Sanna Tuomela

Kokemuksia CoLIS 2 -konferenssista

Tuomela, Sanna, Kokemuksia CoLIS 2 -konferenssista [Experiences from the CoLIS 2 conference]. Informaatiotutkimus 16 (2): 64–68, 1997.

CoLIS 2 -conference was in Copenhagen, in October 1996. Article reviews experiences, and four presentations in the conference. The first presentation is van Rijsbergen's theory of applying probabilistic logic in information searching, followed by the theory of ostensive model of information retrieval. The ostensive model was presented by Iaian Campbell. The ostensive model is developped by van Rijsbergen and Campbell, and based on the theory of probabilistic logic in information searching. Saracevic' presentation is about relevance, and describes how the meaning of the concept has changed in recent decades. Last presentation is Stefano Mizzaro's "A Cognitive Analysis of Information Retrieval"; a formalization of the concepts information, information need and relevance.

Address: Yliopistonkatu 18/204, FIN-90570 Oulu Finland.
Email: satuomel@paju.oulu.fi.

Osallistuin Kööpenhaminassa 13.-16.10 1996 järjestettyyn CoLIS 2 -konferenssiin. Olin Nordplus -opiskelijana Kööpenhaminassa syyslukukauden, ja sain konferenssissa tiiviin paketin tietoa alan ajankohtaisesta tutkimuksesta. Konferenssin teemana oli "Integration in perspective", joka konferenssin puheenjohtajan Peter Ingwersenin mukaan tarkoittaa teknologiakeskeisen ja ihmiskeskeisen tutkimussuunnan tasapainoa ja yhdistymistä informaation tutkimukseksi. Kerron ensin lyhyesti omia mietteitäni konferenssista, minkä jälkeen esittelen neljä konferenssiesitelmää.

Konferenssiin osallistui noin 120 henkilöä, enimmäkseen Pohjois-Euroopasta ja Yhdysvalloista. Suomesta oli esitelmöijiä sekä Oulun että Tampereen yliopiston Informaatiotutkimuksen laitokselta. Konferenssin osallistujista suurin osa oli "vanhoja konkareita", mutta myös uusi, nuorempi sukupolvi on vahvasti valloittamassa tutkimusalaa. Yllätyksekseni yksikään konferenssiesitelmä ei koskenut suoranaisesti Internetiä, vaikka Internet tarjoaa täysin uudenlaisen tietoympäristön, johon entiset informaatiotutkimuksen teoriat ja käsitteet eivät usein sellaisenaan sovellu, ja sen vuoksi Internet olisi voinut tarjota aiheita uudentyyppisiin teorioihin. Yllätyin myös siitä, että seuraava vastaavanlainen konferenssi järjestetään vasta viiden vuoden päästä. Varmaan tieteenalan eri sektoreilla on lukuisia omia konferenssejaan, mutta uskoisin tällaisen kaikkien eri sektoreiden tutkimukset yhteen kokoavan konferenssin olevan tarpeellinen ainakin parin vuoden välein.

Neljä esitelmää, jotka kiinnostivat minua eniten, olivat C.J. van Rijsbergenin "Information, Logic and Uncertainty in Information Science", van Rijsbergenin ja Iaian Campbellin "The Ostensive Model of developing information needs", Tefko Saracevicin "Relevance reconsidered '96" ja Stefano Mizzaron "A cognitive analysis of information retrieval". Nämä esitelmät keskittyvät lähinnä

tiedonhaun mallintamiseen eivätkä anna täyttä kuvaa kokouksen laajasta aihepiiristä.

Relevanssin arviointiin probabilistista logiikkaa klassisen sijasta

Konferenssin pääpuhuja C.J. van Rijsbergen pyrki kuvaamaan tiedonhaun interaktiivista luonnetta käsitteillä "informaatio, logiikka ja todennäköisyys". Hänen mukaansa tiedonhakua on kuvattu yleensä staattisena, vaikka se todellisuudessa on käyttäjän ja IR-järjestelmän dynaamista vuorovaikutusta. Tiedonhakuprosessiin vaikuttaa seitsemän epävarmuutta lisäävää tekijää; tietämättömyys (esim. määriteltäessä dokumentin relevanssia), vaillinaisuus (esim. indeksoinnin tai hakukielen), päättämättömyys, kompleksisuus (vaatii pyöristyksiä ja likiarvoja jotka lisäävät epävarmuutta), satunnaisuus (dokumentti voi olla joissain tilanteissa hyödyllinen ja vastata tiedontarvetta, toisissa ei), epämääräisyys ja epätarkkuus. Näiden epävarmuustekijöiden hallitsemiseen klassisen logiikan päättelysäännöt eivät ole riittäviä, vaan tarvitaan todennäköisyyslogiikkaa.

