaa todellista hyötyä, ja alennetun kumuloidun hyödyn käyriä voidaan käyttää ennustamaan hakumenetelmien suorituskykyä suhteessa otaksuttuun tiedonhakijan kärsivällisyyteen tuloslistojen tutkimisessa. Käyttämällä pientä logaritmin kantalukua relevantin artikkelin tuottama hyöty laskee nopeasti artikkelin sijaluvun kasvaessa – DCG-käyrä kääntyy vaakasuoraan. Tämä mallintaa kärsimättömän hakijan toimintaa, jossa myöhään sijoitettu informaatio ei ole hyödyllistä, koska sitä ei koskaan lueta. Jos CG- ja DCG-käyriä luetaan horisontaalisesti, voimme todeta, että tiedonhakujärjestelmän suunnittelijan pitäisi otaksua hakijoiden tutkivan 50 – 100 % enemmän artikkeleita huonommilla kyselytyypeillä kuin parhailla kyselytyypeillä kootakseen saman hyödyn. Vaikka näin voidaankin otaksua, on tällainen käytös varmasti suhteellisen harvinaista.

Olemme myöhemmässä tutkimuksessamme laajentaneet hypoteesejamme koskemaan artikkeleiden tekstien tilastollisia ominaisuuksia ja pyrkineet siten selittämään sen, miksi vahvarakenteiset kyselyt tuottavat parhaat tulokset. Artikkelissamme (Sormunen & al., 2001) osoitamme, että erittäin relevanteissa uutisdokumenteissa tarkastellaan hakupyynnön aihetta laajemmin, useammasta näkökulmasta, useampaa käsitettä käyttäen ja monisanaisemmin kuin vähemmän relevanteissa dokumenteissa. Laajennetut vahvarakenteiset kyselyt ennustavat tällaisten dokumenttien relevanssin todennäköisyyden suuremmaksi kuin muut kyselytyypit.

Yleisempänä havaintona toteamme, että evaluointimenetelmämme ja tapausesimerkkimme osoittavat, että moniportaisia relevanssiarvioita voidaan käyttää tiedonhakumenetelmien evaluoinnissa hyväksi ja että näin saattaa paljastua mielenkiintoisia ilmiöitä. Binääriset relevanssiarviot saattavat johtaa siihen, että tiedonhakumenetelmiä arvioidaan liian löysin ja sallivin perustein ja menetelmien välisiä eroja ei havaita. Tämä tietysti riippuu myös siitä, kuinka helposti binäärisessä relevanssiarvioinnissa dokumentti arvioidaan relevantiksi.

Kiitokset. Tutkimuksemme on saanut rahoitusta Suomen Akatemialta (tutkimusprojekti nro 44704). Kiitämme Timo Tervolaa ohjelmointityöstä ja Tampereen yliopiston FIRE-tutkimusryhmää neuvokkaista kommenteista.

Hyväksytty julkaistavaksi 8.11.2000.

Lähteet

Allan, J., Callan, J., Croft, B., Ballesteros, L., Broglio, J., Xu, J. & Shu, H. (1997a). INQUERY at TREC 5. In E. M. Voorhees & D. K. Harman (Eds.), *Information technology: The Fifth Text Retrieval Conference (TREC-5).* Gaithersburg, MD: National Institute of Standards and Technology, 119–132.

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.

Borlund, P. & Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York: ACM, 324–331.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: John Wiley & Sons.

Green, R. (1995). The expression of conceptual syntagmatic relationships: A comparative survey. *Journal of Documentation 51*(4), 315–338.

Hersh, W. R. & Hickam, D. H. (1995). An evaluation of interactive Boolean and natural language searching with an online medical textbook. *Journal of the American Society for Information Science, 46*(7): 478–489.

Ingwersen, P. & Willett, P. (1995). An introduction to algorithmic and cognitive approaches for information retrieval. *Libri, 45,* 160–177.

Järvelin, K. (1995). Tekstitiedonhaku tietokannoista: johdatus periaatteisiin ja menetelmiin. Espoo: Suomen ATK-kustannus.

Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In: Belkin, N. & Ingwersen, P. & Leong, M-K. (Eds.), *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '00),* Athens, Greece, July 24-28, 2000. New York, NY: ACM Press, pp. 41-48. (Receiver of the ACM SIGIR '00 Best Paper Award).

Keen, E. M. (1991). The use of term position devices in ranked output experiments. *Journal of Documentation, 47*(1), 1–22.

Kekäläinen, J. (1999). *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval.* Ph.D. dissertation. Acta Universitansis Tamperensis 678. Tampere: TAJU.

Kekäläinen, J. & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *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.

Losee, R.M. (1998). *Text retrieval and filtering: Analytic models of performance.* Kluwer Academic Publishers: Boston.

Rajashekar, T. B. & Croft, W. B. (1995). Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American Society for Information Science 46*(4), 272–283.

Robertson, S. E. & Belkin, N. J. (1978). Ranking in principle. *Journal of Documentation 34*(2), 93–100.

Saracevic, T., Kantor, P., Chamis, A. & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science 39*(3): 161–176.

Smithson, S. (1994). Information retrieval evaluation in practice: A case study approach. *Information Processing & Management 30*(2): 205–221.

Sormunen, E. (1994). *Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanomalehtiaineistoa sisältävässä tekstikannassa* [Free-text searching efficiency and factors affecting it in a newspaper article database]. VTT Julkaisuja 790. Espoo: Valtion Teknillinen Tutkimuskeskus. [In Finnish.]

Sormunen, E. (2000). *A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases.* Ph.D. dissertation. Acta Electronica Universitatis Tamperensis. Tampere: University of Tampere. URL: http://granum.uta.fi/pdf/951-44-4732-8.pdf.

Sormunen, E. , Kekäläinen, J., Koivisto, J. & Järvelin, K. (2001). Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. *Journal of Documentation, 57*: xx-yy, to appear.

Turtle, H. R. (1990). *Inference networks for document retrieval.* Ph.D. dissertation. Computer and information Science Department, University of Massachusetts. COINS Technical Report 90–92.

Viitteet:

[1] Relevanssin asteen ja todennäköisyyden erosta, ks. Robertson & Belkin, 1978.

[2] Tiedonhauntutkimuslaboratorion testikokoelma, ks. http://www.info.uta.fi/tutkimus/labra.html

[3] Tiedonhaun kolme tasoa – käsitteellinen, kielellinen ja merkkien taso – ks. Järvelin, 1995.

[4] Sekä hakuaihe että esimerkkikyselyt ovat lyhennettyjä.

[5] #0-operaattori osoittaa että sanat yhdyssanan yhdysosia, sanassa voi lisäksi esiintyä muitakin yhdysosia.