Mirja Iivonen

Johdonmukaisuuden laskeminen tiedon tallennuksen ja haun tutkimuksessa

Iivonen, Mirja, Johdonmukaisuuden laskeminen tiedon tallennuksen ja haun tutkimuksessa [Calculation of consistency in the domain of information storage and retrieval]. Kirjastotiede ja informatiikka 12 (2): 63–76, 1993.

Calculation of consistency in the domain of information storage and retrieval is considered. The results of previous consistency studies are reviewed briefly. The formulas used in calculating consistency are analyzed. Examples are given of numeric results obtained when using different methods. The phenomena of inter-actor and intra-actor consistency are described. The differences between consistency figures calculated on the basis of terms and on the basis of concepts are discussed.

Address: University of Tampere, Department of Information Studies, P.O. Box 607, SF-33101 Tampere, Finland.

1. Johdanto

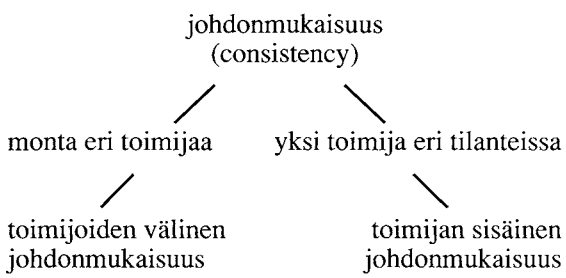
Kirjastotieteessä ja informatiikassa on tarkasteltu tiedon tallennuksessa ja haussa ilmenevää johdonmukaisuutta esittämällä jopa tarkkoja johdonmukaisuusprosentteja. Näiden lukujen ymmärtäminen ja tulkinta (mitä luvut tosiasiassa kertovat) edellyttää kuitenkin, että tiedetään, miten ne on tuotettu ja mitä ongelmia niiden laskemisessa esiintyy.

Tässä artikkelissa tarkastellaan johdonmukaisuuden laskemisessa käytettyjä kaavoja ja annetaan esimerkkejä johdonmukaisuuslukujen laskemisesta. Lisäksi luodaan lyhyt katsaus johdonmukaisuustutkimuksissa saatuihin tuloksiin.

2. Johdonmukaisuuden määrittely

Johdonmukaisuudella tarkoitetaan sitä, missä määrin samalla tavalla samaa tehtävää eri tilanteissa suoritettaessa toimitaan. Johdonmukaisuutta voidaan tarkastella joko niin, että verrataan keskenään usean eri toimijan saman tehtävän suorittamista (toimijoiden välinen johdonmukaisuus) tai niin, että tarkastellaan yhden ja saman toimijan saman tehtävän suorittamista eri aikoina (toimijan sisäinen johdonmukaisuus, ks. kuvio 1).

Johdonmukaisuudessa on siis kyse toiminnan samankaltaisuuden, ei laadun tarkastelusta. Kaksi tai useampi henkilö voi toimia keskenään hyvin johdonmukaisesti mutta kelvottomasti. Ajatellaan-



Kuvio 1. Johdonmukaisuuden kaksi eri puolta

pa kahta pilkkijää, jotka lähtevät kalastamaan samoille heikoille jäille. He saapuvat samalle järvelle ja varoituksista huolimatta lähtevät liikkeelle samaan aikaan samasta poukamasta pudoten molemmat heikkoihin jäihin. He toimivat keskenään hyvinkin johdonmukaisesti tehden samoja ratkaisuja, mutta järkeväksi tai laadukkaaksi heidän toimintaansa tuskin voi sanoa. Vastaavasti yksi ja sama henkilö voi toimia eri tilanteissa hyvinkin johdonmukaisesti, mutta harkitsemattomasti. Siitä esimerkkinä voidaan ajatella autoilijaa, joka aina ajaessaan 80 "lätkällä" merkityn auton perässä pyrkii ohittamaan tämän myös riskitilanteissa. Hän toimii suhteessa itseensä hyvin johdonmukaisesti, mutta joka kerta riskitilanteessa yhtä typerästi.

Tiedon tallennuksen ja haun tutkimuksessa johdonmukaisuudella tarkoitetaan informaation prosessoinnin samanlaisuutta kahdessa tai useammassa eri tilanteessa. Koska informaation prosessoinnissa voidaan erottaa toisistaan ilmaisun taso ja käsitteellisen jäsentämisen taso, tiedon tallennuksen ja haun alueella voidaan puhua toisaalta termijohdonmukaisuudesta, toisaalta käsitejohdonmukaisuudesta. Termijohdonmukaisuudella tarkoitetaan sitä, miten yhdenmukaisesti kohdetta (esim. dokumentin tai hakupyynnön sisältö) kuvaavat termit[1] valitaan, käsitejohdonmukaisuudella sitä, miten yhdenmukaisesti kohteena olevasta informaatiosta esiin nostettavat käsitteet valitaan.

Englanninkielisessä kirjallisuudessa käytetään *johdonmukaisuudesta* tavallisimmin termiä *consistency*. Saracevic käytti tiedonhaun johdonmukaisuudesta aluksi termiä *degree of agreement*, myöhemmin termiä *overlap* (ks. Saracevic 1984, 227-229, Saracevic et al. 1987, 25, Saracevic et al. 1988, 169).

3. Johdonmukaisuustutkimusten tuloksia

Kirjastotieteessä ja informatiikassa on tutkittu erityisesti indeksoinnin johdonmukaisuutta. Tiedonhaun johdonmukaisuutta käsittelevää tutkimusta on huomattavasti vähemmän, ja lisäksi olemassa olevat tutkimukset nojaavat melko pieneen aineistoon. Tutkimusten keskeisenä havaintona on ollut inhimillisten toimijoiden ratkaisujen vaihtelevuus tiedon tallennuksen ja haun prosesseissa.

3.1. Indeksoinnin johdonmukaisuus

Indeksoinnin johdonmukaisuutta on useimmiten tarkasteltu useamman toimijan välisenä termijohdonmukaisuutena. Tutkimuksissa on tarkasteltu toisaalta erilaisen kokemuksen omaavien indeksoijien välistä indeksoinnin johdonmukaisuutta, toisaalta erilaisten apuvälineiden käytön vaikutusta johdonmukaisuuteen. Tutkimusten tavoitteena on ollut paitsi johdonmukaisuutta koskevan tiedon syventäminen, myös hyvin praktisesti antaa välineitä luetteloiden laadun parantamiselle. Vaikka indeksoinnin johdonmukaisuus ja laatu ovat selvästi kaksi eri asiaa, johdonmukaisen indeksoinnin avulla on uskottu tuotettavan laadukkaita luetteloja (ks. esim. Chan 1989, 349).

Leonard (1977, 1-51) ja Markey (1984, 156-161) esittelevät lukuisia (Leonard 34 kpl ja Markey 25 kpl) indeksoinnin johdonmukaisuuteen liittyviä tutkimuksia, joista useimmat ovat opinnäytetöitä. Koska eri tutkimuksissa on käytetty erilaisia johdonmukaisuuden laskemistapoja, niiden tuloksia ei voida suoraan verrata keskenään. Tulosten pääsuunnat ovat kuitenkin nähtävissä. Indeksoijien välinen termijohdonmukaisuus on hyvin vaihteleva. Markeyn katsauksen tutkimuksissa sen keskiarvo vaihteli 4 %:sta 82 %:iin, Leonardin katsauksessa 12,6 %:sta 65 %:iin. Useimmissa tutkimuksissa hakijoiden välisen termijohdonmukaisuuden keskiarvo oli kuitenkin melko alhainen, noin 30 % - 40 %. Cleverdon (1984, 38) toteaakin, että mikäli kaksi kokenutta indeksoijaa indeksoi saman dokumentin käyttäen samaa tesaurusta, ainoastaan 30 % heidän käyttämistään termeistä on samoja.

