Tiedonhaun tutkimuksen nykyvirtauksia

1. SIGIR '91 konferenssi 13–16. 10. 1991

SIGIR '91 -konferenssi (14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval) pidettiin viime vuonna Chicagossa. Konferenssin pääjärjestäjänä toimi ACM SIGIR (Association for Computing Machinery / Special Interest Group on Information Retrieval, USA), joka on ACM:n tiedonhakututkimuksen jaosto. Järjestelyihin osallistuivat myös monet muut tiedonhakututkimuksesta kiinnostuneet yhteisöt ja sponsorit.

Konferenssin sisältöä voidaan hyvin kuvata sen istuntojen otsikoiden avulla. Monissa esitelmissä raportoitu tutkimus liittyi aiheeltaan luonnollisesti useampaankin seuraavista otsikoista:

- Biennal SIGIR Award Acceptance Speech -istunto,
- Dokumenttien rakenteen jäsentäminen,
- Tiedonhakujärjestelmien mallintaminen I - II,
- Tiedon tiivistäminen,
- Hajautetut hakujärjestelmät I - II,
- Käyttöliittymät,
- Toimistoautomaatio ja tietokannat,
- Oliolähestymistavat tiedonhaussa,
- Semanttiset mallit,
- Hakumenetelmät,
- Hypertekstijärjestelmät,
- Luonnollisen kielen käsittely,
- Paneeli: SMART-järjestelmä 30 vuotta.

Konferenssissa esitettiin seuraavat kolme kutsuttua esitelmää. Biennal SIGIR Award Acceptance Speech -esitelmän piti tiedonhakututkimuksen grand-old-man Cyril W. Cleverdon kiitoksena saamastaan tunnustuksesta. Hän tarkasteli 50- ja 60-luvuilla tehtyjen nk. Cranfield-kokeiden merkitystä ja antia tiedonhakututkimukselle. Hän oli näiden kokeiden päätutkija.

Michael Lesk, joka oli SMART-systeemin alkuperäisiä kehittäjiä Prof. Saltonin kanssa 60-luvulla, tarkasteli The CORE Electronic Chemistry Library -systeemiä, joka tarjoaa 10 vuosikertaa American Chemical Societyn lehtiä kokotekstitietokantana (ilman kuvia) ja lisäksi samat lehdet sivu sivulta kokosivukuvatietokantana. Järjestelmällä tutkitaan käyttäjien mieltymyksiä eri tyyppisten käyttöliittymien suhteen sekä heidän kykyään ja tehokkuuttaan kemiallisten tehtävien (ongelmien) ratkaisemisessa eri käyttöliittymien avulla.

Roger C. Schank on tunnettu tekoälytutkija ja tutkinut erityisesti luonnollisen kielen ymmärtämistä, oppimista, muistamista ja tapauspohjaista päättelyä. Hän esitelmöi opetukseen tarkoitetuista vuorovaikutteisista järjestelmistä, jotka sisältävät suuria tapauskirjastoja erilaisista oppimistilanteista ja joiden käyttöliittymä perustuu hypermediaan (integroivat ääntä, tekstiä, still- ja videokuvaa).

Konferenssin päätösistunto oli SMART-systeemin 30-vuotissyntymäpäiväistunto, jossa Prof. Salton muisteli menneitä, kehui luomaansa paradigmaa ja esitteli kollegojaan vuosien varrelta.

Keskusteluissa esitelmien jälkeen näkyi selvä poleeminen rintamalinja kahden keskeisen suuntauksen välillä: Prof. Saltonin aikanaan aloittama tekstien tilastolliseen käsittelyyn perustuva hakujärjestelmätutkimus ja tekstien semantiikkaa jäsentämään ("ymmärtämään") pyrkivä hakujärjestelmätutkimus. Kritiikkiä esitti aktiivisesti lähinnä Prof. Salton, joka piti semanttista suuntausta epärealistisena ja sen puolesta saatua evidenssiä epäuskottavana ja/tai riittämättömänä. Hän vaati myös sellaista näyttöä muilta suuntauksilta, jota hän ei tarjonnut edes oman tutkimuksensa tueksi. Tilanteen huvittavuutta lisää se, että kaupallisen tiedonhakutoiminnan piirissä tutkimus- ja kehitystyötä tekevistä jokseenkin kaikki pitävät (ja ovat kauan pitäneet) Saltonin suuntausta epärealistisena ja käytännössä toimimattomana. Prof. Croftin esitelmän ("Fraasien ja rakenteisten kyselyjen käyttö tiedonhaussa") perusteella näyttää siltä, ainakin muutamat Saltonin empiirisesti todennetut havainnot tehokkaista vs. toivottomista hakutekniikoista pitävät paikkansa vain Saltonin käyttämissä melko vaatimattoman kokoisissa tietokannoissa: niissä eivät vaativampien tekniikoiden edut pääse esiin.