van Rijsbergen kuvaa klassisen logiikan riittämättömyyttä ja jäykkyyttä käsitteiden "aboutness" ja "relevanssi" avulla. Jos esimerkiksi tiedonhakija arvioi dokumentin A ei-relevantiksi, ja katsoo seuraavaksi dokumenttia B arvioiden sen relevantiksi, tiedonhakija voi tämän jälkeen arvioida uudelleen dokumenttia A ja havaita sen, nähtyään dokumentin B, sittenkin relevantiksi. Tiedonhaku on siis jatkuvaa vuorovaikutusta, jossa jokainen aiemmin löydetty dokumentti vaikuttaa tiedonhakijan päätökseen tietyn dokumentin relevanssista. Dokumentit eivät ole yksiselitteisesti "about something", vaan aboutness voidaan nähdä eri valossa eri tilanteissa. Klassisen logiikan mukaan dokumentti voi olla vain joko "about x" tai "not about x", eikä kerran "not about x" -dokumentiksi määritelty voi uudessa valossa olla "about x". Todellisuudessa näin kuitenkin tapahtuu. van Rijsbergen lainaa esimerkin kvanttifysiikasta: havainnoitsija vaikuttaa aina havainnoitavaan. Niin myös tiedonhaussa, havainnoitsija ja hänen toimintansa vaikuttavat aina siihen, miten relevanssi ja "aboutness" kullakin hetkellä määritellään. Tällaista tiedonhaun prosessia on mahdoton mallintaa klassisen logiikan avulla. Klassinen logiikka pyrkii ehdottomaan varmuuteen, mutta probabilistisen logiikan avulla voidaan kuvata esimerkiksi jonkin dokumentin olevan relevantti tietyn asteisella todennäköisyydellä. Ongelmana on tämän relevanssin asteen mallintaminen ja mittaaminen.

van Rijsbergen kuvaa todennäköisyyden mittaamista logiikan Modus Ponens -päättelysäännön avulla, muuttaen kuitenkin päättelyä niin, että premissien todennäköisyyksistä voidaan päätellä johtopäätöksen todennäköisyys. Tällainen kvanttifysiikasta lainattu logiikka kuvaa tiedonhaun interaktiivista luonnetta oikeammin kuin klassinen logiikka.

Ostensiivinen tiedonhaun malli

Iain Campbell ja van Rijsbergen ovat yhdessä rakentaneet edellä esitellyn vuorovaikutuksen logiikan periaatteiden mukaan kehittyviä tiedontarpeita kuvaavan "ostensiivisen mallin". Tiedonhaku tapahtuu dynaamisessa graafisessa tietoavaruudessa dokumentteja selaamalla. Ostensiivisen mallin komponentteja ovat käyttäjän tietämyksen tila (K), informaatio (I), toiminto (A) ja informaatiolle altistuminen, eli tiedontarpeen täyttäminen jonkinasteisesti (E). Komponentit muodostavat tiedonhaun syklin; alkuperäisessä tietämyksen tilassa nousee informaation tarve, jonka tyydyttämiseen tarvitaan toiminto, jonka tarkoituksena on tyydyttää informaation tarve mahdollisimman hyvin, eli toiminnolla pyritään valitsemaan relevantein informaatio, jolle tiedontarvitsija sitten altistaa itsensä, eli tutkii informaationlähdettä ja rakentaa sen tuloksena uuden tietämyksen tilan. Prosessi tapahtuu siis järjestyksessä K-A-I-E => K'. Campbell ja van Rijsbergen esittävät syklistä seuraavia väittämiä.

Tietämyksen tila K: Informaation tarpeeseen lähimmin yhteydessä olevat (tietämyksen tilassa esiintyvät) tekijät vaikuttavat voimakkaimmin tiedonhakijan välittömiin toimintoihin. Toiminnot A: Informaation tarpeen motivoimat toiminnot suuntautuvat

todennäköisimmin kohteisiin, joiden tiedonhakija uskoo parhaiten täyttävän tarpeen. Muutos K => K': Tiedonlähteen "käytön" (E) eli yleensä lukemisen tuloksena tietämyksen tila muuttuu (K=>K'). Valtaosa muutoksista tapahtuu sillä alueella K:TE, jota tiedontarve suorimmin koskee. Komponenteista vain toimintoja ja tiedonlähteiden sisältämää informaatiota voidaan suoraan havainnoida, tietämyksen tila ja informaatiolle altistuminen ovat ei-havainnoitavia komponentteja ja niistä voidaan tehdä johtopäätöksiä vain toimintojen ja informaation perusteella. A ja I ovat indikatiivisia niitä edeltävän tietämyksen tilan suhteen, ja I on myös indikatiivinen sitä seuraavaan uuteen tietämyksen tilaan nähden.