Indeksoijien väliseen termijohdonmukaisuuteen vaikuttaviksi tekijöiksi on eri tutkimuksissa havaittu 1) indeksoinnissa käytettyjen termien määrä, 2) kontrolloidun sanaston käyttö indeksoinnissa, 3) käytetyn kontrolloidun sanaston yksinkertaisuus, 4) indeksoitavan aihealueen perifeerisyys indeksointiin käytetyssä sanastossa, 5) indeksoitavan dokumentin lyhyys, 6) indeksoitava aihealue ja sen sanasto, 7) dokumentin keskeisten aiheiden indeksointi ja 8) indeksoijien indeksointikokemus (Lancaster 1968, Leonard 1977, Funk & Reid & McGoogan 1983, Markey 1984, Iivonen 1989, Lancaster 1991, 62-68).

Indeksoijien välinen käsitejohdonmukaisuus on huomattavasti korkeampi kuin indeksoijien välinen termijohdonmukaisuus (Iivonen 1989, 68-77).

Indeksoijat ovat indeksoitavien käsitteiden valinnassaan keskenään johdonmukaisempia kuin näiden käsitteiden ilmaisussaan. Myös indeksoijien sisäinen johdonmukaisuus, sekä termi- että käsitejohdonmukaisuus, on havaittu selvästi korkeammaksi kuin indeksoijien välinen johdonmukaisuus (Iivonen 1989, 155-167). Vaikka indeksoijat toimivat keskenään epäjohdonmukaisesti, he toimivat sittenkin melko johdonmukaisesti suhteessa itseensä tehden samoja ratkaisuja, valiten myös samoja termejä eri tilanteissa.

3.2. Tiedonhaun johdonmukaisuus

Tiedonhaun tutkimuksessa johdonmukaisuutta on tarkasteltu sekä hakuun käytettyjen termien ja hakukäsitteiden johdonmukaisuutena että löydettyjen hakujoukkojen päällekkäisyytenä.

Hakijoiden välinen termijohdonmukaisuus on tiedonhaun tutkimuksissa vaihdellut 27 %:sta 64 %:iin (Saracevic 1984, 227-230, Saracevic et al. 1987, 182, Saracevic & Kantor 1988, 211-212). Fidel (1985, 69-72) totesi hakupyyntöjen vaikeuden (vaikeus määritelty "sormituntumalta") vaikuttavan hakijoiden väliseen termijohdonmukaisuuteen siten, että helpot hakupyynnöt kuvaillaan johdonmukaisemmin kuin vaikeat. Fidel (1987, 60-61) vertasi myös hakijoiden välistä ja hakijoiden sisäistä termijohdonmukaisuutta. Hän totesi hakijoiden sisäisen johdonmukaisuuden olevan etenkin vaikean hakupyynnön osalta selvästi korkeampi kuin hakijoiden välinen johdonmukaisuus.

Hakijoiden välistä käsitejohdonmukaisuutta on tarkasteltu ainoastaan hyvin pienellä aineistolla. Saracevicin (1984, 227-230) tutkimuksessa 16 tiedonhakuun perehtynyttä opiskelijaa suoritti haun samasta hakupyynnöstä. Tässä tutkimuksessa hakijoiden välinen käsitejohdonmukaisuus osoittautui suuremmaksi kuin termijohdonmukaisuus.

Samasta hakupyynnöstä haettujen useampien hakujoukkojen päällekkäisyys vaihteli Katzerin et al. (1982, 261-274) tutkimuksessa 5,3 %:sta 27,9 %:iin ja Fidelin (1985, 69-72) tutkimuksessa 8 %:sta 70 %:iin. Saracevicin et al. (1987, 182-183, Saracevic & Kantor 1988, 211-212) tutkimuksessa hakujoukkojen keskimääräinen päällekkäisyys oli 17 %. Fidel havaitsi hakupyynnön vaikeuden vaikuttavan myös löydettyjen hakujoukkojen päällekkäisyyteen. Saracevic et al. puolestaan totesivat, että hakutermien epäjohdonmukaisuus ei selitä hakujoukkojen epäjohdonmukaisuutta.

4. Johdonmukaisuuden laskemisesta

Johdonmukaisuutta voidaan tarkastella kvantitatiivisesti. Se edellyttää kuitenkin käyttökelpoista johdonmukaisuuden laskemistapaa. Indeksoinnin johdonmukaisuuden laskemisessa on käytetty erilaisia kaavoja. Osa niistä on kuitenkin melko ongelmallisia. Johdonmukaisuuden laskeminen edellyttää lisäksi päätöksiä siitä, missä tapauksissa eri yksiköt katsotaan samaksi. Tämä ongelma tulee eteen käsitejohdonmukaisuuden laskemisessa. Termijohdonmukaisuuden laskeminen on melko ongelmatonta, koska termejä voidaan silloin verrata toisiinsa merkki merkiltä. Tällöin myös saman termin yksikkö- ja monikkomuodot lasketaan eri termeiksi. Samoin samaan käsitteeseen viittaavat synonyymit lasketaan eri termeiksi. Termien vertaaminen toisiinsa kirjain kirjaimelta saattaa vaikuttaa liian tiukalta menettelytavalta. Sitä voidaan kuitenkin perustella sillä, että tiedon tallennuksessa ja haussa useissa tiedonhakujärjestelmissä yhdenkin merkin erolla on merkitystä, esim. termin katkaisu eri kohdista tuottaa haettaessa eri tuloksia.

4.1. Johdonmukaisuuden laskemiseen käytetyt kaavat

Johdonmukaisuuden laskemisessa on käytetty sekä symmetriseen että epäsymmetriseen laskemiseen perustuvia kaavoja. Symmetrisen laskemisen tuloksena toimijoiden johdonmukaisuus toisiinsa kuvataan yhtenä lukuna. Epäsymmetrisessä laskemisessa kahden toimijan toiminnan samankaltaisuutta verrataan erikseen toimijan 1 ja erikseen toimijan 2 toimintaan. Tällöin kumpikin toimija saa myös oman johdonmukaisuuslukunsa. Johdonmukaisuuden symmetrisen ja epäsymmetrisen laskemisen erot tulevat näkyviin, kun tarkastellaan johdonmukaisuuden laskemiseen käytettyjä kaavoja myös laskuesimerkkien avulla.

Esimerkeissä käytetään aluksi toimijoina indeksoijia, koska kaavat on otettu käyttöön juuri indeksoinnin johdonmukaisuuden laskemisessa. Kaikissa tiedonhaun johdonmukaisuustutkimuksissa (Katzer et al. 1982, Saracevic 1984, Fidel 1985, Fidel 1987, Saracevic et al. 1987, Saracevic et al. 1988) on käytetty johdonmukaisuuden laskemisen epäsymmetristä kaavaa (luku 4.1.3.). Indeksoijien tilalle voitaisiin kuitenkin kaikissa

esimerkeissä sijoittaa myös hakija. Myös hakijoiden välinen johdonmukaisuus voidaan laskea Rodgersin ja Hooperin (luku 4.1.1.) tai Rollingin (luku 4.1.2.) kaavoilla, jos näin jostakin syystä halutaan tehdä.

4.1.1. Rodgersin ja Hooperin kaava

Useissa indeksoinnin johdonmukaisuustutkimuksissa (ks. esim. Lancaster 1968, Leonard 1977, Funk, Reid & McGoogan 1983) indeksoijien välinen termijohdonmukaisuus on laskettu Rodgersin ja Hooperin 1960-luvulla esittelemällä kaavalla (1)[2]. Siinä indeksoijien välinen termijohdonmukaisuusprosentti (JP) lasketaan seuraavasti:

$$(1)\ JP = 100 \cdot \frac{a}{a+m+n}$$

Tässä kaavassa a tarkoittaa niiden termien lukumäärää, joita molemmat indeksoijat ovat käyttäneet, m niiden termien lukumäärää, jota indeksoija M on käyttänyt, mutta indeksoija N ei, ja n niiden termien lukumäärä, joita indeksoija N on käyttänyt, mutta indeksoija M ei.