Kaupallisen tiedonhakutoiminnan tutkimus- ja kehitystyön piirissä on pitkään suhtauduttu vahvasti epäillen kaikkien tilastolliseen käsittelyyn perustuvien hakutekniikoiden toimivuuteen käytännössä. Nyt kuitenkin näyttää siltä, että tilastollisen käsittelyn yleistykset ja laajennokset (esim. juuri Prof. Croftin todennäköisyyskäsitteeseen perustuva hakujärjestelmä) kykenevät pian käsittelemään laajoja tekstitietokantoja. Tämä johtuu ohjel-

misto- ja laitteistotekniikan edistysaskeleista (tehokkaat UNIX/C -työasemat).

Konferenssin osanottajat olivat varsin yksimielinen seuraavista kysymyksistä:

- Hakutulosten järjestäminen todennäköisen relevanssiin mukaan laskevaan järjestykseen on olennaista ja tämän takia Boolen logiikkaan perustuvat perinteiset järjestelmät ovat riittämättömiä.
- Boolen logiikkaan dikotominen täsmäytys (täsmää tai ei täsmää) on riittämätön.
- Relevanssipalaute kohentaa olennaisesti hakutuloksia ja siksi sen käyttö on välttämätöntä.

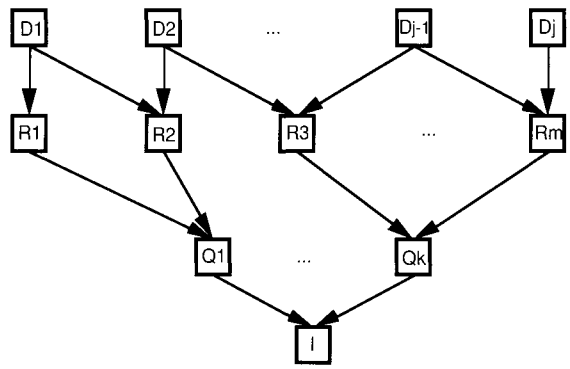
Vaikka tietokannan hakukone siis toimisikin Boolen logiikan puitteissa, pitäisi tulokset järjestää todennäköisen relevanssiin mukaan laskevaan järjestykseen. Jokin relevanssipalautemenetelmä on välttämätön, vaikka Saltonin tutkimusryhmän kehittämä menetelmä ei olekaan ainoa mahdollinen.

Konferenssin proceedings (Bookstein et al., 1991) ja kahden tutoriaalin aineisto (Natural Language Processing and Information Retrieval ; An Overview of Information Retrieval Techniques) on hallussani ja lainaan niitä tarvittaessa lyhyeksi aikaa. Proceedings löytynee myös useimpien suomalaisten tietojenkäsittelyopin laitosten kirjastoista, koska julkaisija on ACM.

Konferenssin aikana tapasin Prof. Edward A. Foxin (Computer Science Dept. / Virginia Institute of Technology & State University). Hän on ollut Prof. Saltonin tutkimusryhmän pitkäaikainen jäsen ja sittemmin tutkinut tiedonhaun vahvistamista tekoälytekniikoilla CODER-järjestelmän puitteissa. Sain häneltä laitoksellamme käytettäväksi tiedonhakututkimuksissa yleisesti käytettyjä standardoituja tekstitietokantoja sisältävän CD-ROM -levyn.

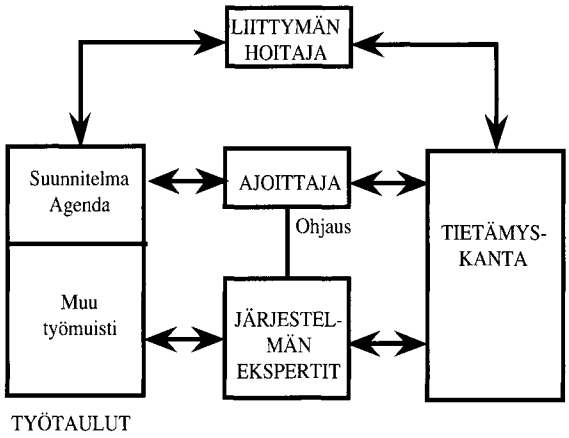
2. *Department of Computer and Information Science, University of Massachusetts, Amherst*