Yksittäisen tiedonhaun syklin A ja I eivät vielä riitä kertomaan tietämyksen tilasta mitään, mutta useiden syklien ketju selventää päättelyä tietämyksen tilasta. Ostensiivisen mallin nimi tuleekin tavasta, jolla malli pyrkii havainnoimaan tiedonhakijan tiedontarpeita sen perusteella, miten tiedonhakija selaa dokumentteja, ja tekemään johtopäätöksiä ei-havaittavissa olevista komponenteista. Systeemi tarkkailee tiedonhakijan toimintoja hakijan huomaamatta. Havaintojen perusteella voidaan hahmottaa tiedonhakijan tarvetta ja arvioida, mitä dokumentteja hakija pitäisi relevantteina, ja asettaa näitä ehdolle. Epävarmuus tietämyksen tilan päättelyn suhteen vähenee sitä mukaa kun tiedonhaun syklejä tulee enemmän. Eli mitä useampia syklejä, sitä useampia aktioita ja informaatiota sisältäviä kohteita systeemi voi käyttää päättelyn perusteina. Tästä johdetaan seuraava väittämä: sitä mukaa kun selaillun dokumentin ikä (selailun suhteen) kasvaa, sen perusteella tietämyksen tilasta tehtyjen päättelyjen epävarmuus, eli toisin sanoen ostensiivinen epävarmuus kasvaa. Tiedonhakija selailee ostensiivisen mallin sovelluksessa dokumentteja liikkuen solmusta toiseen, ja systeemi luo solmuja dynaamisesti hakijan edellisten valintojen perusteella.

Campbell ja van Rijsbergen määrittelevät ostensiivisen relevanssin olevan dynaaminen ja ilmenevän vasta kun tiedonhakija löytää dokumentin, jonka hän arvioi relevantiksi, relevanssia ei siis voi päättää etukäteen. Relevanssi on tietämyksen tilassa tapahtuva arviointi, eikä sitä sellaisena voi suoraan havainnoida. Relevanssista on esitetty seuraava väittämä: ostensiivinen relevanssi kertoo, missä määrin dokumentin voidaan arvioida olevan todiste sen hetkisestä tiedon tarpeesta. Relevanssi ilmaistaan todennäköisyytenä jokaisen dokumentin kohdalla, ja se on käänteen verrannollinen dokumentin ostensiivisen epävarmuuden suhteen.

Ostensiivinen malli on varsin mielenkiintoinen mietittäessä esimerkiksi WWW:n tiedonhakua, vaikkei sellaisenaan ole mahdollinen soveltaa niin heterogeenisen ja laajan tietomäärän selailuun. Hypertekstirakenne tarjoaa kuitenkin puitteet selailuun, ja mallista voisi kehittää yksinkertaistetun version kokeiltavaksi pienessä mittakaavassa esimerkiksi WWW-ympäristössä. Ohjelma seuraisi tiedonhakijan "polkuja", etsisi aiempien dokumenttien perusteella uusia dokumentteja ehdolle ja painottaisi aiempia dokumentteja esim. sen perusteella, kuinka kauan tiedonhakija lukee dokumenttia.

Saracevic ja relevanssi -käsitteen kehittyminen

Konferenssin kutsuttuna puhujana oli Tefko Saracevic, joka arvioi informaatiotutkimuksen peruskäsitettä esitelmässä "Relevance reconsidered '96". Relevanssi on ollut informaatiotutkimuksen keskeisimpiä käsitteitä siitä lähtien, kun ensimmäisiä automaattisia tiedonhakujärjestelmiä alettiin kehittää, ja niiden päätavoitteeksi määriteltiin relevantin tiedon löytäminen. Tiedonhakusysteemien tehokkuutta mitattiin relevanssin avulla. Relevanssin käsitteen selventäminen on tärkeä osa informaatiotutkimusta, mutta usein relevanssin käsitettä käytetään selvittämättä sen merkitystä tarkemmin, koska ymmärrämme relevanssin merkityksen intuitiivisesti ja merkitys tuntuu hyvin itsestään selvältä. Asioiden ja tilanteiden arvottaminen niiden relevanssin perusteella on jatkuvaa ja dynaamista, ja muutamme jatkuvasti toimintaamme ja ajatuksiamme asioiden relevanssin mukaan.