Rodgersin ja Hooperin kaavan mukaan indeksoijien välinen johdonmukaisuus lasketaan symmetrisesti ja se saa saman arvon suhteessa molempiin indeksoijiin. Saracevicin (1984, 227-228) tapaan em. kaavaa voidaankin nimittää myös symmetriseksi johdonmukaisuuden laskemisen kaavaksi, ja se on esitettävissä myös seuraavassa muodossa (2):

$$(2)\ CT_{1\leftrightarrow 2} = 100 \cdot \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

Tässä kaavassa T_1 tarkoittaa indeksoijan 1 käyttämiä termejä, T_2 indeksoijan 2 käyttämiä termejä ja $CT_{1\leftrightarrow 2}$ indeksoijien 1 ja 2 välistä johdonmukaisuusprosenttia. Symmetrisen johdonmukaisuuden laskemisen etuna on se, että näin saadaan näkyviin yhtenä lukuna kahden toimijan välinen johdonmukaisuus. Toisaalta juuri tästä syystä kaava ei paljasta erikseen kummankin toimijan johdonmukaisuutta suhteessa toiseen. Tämä saadaan näkyviin johdonmukaisuuden laskemisen epäsymmetrisellä kaavalla (ks.luku 4.1.3.).

Symmetristä kaavaa käytettäessä useamman indeksoijan välinen johdonmukaisuus lasketaan pareittain ja tämän jälkeen keskiarvo kaikista pa-

Taulukko 1. Esimerkki neljän indeksoijan välisen termijohdonmukaisuuden laskemisesta Rodgersin ja Hooperin kaavalla

Indeksoitava teos: Kauppinen, Timo, Kohtanen Jukka, Työtaistelut ja neuvottelusuhteet Enso-Gutzeit Oy:n Summan tehtailla vuosina 1971-1984. Helsinki: Työelämän suhteiden neuvottelukunta, 1987.

Indeksoijien käyttämät termit			
indeksoija 1	indeksoija 2	indeksoija 3	indeksoija 4
lakot	Suomi	työtaistelut	työriidat
Suomi	työtaistelut	paperiteollisuus	ammattiyhdistys-
puunjalostus-	paperiteollisuus	yritysdemokratia	liike
teollisuus	lakot		
	yritysdemokratia		

A. Indeksoijien 1 ja 2 välinen johdonmukaisuus = 100 · 2 / (2+1+3) = 33,3 %
B. Indeksoijien 1 ja 3 välinen johdonmukaisuus =100 · 0 / (0+3+3) = 0%
C. Indeksoijien 1 ja 4 välinen johdonmukaisuus = 100 · 0 / (0+3+2) = 0%
D. Indeksoijien 2 ja 3 välinen johdonmukaisuus = 100 · 3 / (3+2+0) = 60%
E. Indeksoijien 2 ja 4 välinen johdonmukaisuus = 100 · 0 / (0+5+2) = 0%
F. Indeksoijien 3 ja 4 välinen johdonmukaisuus = 100 · 0 / (0+3+2) = 0%
G. Indeksoijien 1, 2 ja 3 keskinäinen johdonmukaisuus = (33,3+0+60)/3=31,1%
H. Kaikkien indeksoijien keskinäinen johdonmukaisuus =(33,3+0+60+0+0+0)/6=15,6 %

reittain saaduista johdonmukaisuuksista. Esimerkki useamman indeksoijan välisestä symmetrisestä johdonmukaisuuden laskemisesta esitetään taulukossa 1. Siitä huomataan Rodgersin ja Hooperin kaavan kaksi heikkoutta. Ensinnäkin kaava edellyttää korkean termijohdonmukaisuuden saavuttamiseksi huomattavan suurta yhteisten termien määrää. Jo tapauksessa A, jossa molemmat indeksoijat käyttävät kahta yhteistä termiä, mutta joissa tämän lisäksi toinen indeksoija käyttää yhtä ja toinen kolmea "lisätermiä", johdonmukaisuusprosentti jää 33,3 %:iin. Myös tapauksessa D, jossa toisen indeksoijan kaikki termit sisältyvät myös toisen indeksoijan termien joukkoon, johdonmukaisuusprosentti on vain 60.

Rodgersin ja Hooperin kaavan toinen heikkous on siinä, että laskettaessa useamman kuin kahden toimijan välistä johdonmukaisuutta, on yhdellä "poikkeavasti" toimivalla aktorilla erittäin suuri vaikutus kaikkien aktorien keskinäiseen johdonmukaisuusprosenttiin. Taulukon 1 esimerkissä indeksoijalla 4 ei ole yhtään yhteistä termiä muiden indeksoijien kanssa. Hänen ja muiden indeksoijien välinen johdonmukaisuus jää siis nollaan. Koska hänen toimintaansa verrataan pareittain kolmen muun indeksoijan toimintaan, keskiarvoa laskettaessa hän tuottaa tämän nollan moneen kertaan. Kun kolmen ensimmäisen indeksoijan keskinäinen johdonmukaisuus on vielä 31,1 % (kohta G), laskee kaikkien indeksoijien keskinäinen johdonmukaisuus indeksoijan 4 mukana ollessa 15,6 %:iin.

4.1.2. Rollingin kaava

Rolling (1981, 70) esitti useamman indeksoijan välisen johdonmukaisuuden laskemiseksi kahta eri kaavaa. Ensimmäistä niistä voidaan kutsua yksinkertaisen johdonmukaisuuden laskemisen kaavaksi (3). Siinä kaikkien indeksoijien välinen johdonmukaisuusprosentti (JP) saadaan vertaamalla kaikkien indeksoijien käyttämien yhteisten termien määrää kaikkien indeksoijien yhteensä käyttämien termien määrään seuraavasti:

$$(3)\quad JP = 100 \cdot \frac{n \cdot c}{a + b + d + \ldots}$$

$$\smile\ \_\ \_\ n\ \_\ \_\ \smile$$

Tässä kaavassa n tarkoittaa indeksoijien määrää, c kaikkien indeksoijien käyttämien yhteisten termien määrää, a indeksoijan A käyttämien termien määrä, b indeksoijan B käyttämien termien määrä d indeksoijan D:n käyttämien termien määrää. Rollingin yksinkertaisen johdonmukaisuuden laskemisen kaava voidaan esittää myös toisenlaisessa muodossa (4) seuraavasti:

$$(4)\quad JP = 100 \cdot \frac{n \cdot c}{\sum_{i=1}^{n} a_i}$$

Tässä kaavassa JP tarkoittaa n:n indeksoijan välistä johdonmukaisuusprosenttia, n indeksoijien määrää, c kaikkien indeksoijien käyttämien yhteisten termien määrää ja a yhden indeksoijan käyttämien termien määrää.

Rollingin yksinkertaisen johdonmukaisuuden laskemisen kaavassa ylikorostuu kaikkien indeksoijien yhteisesti käyttämien termien rooli. Mikäli ei löydy yhtään termiä, jonka kaikki indeksoijat ovat valinneet edustamaan dokumenttia, indeksoijien välinen termijohdonmukaisuus saa arvon nolla. Näin kävisi taulukossa 1 esitetyssä esimerkissä. Mitä useamman indeksoijan johdonmukaisuutta tarkastellaan, sitä todennäköisempää on, että joku indeksoijista toimii "poikkeavasti" eikä käytä samoja termejä kuin muut. Vaikka kaikki muut olisivat tällöin indeksoineet dokumentit täysin samoilla termeillä, kaikkien indeksoijien keskinäinen johdonmukaisuusprosentti jää kuitenkin nollaksi.

Oletusta siitä, että täysin yhteisiä termejä on vaikea löytää silloin, kun on monia indeksoijia, tukee se havainto, joka tehtiin Aslibin toimesta vuonna 1981 toteutetussa "How do we index" -projektissa. Projektissa 16 vapaaehtoista indeksoijaa indeksoi New Scientist -lehden aineistoa vapailla termeillä. Ainoat indeksoijille annetut ohjeet olivat, että näiden tuli suosia jälkikytkentää ja että termejä ei tulisi yhtä artikkelia kohti käyttää kymmentä enempää. Tarkasteltaessa koeindeksoinnissa useimmin käytettyjä termejä havaittiin, ettei yksikään termi ollut sellainen, jota kaikki indeksoijat olisivat käyttäneet. Lisäksi ainoastaan yksi termeistä oli sellainen, jota oli käyttänyt 15 indeksoijaa. (Jones 1983, 11-19.) Samansuuntainen tulos saatiin suomalaisessa tutkimuksessa, jossa kymmenen indeksoijaa indeksoi kymmenen teosta käyttäen yhteensä 167 erilaista termiä. Ainoastaan kolme

näistä termeistä oli sellaisia, joita jokainen indeksoija käytti. (Iivonen 1989, liite 7.)