Prof. Croftin tiedonhaun tutkimusryhmä on nykyisin yksi arvostetuimmista ja tuotteliaimmista tiedonhaun tutkimuksen alueella. Prof. Croftin johdolla kehitetty tiedonhakujärjestelmä ja käyttöliittymä INQUERY (ennen nimeltään I^{3}R-hakujärjestelmä (Croft & Thompson, 1987)) perustuu todennäköisyyslaskennan ja tekoälytekniikoiden käyttöön. Varsinainen tietokantahaku perustuu siinä Saltonin kehittämän ns. vektorimallin laajennokseen todennäköisyyslaskennan kehyksessä. Tällöin kyselyt ja dokumentit tulkitaan suunnatun päättelyverkon solmuiksi. Verkossa solmuja yhdistävät hakulausekkeet, hakutermit, dokumenteista johdetut termit ja dokumenttien sisältämät termit. Solmujen väliset linkit ja niihin liittyvät todennäköisyydet kuvastavat evidenssiä (näyttöä) siitä todennäköisyydestä, jolla johdetut termit esiintyvät teksteissä, vastaavat hakutermejä, täsmäävät hakulausekkeisiin ja lopulta koko kyselyyn. Useita erilaisia hakustrategioita voidaan käyttää yhtä aikaa haussa, ja jokainen niistä voi tuottaa kumuloituvaa evidenssiä dokumenttien todennäköisestä relevanssista kysymysten suhteen. Hakujärjestelmä on yleistetty siten, että se saadaan sopivilla ohjaustiedoilla toimimaan samoin kuin Saltonin tilastollinen vektorimalli ja toisilla ohjaustiedoilla Boolen logiikan mukaisesti. Kuva 2.1 alla sisältää esimerkin nelitasoisesta päättelyverkosta.



Kuva 2.1. INQUERYn nelitasoinen päättelyverkko. D1-Dj edustavat dokumentteja, R1-Rm ovat käsitesolmuja, Q1-Qk kyselysolmuja ja I edustaa tiedontarvitsijan tarvetta (Croft & Turtle & Lewis, 1991)

INQUERY -järjestelmän käyttöliittymä perustuu asiantuntijajärjestelmätekniikkaan. Järjestelmä sisältää useita erikoistuneita eksperttejä (käyttäjän mallittaja, kyselyn mallittaja, indeksointiekspertti, tesaurusekspertti, haun ohjaaja, selailun ohjaaja, selittäjä), jotka kommunikoivat ositetun työtauluarkkitehtuurin (blackboard architecture) välityksellä (Kuva 2.2.). Haun ohjaaja on ekspertti, joka hallitsee todennäköisyyslaskentaan perustuvan haun.



Kuva 2.2. INQUERYn yleisrakenne.

INQUERY -järjestelmässä käytetään myös tiedontarvitsijalta kerättyä aihealuetietämystä hakukysymysten parantamiseen sekä uutta muunnelmaa relevanssipalautteesta, joka on Prof. Saltonin ryhmässä kehitettyä hienovaraisempi ja tehokkaampi. Näiden keinojen vaikutuksia hakutulosten parantumiseen raportoitiin alustavasti vuoden 1990 SIGIR-konferenssissa (Turtle & Croft, 1990) ja tarkemmin pian julkaistavissa lehtiartikkeleissa, joiden käsikirjoituksia sain käyttööni (Krovetz & Croft, 1991; Turtle & Croft, 1991). Tämän vuoden konferenssiesitelmässä osoitetttiin, että fraasit parantavat hakujen tuloksellisuutta ja että manuaalisesti ja automaattisesti valitut fraasit toimivat lähes yhtä hyvin, kunhan testit suoritetaan riittävän suurissa tietokannoissa.

INQUERY -järjestelmässä haun tarkkuus parani perusmuotoiseen vektorihakuun (tf.idf – termifrekvenssi/käänteinen dokumenttifrekvenssi) verrattuna 30 % (37 -> 48 %) järjestetyn hakutuloksen 10 ensimmäisen dokumentin joukossa, kun käytettiin aihealuetietämyksen yksittäisiä termejä ja asiakkaan antamia painoja (20 kyselyä, CACM-kokoelma 1958-85). Relevanssipalautteessa päästiin selvästi parempaan tulokseen, jos kaikkien termien lisäämisen sijasta asiakas valitsee lisättävät (siis relevantit) termit. Käyttäjät kykenevät tarjoamaan aihealuetietämystä ja sitä pystytään hyödyntämään hakumekanismissa. Relevantit välitulokset osoittautuivat hyviksi lähteiksi valikoitaville relevanssipalautetermeille. (Croft & Das, 1990)

Fraasit ja rakenteiset kyselyt, jotka sisältävät Boolen- ja/tai läheisyysoperaattoreita, parantavat hakujen tuloksellisuutta. Käsin valittujen ja automaattisesti tunnistettujen fraasien välillä ei ollut merkittävää eroa tuloksellisuuden parantumisen suhteen. Fraasit ja läheisyysoperaattoreita sisältävät rakenteiset kyselyt lienevät sitä tehokkaampia, mitä suurempia testitietokantoja käytetään. Aikaisemmissa testeissä niillä ei ole havaittu olevan merkitystä hakujen tuloksellisuuden kannalta, koska on käytetty liian pieniä testikantoja. (Croft & Turtle & Lewis, 1991)