Relevanssin attribuuteiksi voidaan määritellä: se on kognitiivinen toiminto joka vaatii

vuorovaikutusta ja kommunikointia. Se esiintyy ilmaistaessa tilanteiden ja asioiden välisiä suhteita. Intentiot; roolit, odotukset, motivaatio synnyttävät suhteita ja vaikuttavat niihin. Intentiot ilmenevät tietyssä kontekstissa, eikä relevanssia voi ajatella irrallaan tästä kontekstista. Suhteiden arvioiminen edellyttää johtopäätöksiä niiden vaikuttavuudesta tietyssä kontekstissa. Relevanssin päätteleminen on dynaaminen ja interaktiivinen prosessi, ja edellä mainittujen attribuuttien tulkinta muuttuu sitä mukaa kun subjektin havainnot muuttuvat. Saracevic esittelee viisi relevanssin luonnetta kuvaavaa kehystä: systeemi- ja kommunikaatiokehyksen, tilannekohtaisen, psykologisen sekä interaktiivisen kehyksen.

Systeemi- tai järjestelmäkeskeinen kehys muodostui automaattisten tiedonhakujärjestelmien yhteydessä. Tavoitteena oli muuttaa (verbaalinen) tiedontarve järjestelmän hyväksymään muotoon, ja täsmäyttää tämä hakulauseke ja tietokannan sisältämät tekstit. Relevanssi määriteltiin täsmäyttämisen onnistumisen perusteella. Tämän kehyksen mukaan relevanssi on tiedonhakujärjestelmän ominaisuus. Kehys on hyvin yksipuolinen, eikä ota huomioon tiedonhakijaa.

Kommunikaatiokehys taas perustuu Shannonin informaatioteoriaan, jonka mukaan kommunikaatio on lähteen ja kohteen välistä viestien vaihtoa, jossa esiintyy häiriöitä ja mahdollisesti palautetta. Relevanssi on kommunikaation tehokkuuden kriteeri, ja määrittää siis viestin lähettäjän ja kohteen välistä suhdetta. Viestin lähettäjä ja kohde sisältävät useita elementtejä, joiden välille syntyy siis useita suhteita. Relevanssi voidaan tulkita näistä miksi tahansa.

Tilannekohtaisessa kehyksessä relevanssi määritellään eri tilanteen, ajan ja sosiaalisen kontekstin mukaan. Relevanssin luonne on dynaaminen ja näkökulma on hyvin käyttäjäkeskeinen. Psykologisessa kehyksessä keskitytään tiedontarvitsijan kognitiivisiin prosesseihin ja tilaan. Kognitiivisen relevanssin keskeinen ongelma on tiedontarpeen muotoileminen suulliseksi kuvaukseksi. Sekä tilannekohtaisessa että psykologisessa kehyksessä relevanssia tutkitaan tiedontarvitsijaa ja hänen tilannettaan analysoimalla. Vastakohtana on systeemikeskeinen kehys, joka unohtaa tiedontarvitsijan vaikutuksen relevanssiin. Kumpikaan ääripää ei ole hedelmällinen ja vastaukseksi siihen Saracevic ehdottaa interaktiivista kehystä.

Interaktiivinen kehys pyrkii tarkastelemaan sekä automaattista tiedonhakusysteemiä että sen käyttäjää. Kehyksessä on useita kerroksia; käyttäjä toimii kognitiivisella, tilannekohtaisella ja tunnetasolla. Tiekoneella on tekninen, sisältö- ja käsittelytaso. Kahden osapuolen tasot ovat vuorovaikutuksessa pintatason (käyttöliittymä) kautta. Kaikilla tasoilla ilmenee oletuksia ja johtopäätöksiä relevanssista. Tuloksena on dynaaminen relevanssien järjestelmä. Tiedonhaun evaluointi on eri tasojen relevanssien vertailua, eikä yhden tason relevanssia tule pitää koko tiedonhaun interaktion relevanssina. Esitelmä kuvaa yhden tavan rakentaa siltaa teknologiakeskeisen ja käyttäjäkeskeisen tutkimuksen välille, konferenssin teeman mukaisesti.