Yksinkertaisen johdonmukaisuuden laskemisen kaavaa käyttäen on siis mahdollista, että jaettava saa useassa tapauksessa arvon nolla, ja mahdolliset, suhteellisen selvätkään erot indeksointitulosten johdonmukaisuudessa eivät tule esiin. Esimerkiksi jos yhdeksän indeksoijaa kymmenestä indeksoi teoksen A täysin samalla tavalla, mutta teoksen B siten, ettei löydy yhtään sellaista termiä, jota edes kaksi indeksoijaa kymmenestä käyttäisi, on teosten A ja B indeksoinnin johdonmukaisuudessa selvä ero, mutta em. kaavalla laskettaessa johdonmukaisuusprosentti jää molemmissa tapauksissa nollaksi. Kaava, joka kätkee näinkin selvät johdonmukaisuuserot, on melko hyödytön.

Niinpä Rolling (1981, 75) esittikin myös toisen kaavan, jolla voidaan laskea indeksoinnin painotettu johdonmukaisuus. Sen avulla on tarkoitus ottaa huomioon useamman henkilön välistä johdonmukaisuutta laskettaessa myös se indeksoinnin johdonmukaisuus, joka vallitsee vain joidenkin indeksoijien (vähintään kahden) kesken. Tätä kaavaa voidaan kutsua painotetun johdonmukaisuuden laskemisen kaavaksi. Rolling esittää siitä esimerkkinä neljän eri indeksoijan suorittaman indeksoinnin johdonmukaisuuden (JP) laskemisen kaavan (5)

$$(5)\ JP = 100 \cdot \frac{4c + 3/4(c_{abe} + c_{abd} + c_{aed} + c_{bed}) + 2/4(c_{ab} + c_{ae} + c_{ad} + c_{be} + c_{bd} + c_{ed})}{a + b + d + e}$$

jossa

a = indeksoijan A käyttämien termien määrä
b = indeksoijan B käyttämien termien määrä
d = indeksoijan D käyttämien termien määrä
e = indeksoijan E käyttämien termien määrä
c = kaikkien neljän indeksoijan käyttämien yhteisten termien määrä
c_{abe} = indeksoijien A, B ja E (mutta ei D) käyttämien yhteisten termien määrä
c_{abd} = indeksoijien A, B ja D (mutta ei E) käyttämien yhteisten termien määrää
c_{aed} = indeksoijien A, E ja D (mutta ei B) käyttämien yhteisten termien määrä
c_{bed} = indeksoijien B, E ja D (mutta ei A) käyttämien yhteisten termien määrä
c_{ab} = indeksoijien A ja B (mutta ei D ja E) käyttämien yhteisten termien määrä
c_{ae} = indeksoijien A ja E (mutta ei B ja D) käyttämien yhteisten termien määrä
c_{ad} = indeksoijien A ja D (mutta ei B ja E) käyttämien yhteisten termien määrä
c_{be} = indeksoijien B ja E (mutta ei A ja D) käyttämien yhteisten termien määrä
c_{bd} = indeksoijien B ja D (mutta ei A ja E) käyttämien yhteisten termien määrä
c_{ed} = indeksoijien E ja D (mutta ei A ja B) käyttämien yhteisten termien määrä

Rollingin esittämä kaava on luettavissa ja esitettävissä vielä neljän indeksoijan välisen johdonmukaisuuden kaavana. Mutta mitä useampi toimija tulee ottaa huomioon, sitä hankalammaksi kaavan esittäminen Rollingin mallin mukaisesti käy. Se voidaan kuitenkin muuntaa myös n:n indeksoijan välisen johdonmukaisuuden laskemisen kaavaksi (6) ja esittää taloudellisesti, jolloin n:n indeksoijan välinen johdonmukaisuusprosentti (JP) lasketaan seuraavasti:

$$(6)\ JP = [100 / \sum_{i=1}^{n} h_i] \cdot [(n-1) \cdot c + \sum_{j=2}^{n} \binom{n}{j} (j/n \cdot \sum_{k=1} c_{jk})]$$

Tässä kaavassa n on indeksoijien lukumäärä, h yhden indeksoijan käyttämien termien määrä, c indeksoijien yhteisesti käyttämien termien määrä ja c_{jk} j:n indeksoijan yhteisesti käyttämien termien määrä.

Myös painotetun johdonmukaisuuden kaavassa kaikkien indeksoijien käyttämät yhteiset termit ovat keskeisiä ja saavat suuren painoarvon. Pienemmän, lähinnä "lohdutusvoittoarvon" saavat ne termit, joita on käyttänyt useampi indeksoija, mutta eivät kuitenkaan kaikki. Jälleen yksi, poikkeavalla tavalla toimiva indeksoija voi laskea indeksoinnin johdonmukaisuusprosenttia huomattavasti vaikka muut indeksoijat toimisivatkin toisiinsa nähden hyvin johdonmukaisesti. Laskettaessa aiemmin taulukossa 1 esitetyn esimerkin neljän indeksoijan termijohdonnmukaisuutta painotetun kaavan mukaisesti se osoittautuu hiukan paremmaksi kuin Rodgersin ja Hooperin kaavalla laskettuna, mutta jää edelleen melko alhaiseksi (taulukko 2).

Painotetun johdonmukaisuuden laskemisen kaava osoittautuu käyttökelpoisemmaksi kuin yksinkertaisen johdonmukaisuuden laskemisen kaava. Johdonmukaisuusprosentti jää nollaksi ainoastaan

Taulukko 2. Esimerkki neljän indeksoijan välisen termijohdonmukaisuuden laskemisesta Rollingin painotetun johdonmukaisuuden kaavalla

Indeksoitava teos: Kauppinen, Timo, Kohtanen Jukka, Työtaistelut ja neuvottelusuhteet Enso-Gutzeit Oy:n Summan tehtailla vuosina 1971-1984. Helsinki: Työelämän suhteiden neuvottelukunta, 1987.

	Indeksoijien käyttämät termit		
indeksoija 1 lakot Suomi puunjalostus- teollisuus	indeksoija 2 Suomi työtaistelut paperiteollisuus lakot yritysdemokratia	indeksoija 3 työtaistelut paperiteollisuus yritysdemokratia	indeksoija 4 työriidat ammattiyhdistys- liike

painotettu johdonmukaisuus-prosentti $= \frac{100}{3+5+3+2} \cdot [(4-1) \cdot 0 + (2/4 \cdot 5 + 3/4 \cdot 0 + 4/4 \cdot 0)] = 19{,}2$

silloin, kun ei löydy yhtään termiä, jota vähintään kaksi eri indeksoijaa olisi käyttänyt. Tällöin nolla prosenttia lienee jopa käyttökelpoinen luku kuvaamaan indeksoijien välistä johdonmukaisuutta. Vaikka painotetun johdonmukaisuuden kaavallakin laskettaessa prosenttiluvut jäävät alhaisiksi, saadaan kuitenkin jonkinlaisia lukuja. Pieniäkin lukuja voidaan käyttää vertailussa apuna, koska suurempi johdonmukaisuus saa suuremman prosenttiluvun. Aina voidaan sanoa, että kuusi on suurempi kuin viisi ja viisi on suurempi kuin neljä.