******

Prof. Croftin johtama aivan viimeaikainen tutkimustyö on keskittynyt hakumenetelmien kehittämiseen ja käyttöliittymän kehittäminen on jäänyt vähemmälle huomiolle. Prof. Croft on saanut tänä vuonna suuren, kolmivuotisen tutkimussopimuksen INQUERY -järjestelmän kehittämiseen. Apuraha kattaa 12 hengen palkat kolmeksi vuodeksi. Projektiin palkataan tällä rahalla mm. 4 kokenutta C-kielen ohjelmoijaa. Järjestelmän nykyinen versio kykenee käsittelemään tekstitietokantoja aina 500 megatavuun asti. Kehitettävän version on tarkoitus kyetä käsittelemään jopa 4 gigatavun tekstitietokantoja sekä englannin että japanin kielellä. Japanin kieli on valittu toiseksi testikieleksi, jotta voidaan vakuuttua kehitettävien menetelmien yleispätevyydestä. (Croft, 1991)

Croftin tutkimusryhmällä on runsaasti kokemusta hakumenetelmien ja käyttöliittymien testaamisesta. Keskustelussa sain arvokasta tietoa testien järjestämisestä ja löydösten tilastollisen merkitävyyden testaamisesta. Mm. sain selityksen sille, ettei tilastollisia merkitsevyystestejä kovinkaan usein esitetä tutkimusraporteissa: tärkeimmät käytetyt testit antavat merkitsevän tuloksen aina, jos verrattavien hakumenetelmien ero on vaikka pienikin, mutta systemaattinen. Niinpä merkitsevyyttä kannattaakin lähestyä toisella tavalla. Tutustuin myös uusiin testiasetelmiin: kokonaisten hakutulosten tutkimisen sijasta testataankin vain 10-20 parhaan (siis relevanssin todennäköisyyden mukaan järjestettynä 10-20 ensimmäisen) hakutuloksen relevanssia. Tämä on monen tiedontarvitsijan kannalta kaikkein mielekkäin testiasetelma.

INQUERY -järjestelmän kyselyn mallittaja ja tesaurusekspertti käyttävät tiedontarvitsijalta kerättyä aihealuetietämystä hakukysymysten parantamiseen. Tämä liittyy läheisesti laitoksellame meneillään olevaan Jaana Kristensenin tutkimusprojektiin, jossa tutkitaan ns. hakutesauruksen käyttöä tekstihaun tulosten parantajana. Prof. Croft oli kiinnostunut havainnosta, että hakutesaurus paran-

taa selvästi hakutuloksen saantia heikentämättä sen tarkkuutta kohtuuttomasti. Nämä tulokset on saatu Boolen logiikkaan perustuvassa hakujärjestelmässä. Croft esitti, etteivät tulokset INQUERYn kaltaisessa järjestelmässä välttämättä ole samanlaisia, ja toivoi, että voisimme tämän seikan tutkia. Oma hypoteesini on, että saantia voidaan oleellisesti kohentaa ilman, että hakutuloksen 10-20 parhaan (ensimmäisen) dokumentin tarkkuus laskee lainkaan. Pikemminkin se nousee.

Tampereen yliopiston Kirjastotieteen ja informatiikan laitokselle perustetun tiedonhaun tutkimuslaboratorion käynnistämisessä on ollut vaikeuksia, koska kaikki tutkimamme kaupalliset tekstihakuohjelmistot ovat neuvottelujen edetessä osoittautuneet liian kalliiksi (lähes 200 000 mk). Prof. Croft lupasi INQUERY-ohjelmiston lähdekielisenä dokumentaatioineen kaikkineen ilmaiseksi käyttöömme aivan lähiaikoina. Samalla saamme suurehkon TIME-lehden artikkeleita sisältävän standardoidun testitietokannan käyttöömme. Tätä tekstiä viimeisteltäessä ohjelmisto on jo saapunut ja asennettu laboratorion SUN SparcStation-laitteistoon.

3. National Library of Medicine ja Lister Hill Center for Biomedical Communication, Bethesda, MD , USA

Tutustuin Bethesdassa Unified Medical Language System (UMLS) -projektiin ja MedIndEx-projektiin, jonka entinen nimi on IAP (The Indexing Aid Project) -projekti. National Library of Medicine (NLM) on USA:n lääketieteellinen keskuskirjasto, joka on panostanut voimakkaasti tiedonhakujärjestelmien kehittämiseen 60-luvulta lähtien. Suuri osa tutkimus- ja kehitystyöstä tapahtuu yhteistyössä Lister Hill Center for Biomedical Communicationin kanssa. Molemmat projektit toimivat molemmissa organisaatioissa.

UMLS-projekti: UMLS (Unified Medical Language System) -järjestelmän kehittäminen alkoi NLM:n (National Library of Medicine, Bethesda, MD) pitkän aikavälin tavoitteista (1) helpottaa monesta lähteestä saatavan lääketieteellisen tiedon välittämistä ja käyttöä, (2) parantaa MeSH -tesauruksen (Medical Subject Headings) ominaisuuksia lääketieteellisen kirjallisuuden tallennuksessa ja haussa sekä (3) luoda käännösmekanismi lääketieteellisten sanastojen välille.