Informaation, tiedontarpeen ja relevanssin formaali määrittely

Konferenssipäivällisillä "Young Scientist of the Year" -palkinnon saanut Stefano Mizzaro pyrki kognitiivisen tiedonhaun mallin pohjalta määrittelemään formaalisti käsitteet informaatio, informaation tai tiedon tarve sekä relevanssi. Kognitiivisen mallin peruskäsitteitä ovat kognitiivinen agentti, agentin tietämyksen tila (knowledge state, KS), ulkoinen havaittu maailma, josta agentti muodostaa representaatioita tietämyksen tilaansa (representaatio ei välttämättä täysin vastaa ulkoista maailmaa) sekä oliot (knowledge items, KI), jotka muodostavat tietämyksen tilan. Agentin tietämys voi jakaantua useisiin alaryhmiin (SubKS), jotka muodostuvat keskenään vahvasti kytkeytyneistä olioista. Eri alaryhmien ja niiden olioiden välillä voi olla eri vahvuisia yhteyksiä. Myös näiden yhteyksien välillä voi esiintyä yhteyksiä. Tietämys on siis monin tavoin verkottunut, ja eri alaryhmiä on mahdoton täysin erottaa toisistaan. Agentin tietämyksen tila muuttuu joko sisäisen päättelyn johdosta tai ulkoisen havainnoinnin kautta saadun informaation ansiosta. Tietämyksen tilaan joko

tulee jotain jota se ei ole aiemmin sisältänyt (K+), tai siitä poistetaan jotain (K-). Nämä muutokset muuttavat koko tietämyksen tilaa linkkien kautta, tietämys on jokaisen muutoksen jälkeen rakennettava uudelleen.

Mizzaro määrittää informaation olevan aikajärjestyksessä esitetty pari, joka kuvaa eron kahden tietämyksen tilan välillä. Hän ei siis määrittele informaation olevan jokin, joka aiheuttaa eron, vaan informaatio on nimenomaan ero. Data kantaa informaatiota, mutta se voidaan mitata vasta kun havainnoitsijan tietämyksen tilassa tapahtuu muutos. Informaatio on siis subjektiivista ja kontekstisidonnaista. Sama data voi kantaa erilaista informaatiota eri havainnoitsijoille, ja eri datat voivat kantaa samaa informaatiota. Kuitenkin keskimäärin sama data kantaa samaa informaatiota eri agenteille. Se, millaista informaatiota ja minkä verran data kantaa, riippuu havainnoitsijan tietämyksen tilan siitä osasta, johon data liittyy ja mahdollisesti vaikuttaa, ei siis koko tietämyksen tilasta. Tiedon tarve nousee, kun agentin tietämyksen tila ei ole riittävä, jotta agentti voisi ratkaista ongelman tai saavuttaa jonkin tavoitteen. Tavoitteena oleva tietämyksen tila voidaan saavuttaa eri tavoin ja erilaisten välivaiheiden kautta.

Tiedon tarve määritellään pienimmäksi mahdolliseksi tavoiteltuun tietämyksen tilaan tarvittavien informaatio-olioiden joukoksi. Informaation tarve esiintyy kahdenlaisena: todellinen tiedon tarve ja agentin itsensä havainnoima ja tiedostama tiedon tarve. Usein nämä eroavat toisistaan. Agentin toiminta hänen pyrkiessään tyydyttämään informaation tarpeensa voidaan jakaa kolmeen vaiheeseen: informaation etsiminen, informaation vastaanottaminen ja informaation käyttö. Nämä kolme vaihetta yhdessä muodostavat tiedonhaun (IR) tutkimuskohteen, ei siis pelkästään informaation etsimisvaihe. Lopuksi Mizzaro määrittelee informaatio-olion (tai -yksikön) relevantiksi, jos sen ja tiedontarpeen välillä on leikkauskohta (tiedontarve / \ informaatio-olio). Informaatio-olio on siis relevantti vain, jos se auttaa saavuttamaan tavoitellun tietämyksen tason. Mizzaron tavoitteena on selkeyttää tiedonhaun tutkimusta määrittelemällä kolme peruskäsitettä formaalisti. Tämä väistämättä yksinkertaistaa tiedonhaun prosessin kompleksisuutta, mutta auttaa ehkä jatkossa tekemään tutkimusta selkeästi määriteltyjen käsitteiden pohjalta.

Lopuksi

Konferenssin tiukan aikataulun vuoksi keskustelut esitelmistä jäivät lyhyiksi, ja varsinainen paneelikeskustelu "Maps, Paradigms and Browsers: Representing the Knowledge Base of LIS" ei ehtinyt edetä varsinaiseksi keskusteluksi lainkaan, alustukset veivät paneelikeskusteluun varatun ajan. Kaiken kaikkiaan useissa konferenssiesitelmissä heijastui pyrkimys yhdistää aiemmin vastakkaisina pidettyjä vaihtoehtoja tai löytää "kultainen keskitie" niiden välillä. Tiedonhaun tutkimuksessa tiedonhakijan ja tietojärjestelmän interaktio oli keskeisellä sijalla.

Hyväksytty julkaistavaksi 21.5.1997