4.1.3. Johdonmukaisuuden laskemisen epäsymmetrinen kaava

Rodgersin ja Hooperin kaavan yhtenä puutteena oli se, että se antoi saman tuloksen molemmille osapuolille, vaikka tilanne saattoi olla se, että henkilö x oli johdonmukaisempi suhteessa henkilöön y kuin henkilö y suhteessa henkilöön x. Johdonmukaisuus voidaankin laskea paitsi symmetrisesti myös epäsymmetrisesti, jolloin kumpikin toimija saa oman johdonmukaisuuslukunsa. Johdonmukaisuutta epäsymmetrisesti laskettaessa yhteisten termien määrää verrataan indeksoijan/hakijan omien termien määrään seuraavasti (7.1 ja 7.2):

$$(7.1)\ CT_{1,2} = \frac{|T_1 \cap T_2|}{|T_1|} \quad \text{ja} \quad (7.2)\ CT_{2,1} = \frac{|T_1 \cap T_2|}{|T_2|}$$

Tässä kaavassa $CT_{1,2}$ tarkoittaa indeksoijan 1 johdonmukaisuusprosenttia suhteessa indeksoijaan 2 ja $CT_{2,1}$ indeksoijan 2 johdonmukaisuusprosentti suhteessa indeksoijaan 1. T_1 tarkoittaa indeksoijan 1 käyttämiä termejä ja T_2 indeksoijan 2 käyttämiä termejä.

Laskettaessa useamman indeksoijan välistä johdonmukaisuutta tulee aluksi verrata jokaista paria epäsymmetrisesti toisiinsa. Tämän jälkeen lasketaan kaikkien henkilökohtaisten johdonmukaisuuksien keskiarvo. Esimerkki epäsymmetrisestä johdonmukaisuuden laskemisesta on esitetty taulukossa 3.

Usemman henkilön välinen, epäsymmetrisesti laskettu johdonmukaisuus voidaan havainnollisesti esittää myös matriisina. Taulukon 3 esimerkissä esiintyvät johdonmukaisuusluvut on kuvattu matriisina taulukossa 4. Siitä näkee heti, että indeksoija 4, joka saa johdonmukaisuusprosentiksi nollan, on epäjohdonmukaisin verrattiin hänen toimintaansa kehen tahansa toiseen indeksoijaan. Johdonmukaisimmin suhteessa toisiin toimii indeksoija 2, joka valitsee samoja termejä kuin sekä indeksoija 1 että indeksoija 3. Kun matriisin avulla on tunnistettu johdonmukaisimmin toimiva(t) henkilö(t),

Taulukko 3. Esimerkki neljän indeksoijan välisen termijohdonmukaisuuden laskemisesta käyttäen epäsymmetristä kaavaa (7)

Indeksoitava teos: Kauppinen, Timo, Kohtanen Jukka, Työtaistelut ja neuvottelusuhteet Enso-Gutzeit Oy:n Summan tehtailla vuosina 1971-1984. Helsinki: Työelämän suhteiden neuvottelukunta, 1987.

	Indeksoijien käyttämät termit		
indeksoija 1 lakot Suomi puunjalostus- teollisuus	indeksoija 2 Suomi työtaistelut paperiteollisuus lakot yritysdemokratia	indeksoija 3 työtaistelut paperiteollisuus yritysdemokratia	indeksoija 4 työriidat ammattiyhdistys- liike
CT1,2 = 100 · 2/3 = 67	CT1,4 = 100 · 0/3 = 0	CT2,4 = 100 · 0/5 = 0	
CT2,1 = 100 · 2/5 = 40	CT4,1 = 100 · 0/2 = 0	CT4,2 = 100 · 0/2 = 0	
CT1,3 = 100 · 0/3 = 0	CT2,3 = 100 · 3/5 = 60	CT3,4 = 100 · 0/3 = 0	
CT3,1 = 100 · 0/3 = 0	CT3,2 = 100 · 3/3 =100	CT4,3 = 100 · 0/2 = 0	
kaikkien indeksoijien välinen johdonmukaisuus on 267/12 = 22,3			

voidaan aineistosta etsiä syitä muita korkeampiin johdonmukaisuuslukuihin. Esimerkkitapauksessa syy on varsin yksinkertainen. Indeksoija 2 käyttää termejä enemmän kuin muut indeksoijat. Osa hänen termeistään sopii yhteen indeksoijan 1 ja osa indeksoijan 3 käyttämiin termeihin. Sensijaan indeksoijien 1 ja 3 termit eivät kohtaa.

4.2. Toimijoiden sisäisen johdonmukaisuuden laskeminen

Samoja kaavoja, joita käytetään toimijoiden (indeksoijien ja hakijoiden) välistä johdonmukaisuutta laskettaessa, voidaan käyttää myös toimijoiden sisäistä johdonmukaisuutta laskettaessa. Tällöin toimijan 1 korvaa kaavassa toimijan x toiminta tilanteessa 1 ja toimijan 2 korvaa toimijan x toiminta tilanteessa 2. Esimerkiksi indeksoijan x kahdessa eri tilanteessa suorittaman indeksoinnin johdonmukaisuudesta on esitetty taulukossa 5.

Toimijan sisäistä johdonmukaisuutta tarkasteltaessa kannattaa huomio kiinnittää siihen, milloin kaksi eri tilannetta voidaan käsitellä selvästi eri tilanteina. Saman tehtävän suorittamista kahdesti saman päivän aikana ei tutkimusaineistoa kerättäessä voida vielä käsitellä kahtena eri tilanteena, koska jälkimmäisellä kerralla toimijan voi olettaa muistavan ensimmäisen kerran suorituksensa. Toimijoiden sisäistä johdonmukaisuutta tarkasteltaessa kahden eri tilanteen väliaikana on käytetty sekä

Taulukko 4. Johdonmukaisuusmatriisi taulukon 3 esimerkissä esitetyistä johdonmukaisuuksista

INDEKSOIJA	Indeksoija 1	Indeksoija 2	Indeksoija 3	Indeksoija 4
Indeksoija 1	-	67	0	0
Indeksoija 2	40	-	60	0
Indeksoija 3	0	100	-	0
Indeksoija 4	0	0	0	-

Taulukko 5: Esimerkki indeksoijan sisäisen termijohdonmukaisuuden laskemisesta

Indeksoitava teos: Kauppinen, Timo, Kohtanen Jukka, Työtaistelut ja neuvottelusuhteet Enso-Gutzeit Oy:n Summan tehtailla vuosina 1971-1984. Työelämän suhteiden neuvottelukunta, Helsinki, 1987.

Indeksoijan käyttämät termit	
tilanne 1	tilanne 2
työtaistelut	lakot
Suomi	Suomi
paperiteollisuus	

Indeksoijan sisäinen johdonmukaisuus eri kaavoilla laskettuna
Rodgersin ja Hooperin kaava: $100 \cdot 1/(1+2+1) = 25$
Rollingin yksinkertainen johdonmukaisuuden kaava : $100 \cdot 1/(3+2) = 20$
Epäsymmetrinen kaava:
$(100 \cdot 1/3 + 100 \cdot 1/2) / 2 = 41{,}7$

yhden kuukauden (Iivonen 1989, 9-10) että kahden kuukauden (Fidel 1987, 60, Iivonen 1992, 117) jaksoja.

4.3. Käsitejohdonmukaisuuden laskeminen

Termien perusteella laskettu johdonmukaisuus jää monesti melko alhaiseksi. Se, mitä pidetään alhaisena, on toki sopimuksenvarainen asia. Voimme hyvinkin sopia, että esimerkiksi 15,6 %:n johdonmukaisuutta pidetään jo korkeana johdonmukaisuutena. Jos asteikko kuitenkin on nollasta sataan, niin 15,6 assosioituu eittämättä alhaiseksi prosenttilukemaksi. Taulukoissa 1, 2 ja 3 esitetyssä esimerkissä kuvatut indeksoijien indeksointitulokset eivät kuitenkaan ole niin kaukana toisistaan, että ilman muuta voitaisiin sanoa indeksoijien toimineen keskenään hyvin epäjohdonmukaisesti. Niinpä termijohdonmukaisuuden lisäksi kannattaakin tarkastella käsitejohdonmukaisuutta, jolloin edellä esitetyissä kaavoissa termit korvataan käsitteillä. Tällöin pitää kuitenkin määritellä se, milloin eri termien voidaan katsoa viittaavan samaan käsitteeseen.