UMLS on yritys luoda älykäs järjestelmä, joka ymmärtää biolääketieteen termejä ja niiden keskinäisiä suhteita ja kykynee siten auttamaan tiedontarvitsijoita hakemaan ja järjestämään tieto- ja tietämyskannoissa olevaa tietoa. UMLS pyrkii tukemaan tiedon integrointia monenlaisista lähteistä, kuten biolääketieteen kirjallisuusviitekannat, potilaskertomukset, faktakannat ja lääketieteelliset tietämyskannat. Se yhdenmukaistaa eri lähteiden ja eri tiedontarvitsijoiden vaihtelevaa sanankäyttöä.

UMLS ei ole yritys kehittää yhtä standardoitua sanastoa lääketieteen tiedonhakua varten. Se ei myöskään ole suunnitelma potilaskertomusten muodon standardointiin eikä lääketieteellisten tietämyskantojen rakentamiseen (UMLS, 1990)

UMLS tulee sisältämään ainakin kolme tietämyslähdettä (UMLS, 1990): Metatesaurus sisältää tietoa biolääketieteen käsitteistä ja niiden esitystavoista eri sanastoissa ja teseuruksissa. Se antaa myös termien tyypit (kategoriat). Se tukee tiedontarvitsijan termien kääntämistä sopivien sanastojen termeiksi. Meta-1 kattaa yli 66.000 käsitettä ja yli 100.000 termiä. Semanttinen verkko sisältää tietoa termityypeistä tai kategorioista metatesauruksessa sekä tietoa tyyppien välillä sallituista suhteista (esim. virus voi aiheuttaa taudin tai oireryhmän). Se ei sisällä varsinaisia termejä. Tietolähdehakemisto tarjoaa kuvauksia ihmisten ja ohjelmien käyttöön kaikentyyppisten biolääketieteen tietokantojen katteesta, sijainnista, sanastoista, syntaktisista säännöistä, ja käyttöehdoista.

Metatesauruksen ja semanttisen verkon ensimmäiset versiot ovat olleet kokeiltavina syksystä 1990 alkaen. Ne ovat saatavana Macintosheille CD-ROM-versioina. Metatesauruksen koko on noin 250-300 MB ja semanttisen verkon noin 65 KB. Tietolähdehakemisto tulee koekäyttöön loppuvuodesta 1991. Tietämyslähteiden integrointia loppukäyttäjien hakujärjestelmiin, kuten Grateful Med, suunnitellaan (Lindberg & Humphreys, 1990).

Metatesauruksen ja semanttisen verkon jäsentämisessä erotetaan tyypit (types) ja ilmentymät (tokens) toisistaan ja sovelletaan kolmea eri abstrahointityyppiä: (1) ilmentymien luokittamista (classification / instantiation) tyyppeihin; (2) yleistämistä (generalization / specialization) tyyppien ryhmittämiseksi abstraktimmiksi tyypeiksi; sekä (3) aggregointia (aggregation / stepwise refinement) osien ryhmittämiseksi kokonaisuuksiksi.

UMLS tulee todennäköisesti sisältämään ainakin seuraavat toiminnalliset osat (Barr & al., 1988; UMLS, 1990): Kyselytulkki kääntää asiakkaan

kyselyn ohjelmalle sopivaan muotoon. Graafinen visualisoija havainnollistaa metatesauruksen ja semanttisen verkon termien ja termityyppien suhteita. Vuorovaikutteinen haun muotoilija avustaa hakujen muotoilua ja kääntämistä sopivaan muotoon eri tietolähteistä tapahtuvaa hakua varten. Haun suorittaja lähettää hakulausekkeet, suoritettavaksi valvontaa ja tulosten vastaanottoa varten. Tulosten jälkikäsittelijä eri lähteistä saatujen tulosten yhdistämiseen, organisointiin, arviointiin ja järjestämiseen.

Saamme laitoksellemme lähiaikoina ULMS-metatesauruksen, jossa on useita tuhansia lääketieteen käsitteitä ja niiden käännöksiä muiden tesaurusten (käsitemallien) käsitteiksi (noin 300 megatavua), sekä UMLS:n semanttisen verkon (noin 65 kilotavua) Macintosh / HyperCard -versiona.

MedIndEx-projekti: Projekti tutkii lääketieteellisten viite- ja tekstitietokantojen manuaalisen (intellektuaalisen) indeksoinnin tukemista ns. kehysesitykseen (frame representation) perustuvalla tekoälyjärjestelmällä. Projekti tutkii tietämysperusteista indeksointia (knowledge-based indexing). Sen päätarkoitus on tukea ihmisasiantuntijan indeksointia MEDLINE-tiedonhakujärjestelmää varten. Tätä varten järjestelmä sisältää käsikirjoituksen (script), joka mallintaa indeksointiprosessia ja automaattisesti valitsee kehyksiä (frame) indeksoijan täytettäväksi. Indeksoijan tehtävä on antaa esitettyjen kehysten kolojen (slot) arvoja, joiden mukaan indeksointiprosessi etenee. Käsikirjoitus valitsee automaattisesti MeSHin alaotsikot (subheadings) indeksitermeille. Tällä tavalla MeSH-alaotsikoiden unohtamisesta johtuvia ongelmia yritetään lieventää tai välttää kokonaan. MedIndEx kirjaa myös joitakin MeSHin termejä (ei alaotsikoita) automaattisesti. Esim. jos dokumentissa tarkastellaan useita erilaisia hoitomuotoja jollekin sairaudelle, voi MedIndEx automaattisesti päätellä termin COMBINED MODALITY THERAPY tarpeellisuuden.

MedIndEx:n tietämyksen esitys koskee kolmea oliotyyppiä: dokumenttityyppi, tietämystyyppi ja aikakausilehtityyppi. Dokumentti- ja aikakausilehtityypille on vain yksi geneerinen kehys. Tietämystyypille on useita, ja tietämyskanta koostuu niistä. Tietämyskehykset esittävät indeksoitavaa tietämystä lääketieteellisistä prosesseista, menettelytavoista, biologisista rakenteista ja kemiallisista yhdisteistä. Tietämyskehysten kesken on semanttinen verkko. Kuvassa 3.1 esitetään tietämyskehykset taudeista, kasvaimista ja kysta-tyyppisistä kasvaimista.

```
(disease    (is-a (value medical_subject))
            (instances (value neoplasm))
            (body-part
                  (restrictions (<Lisp-function>))
                  (if-added (<Lisp-function>)))
            (procedure
                  (restrictions (<Lisp-function>))
                  (if-added (<Lisp-function>)))
            (symptom
                  (restrictions (<Lisp-function>))
                  (if-added (<Lisp-function>))))
-------------------------------------------------
(neoplasm (is-a (value disease))
            (instances (value cyst)))
-------------------------------------------------
(cyst (is-a (value neoplasm)))
```

Kuva 3.1. Esimerkki tietämyskehyksestä

Kehys kertoo seuraavaa: kysta on kasvain (neoplasm); kasvain on tauti, jonka alalajina on kysta; tauti on lääketieteellinen seikka, ja sen alalajina on kasvain. Lisäksi taudeille voidaan määritellä ruumiinosa (body-part), toimenpide (procedure) ja oireet, joita voidaan tarkemmin kuvata annettavien rajoitteiden (restrictions) ja lisäysehtojen (if-added) avulla, joita ei kuitenkaan kuvassa tarkemmin eritellä. Rajoitteiden avulla voidaan asettaa taudin esiintymispaikalle, mahdollisille toimenpiteille tai oireille ehtoja, jotka rajaavat kyseeseen tulevat kaikkien (eri taudeissa yhteensä) esiintyvien mahdollisuuksien joukosta. Ne määritellään Lisp-ohjelmointikielen funktioina.

MedIndEx -järjestelmän uusin version on toteutettu Common Lispillä SUN-työasemaympäristössä. MedIndEx-projektia kuvataan seuraavissa julkaisuissa: (Humphrey & Miller, 1987; Humphrey, 1989a ; Humphrey, 1989b; Humphrey, 1991).

4. Department of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ , USA

Rutgers Universityn informatiikan laitos, Department of Communication, Information and Library Studies, on laajalti tunnettu. Tapasin siellä seuraavat tutkijat: Prof. Nicholas Belkin, Prof. Tefko Saracevic, Prof. Paul Kantor ja Assoc. Prof. James

Anderson. Lisäksi osallistuin Prof. Andersonin luennolle sekä laitoksen tutkimuskollokvioon.

Prof. Belkinin tutkimus suuntautuu seuraaviin peruskysymyksiin: ihmisten ongelmanratkaisukäyttäytyminen eri tilanteissa, tätä tukevien tietotukijärjestelmien luonne ja toiminnot sekä näihin tilanteisiin soveltuvat esitys- ja hakutekniikat. Parhaillaan hän pyrkii luonnehtimaan ja luokittamaan ihmisten tietoon liittyviä ongelmia (information related problems), kuvaamaan ja analysoimaan ihmisten keskinäistä informaation vaihtoa ja suunnittelemaan informaation vaihtoa ihminen-kone-työpareissa. Tätä työtä sovelletaan myös hakujärjestelmien kehitystutkimuksessa (ks. Belkin & Marchetti, 1991). Belkinin toinen mielenkiintoalue on näyttöluetteloiden suunnittelu. Hän vetää yhtä projektia tällä alaalla (ks. Belkin, N.J. et al., 1990).