Tiukasti ottaen ainoastaan synonyymiset ilmaisut voivat viitata samaan käsitteeseen (Karlsson 1980, 248-249, Häkkinen 1990, 86). Aivan oma kysymyksensä on, onko todellisia synonyymejä olemassakaan tai ovatko ne kovin yleisiä (Hutchins 1975, 37, Lancaster 1986, 60). Jos oletetaan, että todelliset synonyymit ovat hyvin harvinaisia, ei eri synonyymien katsominen samaksi käsitteeksi vielä nosta paljoakaan toimijoiden välistä johdonmukaisuusprosenttia siirryttäessä termijohdonmukaisuudesta käsitejohdonmukaisuuteen.

Tiedon tallennuksen ja haun kontekstissa eri termien voi katsoa viittaavan samaan käsitteeseen myös väljemmin kriteerein. Käsitejohdonmukaisuuden laskemisen kaavaa esitellessään Saracevic (1984, 221) ei täsmennä, milloin eri termit lasketaan samaksi käsitteeksi vaan tyytyy toteamaan, että käsitteet ovat hakupyynnöstä peräisin ("derived from a request"). Perusteet sille, milloin eri termit lasketaan samaksi käsitteeksi, on kuitenkin selkeästi ilmoitettava.

Sievert ja Verbeck (1987, 86-98) tutkivat onlinehakua käsittelevän kirjallisuuden indeksointia LISA:ssa ja ERIC:ssä. He tarkastelivat mm. sitä, montako käsitettä artikkeleista oli indeksoitu. Koska yhdessä artikkelissa useampi termi saattoi viitata samaan käsitteeseen, Sievers ja Verbeck joutuivat "supistamaan" useammalla termillä indeksoidun käsitteen yhdeksi. He laskivat useamman termin yhdeksi käsitteeksi seuraavissa tapauksissa:

1) Termit olivat toistensa kieliopillisia tai syntaktisia muunnoksia. Esim. termit information storage and retrieval, information retrieval, computerised information retrieval, online information retrieval ja computerised information storage and retrieval laskettiin yhdeksi käsitteeksi.

2) Termi oli toisen termin suppeampialainen termi, ja suppeampialaisessa ja sen laajempialaisessa termissä esiintyi yksi yhteinen sana. Esim. termit reference services ja library services laskettiin yhdeksi käsitteeksi.

3) Termit olivat synonyymejä.

4) Jos useampaa eri termiä oli käytetty jonkin populaation tai alan indeksointiin, ne laskettiin yhdeksi käsitteeksi. Sievers ja Verbeck laskivat esim. termit medical education ja medical services samaksi käsitteeksi, vaikka ne viittaavatkin selvästi eri käsitteeseen. He perustelivat ratkaisunsa sillä, että tutkituissa lehdissä (Online, Online Review ja Database) nämä termit hyvin todennäköisesti viittasivat samaan käsitteeseen.

Sieversin ja Verbeckin luettelo sisältää eri tason

tekijöitä. Mukana on sekä kielen syntaktinen rakenne (kohta 1), termien väliset semanttiset suhteet (kohdat 2 ja 3) että johdonmukaisuuden tarkastelun kohteena oleva aihealue (kohta 4). Etenkin Sieversin ja Verbeckin kohta 4 saattaa tuottaa joissakin tilanteissa ongelmia. Ajatellaanpa esimerkkinä naistutkimusta käsittelevää tietokantaa. Siihen indeksoidaan käyttäen sanastoa, jossa on paljon "naistermejä" (naisautoilijat, naisen asema, naiset, naisjohtajat, naisnäkökulma, naistutkimus jne.). Sieversin ja Verbeckin ohjeen mukaisesti eri naistermit voitaisiin laskea samaksi käsitteeksi, koska kyseessä on tietty populaatio (naiset). Kuitenkin tietokannan konteksti huomioon ottaen esim. termit naisjohtajat ja naisautoilijat on syytä käsitellä eri käsitteinä johdonmukaisuutta laskettaessa. Jotkut naisjohtajat voivat toki olla naisautoilijoita, ja saattaapa joku naisautoilija olla naisjohtajakin. Naisjohtajia ja naisautoilijoita koskevat haut ovat kuitenkin käsitteellisesti eri hakuja. Haettaessa naisjohtajia käsitteleviä dokumentteja termillä naisautoilijat toimitaan epäjohdonmukaisesti suhteessa niihin hakijoihin, jotka kutsuvat naisjohtajia naisjohtajiksi.

Yksinkertaisemman, ja ehkä helpommin sovellettavissa olevan lähtökohdan eri termien samaksi käsitteeksi laskemiselle saa tyytymällä tarkastelemaan niitä tekijöitä, joilla termien väliset suhteet normitetaan dokumentaatiokielissä (ks. esim. Hutchins 1975, 22-24, 37-42, Documentation 1984, 13-15, 30-32, Lancaster 1986, 35-71, Aitchison & Gilchrist 1987, 34-50). Tällöin voidaan kiinnittää huomio saman termin vapaa termi- ja asiasanavariaatioihin, yksikkö- ja monikkomuotoihin sekä termien välisiin semanttisiin suhteisiin.

Eri termien samaksi käsitteeksi hyväksymisen alaa voidaan laajentaa vaiheittain. Ensimmäisessä vaiheessa voidaan samaksi käsitteeksi katsoa selvästi samaan käsitteeseen viittaavat eri termit. Tällaisia tapauksia ovat saman termin yksikkö- ja monikkomuodot, saman termin vapaa termi- ja asiasanavariaatiot, saman termin eri kohdasta katkaistut variaatiot sekä synonyymit ja kvasisynonyymit.

Toisessa vaiheessa voidaan samaksi käsitteeksi hyväksyä myös ne termit, joiden välillä on selvä hierarkkinen suhde. Tällöin johdonmukaisuutta laskettaessa pysytään vielä samassa käsitehierarkiassa, mutta käsitteen ilmaisu hierarkian eri tasoilla hyväksytään yhdeksi ja samaksi yksiköksi.

Kolmannessa vaiheessa voidaan käsitejohdonmukaisuutta laskettaessa ottaa mukaan myös yleisimmin tunnetut assosiaatiosuhteet[3] ja hyväksyä samaksi käsitteeksi ne termit, joiden välillä em. yleisesti tunnettu assosiaatiosuhde esiintyy. Tällöin käsite ymmärretään jo aika väljästi. Käsitejohdonmukaisuuden asemasta voitaisiin tässä tapauksessa puhua myös aspektijohdonmukaisuudesta. Ilmaisun kohteena on tällöin tietyn kohteen (dokumentin, hakupyynnön) tietty aspekti.

Termi- ja käsitejohdonmukaisuuden ero saadaan näkyväksi esimerkin avulla. Aiemmissa esimerkeissä hakijoiden välinen termijohdonmukaisuus jäi melko alhaiseksi siitä huolimatta, että indeksoijien valinnat olivat samansuuntaisia. Jokaisella indeksoijalla esiintyy työtaisteluihin viittaava termi, mutta indeksoijat lähestyvät sitä hierarkian eri tasoilla "puhuen" työtaisteluiden ohella lakoista ja työriidoista. Termit lakot, työtaistelut ja työriidat voidaan käsitejohdonmukaisuutta laskettaessa hyväksyä samaksi käsitteeksi. Lakot ovat aina työtaisteluita, jotka ovat työriitoja. Samoin termien puunjalostusteollisuus ja paperiteollisuus välillä vallitsee hierarkkinen suhde, paperiteollisuushan on puunjalostusteollisuutta. Indeksoija 2 ilmaisee käsitteen työtaistelut kahdella eri termillä, joten vaikka hän indeksoinnissa käyttää viittä termiä, indeksoituja käsitteitä hänellä on kuitenkin vain neljä. Taulukossa 6 on esitetty aiemmissa esimerkeissä (taulukot 1, 2 ja 3) esitetyn indeksoinnin käsitejohdonmukaisuudet eri kaavoilla laskettuna. Myös käsitejohdonmukaisuudet vaihtelevat sen mukaan, mitä kaavaa niiden laskemiseen on käytetty. Oleellista tässä yhteydessä on kuitenkin se, että käsitejohdonmukaisuusluvut ovat selvästi termijohdonmukaisuuslukuja korkeammat. Tarkastelemalla termien asemasta käsitteitä ja laskemalla indeksoijien (hakijoiden) välinen johdonmukaisuus käsitteiden perusteella, päästään siis selvästi korkeampiin johdonmukaisuusprosentteihin, kuin vertaamalla indeksoijien (hakijoiden) termejä merkki merkiltä.