Prof. Saracevic jatkaa pitkän ajan tutkimustaan tiedon hankinnasta ja online-hakujärjestelmien käytöstä. Projektissa on väitöskirjan tekijöitä ja sitä on kuvattu esitelmässä (Saracevic & al., 1991). Saracevicin toinen päätutkimusalansa on kehitysmaihin suuntautuva "Comprehensive Information System in Public Health" -projekti (Kansanterveyden kokonaisvaltainen tietojärjestelmä). Järjestelmää kokeillaan Meksikossa, Brasiliassa, Kiinassa ja Zimbabwessa. Prof. Saracevic on myös Information Processing & Management : An International Journal -lehden päätoimittaja.

Prof. Kantor on osallistunut Saracevicin tiedonhakututkimuksiin, mutta on lähinnä kiinnostunut tiedon taloudesta (economics of information). Tällä hetkellä hänellä on laaja projekti (Alexandria-project), jossa tutkitaan laboratorio-olosuhteissa kirjastojen suoritteita ja hyötyjä. Projekti on avoin ulkomaisille osanottajille. Tämä on kiinnostava yhteysmahdollisuus suomalaisille kirjastojen suoritteiden ja hyötyjen tutkijoille.

Prof. Anderson tunsi laitoksemme hakutesaurustutkimushankkeen. Hän on kehittänyt PC-mikrolla toimivan ohjelmiston (IOTA), joka tukee hakutesauruksen rakentamista luonnollisen kielen tekstien perusteella. Ohjelma on lähinnä tesauruksen rakentamista tukeva kirjanpitoväline, sillä se ei tee minkäänlaista käsiteanalyysia (ei osaa yhdistää toisiinsa liittyviä, erilaisia sanoja). Ohjelma ei tietenkään osaa palauttaa suomenkielen taipuneita sanoja perusmuotoihinsa eikä osanne käsitellä skandinaavisia aakkosiakaan, joten sen sovellettavuus Suomessa lienee kyseenalainen, vaikka se olisikin saatavilla. Käsikirjoitukset (Anderson, 1987) ja (Anderson & Rowley, 1991) kuvaavat IOTAa.

Prof. Belkin järjesti kollokvion, jossa analysoitiin yhden väitöskirjantekijän tutkimusaineistoa monen eri tieteenalan näkökulmasta. Aineisto käsitti videoidun tiedonhakutilanteen, kattaen ns. hakuhaastattelun ja varsinaisen haun, sekä haun aikana kertyneen hakujärjestelmän tapahtumalokin (annetut komennot ja saadut vastaukset). Näitä analysoivat tutkijat, jotka edustivat tiedotustutkimusta, sosiaalipsykologiaa, politologiaa ja informatiikkaa. Analyysit olivat hyvin valaisevia ko. oppiaineiden näkökulmien suhteen ja hyvin erilaisia. Tekijät myönsivätkin toistensa analyysien olevan aivan oikeassa, mutta kuitenkin totesivat niiden olevan vääriä! Vuorovaikutus oli siten hedelmällistä. Tilaisuus oli hyvin opettavainen ja vastaavan tilaisuuden järjestämistä meillä kannattaa harkita.

5. Lopuksi

Tiedonhaun tutkimusta tehdään kolmella tasolla: käsitetaso (käsitteiden välisten suhteiden hallinta), ilmentymätaso (käsitteitä edustavat kielelliset ilmaisut) ja esiintymätaso (kielellisten ilmaisujen esiintymäpaikat ja -määrät dokumenteissa). Eri tutkimusperinteet näyttävät keskittyvän eri tasoille. Esim. Saltonin perinnettä seuraava tutkimus painottaa sanojen esiintymien tilastollisia ominaisuuksia antamatta merkittävää sijaa kielellisten ilmaisujen käsittelylle tai käsitteiden suhteiden hallinnalle. Mielestäni eri tasoilla tehtävän tutkimuksen keskinäinen vuorovaikutus on liian vähäistä.

Toisaalta näyttää myös siltä, että laboratorioissa kauan tutkituilla tiedonhaun menetelmillä on pian annettavaa käytännön tiedonhakutoiminnallekin, joten nämä kaksi kauan varsin kaukana toisistaan ollutta maailmaa voivat hyötyä toisistaan. Suurten tietokantojen tutkiminen laboratorio-olosuhteissa vahvistaa tätä.

Niin käytännön toiminta kuin tutkimuskin näyttävät hyödyntävän samaa lähestymistapaa: tiedonhaun tehtävä on löytää relevantteja (tai todennäköisesti relevantteja) dokumentteja niitä tarvitseville. Uskon tämän suuntauksen saavan muita rinnalleen lähitulevaisuudessa. Uskon, että dokumenttien tarvitsijoita on paljon vähemmän kuin helposti ajatellaan – tiedon tarvitsijoita sitäkin enemmän. Tarvitaan järjestelmiä ja menetelmiä, jotka hakevat vastauksia kysymyksiin suoremmin ja jotka paremmin sopeutuvat, jopa sulautuvat, muuhun tietotyöhön.

Kalervo Järvelin

Lähdeviitteet

Anderson, J.D. (1987), Information Organization Based on Textual Analysis (IOTA): Instructional Programs for Database Design. New Brunswick, NJ : Rutgers Univ, School of Communication, Information & Library Studies, manuscript. (Saatavana Järveliniltä)

Anderson, J.D. & Rowley, F.A. (1991), Building End-User Thesauri from Full-Text. New Brunswick, NJ : Rutgers Univ, School of Communication, Information & Library Studies, manuscript. (Saatavana Järveliniltä)

Barr, Charles E. & Komorowski, Henryk Jan ; Pattison-Gordon, Edward ; Greenes, Robert A. (1988), Conceptual Modelling for the Unified Medical Lanfuage System. IN : Proceedings of the 12th Symposium on Computer Applications in Medical Care, Washington D.C., Nov. 1988. IEEE Computer Society, 1988.

Belkin, N.J. & al. (1990), Taking Account of User Tasks, Goals and Behavior for the Design of Online Public Access Catalogs. New Brunswick, NJ : Rutgers Univ, School of Communication, Information & Library Studies, SCILS Res. Rep. No. 90-14.

Belkin, N.J. & Marchetti, P.G. (1991), Interactive Online Search Formulation Support. New Brunswick, NJ : Rutgers Univ., School of Communication, Information & Library Studies, SCILS Res. Rep. No. 91-31.

Bookstein, A. & Chieramella, Y. & Salton, G. & Raghavan V.V. (Eds.) (1991), Proceedings of the 14th International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, IL, Oct. 13-16, 1991. New York, NY: The Association for Computing Machinery. (myös : Special Issue of the SIGIR Forum, October 13-16, 1991.)

Croft, W. Bruce (1991), Text Representation and Retrieval Techniques for Document Detection. Vol I : Technical. Research Proposal, Dept. of Computer and Information Science, Univ. of Massachusetts, Amherst, MA. (Saatavana K. Järveliniltä)

Croft, W. Bruce & Das, Raj (1990), Experiments with Query Acquisition and Use in Document Retrieval Systems. IN: Vidick, 1990 : pp. 349-368.

Croft, W. Bruce & Thompson (1987), R.H., I^{3}R : A New Approach to the Design of Documentt Retrieval Systems. Journal of the American Society of Information Science. 38(6) : 389-404.

Croft, W. Bruce & Turtle, Howard R ; Lewis, David D. (1991), The Use of Phrases and Structured Queries in Information Retrieval. IN : Bookstein & al., (1991), pp. 32-45

Humphrey, Susanne M. (1989a), A Knowledge-Based Expert System for Computer-Assisted Indexing. IEEE Expert, Fall 1989 : 25-38.

Humphrey, Susanne M.(1989b), MedIndEx System : Medical Indexing Expert System. Information Processing and Management 25(1) : 73-88.

Humphrey, Susanne M. (1991), Evolution Toward Knowledge-Based Indexing for Information Retrieval. IN : Proc. Workshop on Future Directions in Text Analysis, Retrieval and Understanding, Oct. 10-11, 1991, Chicago, IL. pp. 132-139.

Humphrey, Susanne M. & Miller, Nancy E. (1987), Knowledge-Based Indexing of the Medical Literature : The Indexing Aid Project, JASIS 38(3) : 184-196.

Krovetz, R. & Croft, W. Bruce (1991), Lexical Ambiquity and Information Retrieval. ACM Transactions on Information Systems, to appear.

Lindberg, D.A.B. & Humphreys, B.L. (1990), The UMLS Knowledge Sources : Tools for building better user interfaces. IN: Proc. of the 14th Annual Symposium on Computer Applications in Medical Care, Washington DC, November 4-7, 1990. Los Alamitos, CA : IEEE Computer Society : pp. 121-125.

Saracevic, T. & Mokros, H. & Su L.T. & Spink, A.(1991), Interaction between users and intermediaries in online searchining. IN : Williams, M.E., Proc. 12th National Online Meeting, May 7-9, 1991, New York. Medford, NJ: Learned Information, pp. 329-340.

Turtle, H.R. & Croft, W. Bruce (1990), Inference Networks for Document Retrieval . IN: Vidick (1990) : pp. 1-24.

Turtle, H.R. & Croft, W. Bruce (1991), Evaluation of an Inference Network-Based Retrieval Model. ACM Transactions on Information Systems (to appear).

UMLS, (1990), Unified Medical Language System. Fact Sheet. National Library of Medicine, Office of Public Information, Bethesda, MD, November 1990.

Vidick, J.L. (Ed.), (1990), Proc. of the 13th International Conference on Research and Development in Information Retrieval, Brussels, Belgium, Sept 5-7, 1990. Bruxelles : ACM.