Hakukäsitteiden johdonmukaisuutta laskettaessa joudutaan vastaamaan paitsi kysymykseen, milloin eri termit viittaavat samaan käsitteeseen, myös kysymykseen, miten Boolen logiikan JA, TAI ja EI -operaattorit vaikuttavat siihen, onko kyseessä sama vai eri hakukäsite. Saracevicin (1984, 228) mukaan eri käsitteet voivat olla hakulausekkeessa yhdistetty toisiinsa millä tahansa operaattorilla.

Boolen operaattorien vaikutus siihen, onko kyseessä yksi vai useampi hakukäsite, voidaan määritellä myös operaattoreihin sisältyvän logiikan kautta. TAI -operaattorilla lasketaan eri vaihtoeh-

Taulukko 6. Esimerkki neljän indeksoijan välisen käsitejohdonmukaisuuden laskemisesta

Indeksoitava teos: Kauppinen, Timo, Kohtanen Jukka, Työtaistelut ja neuvotteluyhteet Enso-Gutzeit Oy:n Summan tehtailla vuosina 1971-1984. Helsinki: Työelämän suhteiden neuvottelukunta, 1987.

	Eri indeksoijien indeksoimat käsitteet		
indeksoija 1	indeksoija 2	indeksoija 3	indeksoija 4
työtaistelut*	Suomi	työtaistelut	työtaistelut*
Suomi	työtaistelut*	paperiteollisuus	ammattiyhdistysliike
paperiteollisuus*	paperiteollisuus	yritysdemokratia	
	yritysdemokratia		

* Käsiteanalyysin avulla muunnettu termi

Rodgersin ja Hooperin kaava: 45 %
Rollingin yksinkertainen johdonmukaisuuden kaava : 8,3 %
Rollingin painotetun johdonmukaisuuden kaava : 47,9 %
Epäsymmetrinen kaava: 60,4 %

toja yhteen. Jos TAI -operaattoria on käytetty samaan käsitteeseen viittavien eri termien yhdistämiseen, voidaan uusi joukko mieltää yhdeksi ja samaksi hakukäsitteeksi. JA -operaattorilla haetaan useamman käsitteen leikkausta. Jos hakija liittää JA -operaattorilla yhteen samaan käsitteeseen viittaavia eri termejä, ne tulee laskea eri hakukäsitteiksi, koska hakija ilmoittaa hakevansa käsitteiden leikkausta. Tässä tapauksessa hakija oletettavasti tekee tyypillisen logiikkavirheen, mikä omalta osaltaan alentaakin johdonmukaisuutta. EI -operaattorilla rajataan jotakin jostakin pois. Jos hakija käyttää EI -operaattoria yhdistämään samaan hakukäsitteeseen viittaavia termejä, hän haluaa rajata käsitteestä pois toisen käsitteen/muita käsitteitä, ja samaan käsitteeseen viittaavat eri termit lasketaan tällöin useammaksi hakukäsitteeksi. Esimerkki Boolen operaattorien vaikutuksesta käsitejohdonmukaisuuteen on esitetty taulukossa 7.

5. Lopuksi

Johdonmukaisuuslukuja laskemalla saadaan näkyviin useamman eri toimijan tai yhden ja saman toimijan eri tilanteissa suorittaman toiminnan (epä)johdonmukaisuuden suuruus. Pelkät luvut kertovat vasta, onko (epä)johdonmukaisuutta. Ne auttavat kuitenkin suunnistamaan etsittäessä toimijoiden välisiä yhtäläisyyksiä ja erilaisuuksia. Johdonmukaisuuslukujen lisäksi tarvitaan kuitenkin myös toisenlaista aineistoa (esim. toimijoiden sanallisia selityksiä) sen selvittämiseen, mitkä tekijät (epä)johdonmukaisuutta aiheuttavat.

Johdonmukaisuutta on tiedon tallennuksen ja haun tutkimuksessa tarkasteltu ennenkaikkea käytettyjen termien johdonmukaisuutena, jonkin verran myös valittujen hakukäsitteiden johdonmukaisuutena sekä löydettyjen hakujoukkojen päällekkäisyytenä. Mikään ei kuitenkaan estä käyttämästä johdonmukaisuuden laskemiseen kehiteltyjä kaavoja myös tiedon tallennuksen ja haun muiden ilmiöiden tarkasteluun. Yksi kokeilemisen arvoinen alue voisi olla eri hakuympäristöissä työskentelevien hakijoiden saaman tiedonhaun koulutuksen yhdenmukaisuus. Tällöin pitäisi vain valita yhdenmukaisuuden laskemiseen käytettävä kaava ja päättää, mikä on tarkasteltava yksikkö (esim. kurssi) ja milloin eri kurssit voidaan laskea samaksi (esim. lasketaanko Tampereen yliopiston täydennyskoulutuskeskuksen eri aikoina järjestämät tiedon tallennus ja haku -kurssit samaksi).

Kirjastotieteen ja informatiikan muilla alueilla johdonmukaisuuden laskemisessa käytettyjen kaavojen avulla voitaisiin laskea esimerkiksi eri ammateissa toimivien henkilöiden tiedonhankintakanavien tai eri kirjastojen aikakauslehtikokoelmien päällekkäisyys. Jälleen pitäisi vain valita kaava ja tehdä laskemista koskevat päätökset (mikä on yksikkö, milloin eri yksiköt voidaan laskea samaksi).

Taulukko 7. Esimerkki Boolen operaattorien vaikutuksesta neljän hakijan väliseen johdonmukaisuuteen (johdonmukaisuus laskettu epäsymmetrisellä kaavalla)[1]

Hakupyyntö: Euroopan yhdentyminen			
		Hakijoiden käyttämät termit	
hakija 1 Eurooppa ja (integraatio tai taloudellinen integraatio tai poliittinen integraatio)	hakija 2 Eurooppa ja yhdentyminen	hakija 3 Eurooppa ja integraatio ja taloudellinen integraatio	hakija 4 (Eurooppa ei Suomi) ja (integraatio tai yhdentyminen)

A. Termijohdonmukaisuus: 58.3 %
- Termejä verrataan merkki merkiltä.
- Hakijan 1 termien määrä on neljä, hakijan 2 kaksi, hakijan 3 kolme ja hakijan 4 neljä.

B. Käsitejohdonmukaisuus vaiheessa 1: 75.7 %
- Samaan käsitteeseen viittaavat synonyymit integraatio ja yhdentyminen lasketaan samaksi hakukäsitteeksi.
- Hakija 4 käyttää termejä integraatio ja yhdentyminen, ja yhdistää ne toisiinsa TAI-opertaattorilla (viittaa samaan käsitteeseen kahdella termillä), joten ne lasketaan hänellä yhdeksi hakukäsitteeksi. Tässä vaiheessa hakijan 4 hakukäsitteiden määrä on kolme. Hakijan 1 hakukäsitteiden määrä on neljä, hakijan 2 kaksi ja hakijan 3 kolme.

C. Käsitejohdonmukaisuus vaiheessa 2: 83.4 %
-Samaan käsitteeseen viittaavat synonyymit (integraatio/yhdentyminen) lasketaan edelleen samaksi hakukäsitteeksi (ks. kohta B). Samoin samaan käsitteeseen hierarkian eri tasoilla viittaavat termit lasketaan samaksi hakukäsitteeksi.
- Hakijan 1 termit integraatio, taloudellinen integraatio ja poliittinen integraatio lasketaan hänellä yhdeksi hakukäsitteeksi, koska niiden välissä on TAI-operaattori. Hakijalla 1 on tässä vaiheessa vain 2 hakukäsitettä.
- Hakijan 2 hakukäsitteiden määrä on edelleen kaksi.
- Hierarkkisesta suhteesta huolimatta hakijalla 3 olevat termit integraatio ja taloudellinen integraatio lasketaan kahdeksi eri hakukäsitteeksi, koska niiden välissä on JA-operaattori. Hakijan 3 hakukäsitteiden määrä on 3.
- Hierarkisesta suhteesta huolimatta hakijalla 4 olevat termit Eurooppa ja Suomi lasketaan kahdeksi eri hakukäsitteeksi, koska niiden välissä on EI-operaattori. Hakijan 4 hakukäsitteiden määrä on 3.

1. Termien katkaisut sekä vapaa termi- ja asiasanavaihtelut on jätetty pois esimerkistä. Niiden tuoman vaihtelun mukaanotto olisi luonnollisesti alentanut johdonmukaisuuslukuja.

Kirjastotieteen ja informatiikan ulkopuolella johdonmukaisuuden laskemisessa käytettyjen kaavojen avulla voitaisiin tarkastella vaikkapa poliitikkojen esiintymistä julkisuudessa tai eri tutkimusten metodologisia ratkaisuja, kunhan ensin päätettäisiin, miten tarkasteltava toiminta operationalisoidaan ja mitkä toimintaa kuvaavat ilmiöt lasketaan samoiksi. Tutkimuskohteen valinnassa tarvitaan aina tieteellistä mielikuvitusta, itse tutkimuksen suorittamisessa myös raakaa työtä, selviä sääntöjä ja kurinalaisuutta.

Hyväksytty julkaistavaksi 26.4.1993.

Viitteet

1 Termiä *termi* käytetään tässä väljemmin kuin mitä esim. Haarala (1981, 16) sen käytöltä edellyttää. Haarala esittää, että *termi* on 1) tarkasti määritellyn käsitteen nimi, 2) alalla yleisesti tunnettu ja hyväksytty ja 3) käyttöön vakiintunut. Kaikki tiedon tallennuksessa ja haussa käytetyt termit (sanat, sanaliitot, merkkijonot) eivät tätä tietenkään ole. Mutta termi *termi* viittaa kirjastotieteessä ja informatiikassa eri käsitteeseen kuin kielitieteessä. *Termi* on tiedon tallennuksen ja haun tutkimuksessa yleisesti tunnettu ja hyväksytty sekä käyttöön vakiintunut termi puhuttaessa sanoilla ja sanaliitoilla tapahtuvasta tiedon tallennuksesta ja hausta.

2 Tähän kaavaan viitataan johdonmukaisuustutkimuksissa toisinaan vain Hooperin kaavan nimellä (esim. Funk, Reid & McGoogan 1983). Robert Hooper perusti kuitenkin oman kaavansa (vuonna 1965) Dorothy Rodgersin tavalle (esitetty jo vuonna 1961) määritellä ja laskea indeksoijien välinen johdonmukaisuus (Leonard 1977, 3). Lancaster (1968, 178) nimeää kaavan Rodgersin ja Hooperin kaavaksi. Tässä artikkelissa noudatetaan Lancasterin esimerkkiä.

3 Assosiaatiosuhteet voivat olla mitä moninaisempia. Kuitenkin useampi auktoriteettijulkaisu luettelee melko yhdenmukaisesti joitakin assosiaatiosuhteiden lajeja. Tällaisia ovat 1) tieteenala ja sen kohteet tai ilmiöt, 2) toiminta ja sen suorittaja, 3) toiminta ja sen väline, 4) toiminta ja sen tuote, 5) toiminta ja sen kohde, 6) käsite ja sen ominaisuus, 7) toiminta ja sen (siihen liittyvä) ominaisuus, 8) käsite ja sen alkuperä, 9) käsitteet, jotka liittyvät toisiinsa kausaalisuhteen perusteella, 10) toisensa poissulkevat vaihtoehdot (antonyymit), 11) asia ja sen vastatoimija, 12) suureet ja niiden mittayksiköt ja 13) raaka-aine ja tuote (Hutchins 1975, Lancaster 1986, Aitchison & Gilchrist 1987, Documentation 1986, Suomenkielisen 1987).

Lähteet

Aitchison, Jean, Gilchrist, Alan (1987). Thesaurus Construction: a Practical Manual. London: Aslib, the Association for Information Management.

Chan, Lois Mai (1989), Inter-Indexer Consistency in Subject Cataloging. - Information Technology and Libraries (December): 349-358.

Cleverdon, Cyril (1984). Optimizing Convient Online Access to Bibliographic Databases. - Information Services & Use (4): 37-47.

Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri, ISO 2788-1986.

Fidel, Raya (1985). Individual Variability in Online Searching Behavior. In: ASIS '85: Proceedings of the American Society for Information Science 48th Annual Meeting: Vol. 22: 1985 October 20-24, Las Vegas, Nevada. Ed. by Carol A. Parkhurst, p. 69-72. White Plains, NY: Knowledge Industry Publications.

Fidel, Raya (1987). What Is Missing in Research about Online Searching Behavior. - Canadian Journal of Information Science 12 (3-4): 54-61.

Funk, Mark E., Reid, Carolyn Anne, McGoogan Leon S. (1983). Indexing Consistency in Medline. - Bulletin of Medical Library Associations 71 (2): 176-183.

Haarala, Risto, Sanastotyön opas. Hki: Kotimaisten kielten tutkimuskeskus, 1981

Hutchins, W. J. (1975). Languages of Indexing and Classification. Stevenage: Peter Peregrinus.

Häkkinen, Kaisa (1990). Yleisen kielitieteen peruskurssi. Turku: Turun yliopisto.

Iivonen, Mirja (1989). Indeksointituloksen riippuvuus indeksointiympäristöstä. Tampere, Tampereen yliopisto.

Iivonen, Mirja (1992). Factors Affecting the Analysis of Requests and the Formulation of Query Statements. In: Cognitive Paradigms in Knowledge Organisation, p. 112-129. Bangalore: Sarada Ranganathan Endowment for Library Science.

Jones, Kevin P. (1983). How Do We Index: a Report of Some Aslib Informatics Group Activity. - Journal of Documentation 39 (1): 1-23.

Karlsson, Fred (1980). Johdatus yleiseen kielitietee-

seen. Vaasa: Gaudeamus.

Katzer, J., McGill, M.J., Tessier, J. A., Frakes, W., DasGupta, P. (1982). A Study of the Overlap among Document Representations. - Information Technology: Research and Development 1 (4): 261-274.

Lancaster, F.W. (1968). Evaluation of the Medlars Demand Search Service. Washington: National Library of Medicine.

Lancaster, F. W. (1986). Vocabulary Control for Information Retrieval. Arlington, Virginia: Information Resources Press.

Lancaster, F.W. (1991). Indexing and Abstracting in Theory and Practice. London: The Library Association.

Leonard, Lawrence E. (1977). Inter-Indexer Consistency Studies 1954-1975: a Review of the Literature and Summary of Study Results. Champaign: University of Illinois.

Markey, Karen (1984). Interindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials. - Library and Information Science Research (6): 155-167.

Rolling, R. (1981). Indexing Consistency, Quality and Efficiency. - Information Processing & Management 17 (2): 69-76.

Saracevic, Tefko (1984). Measuring the Degree of Agreement Between Searchers. In: ASIS '84: Proceedings of the American Society for Information Science 47th annual meeting: Vol. 21: 1984 October 21-25, Philadelphia, Pennsylvania. Compiled by Barbara Flood, Joanne Witiak, Thomas H. Hogan, p. 227-230. White Plains, NY: Knowledge Industry Publications.

Saracevic, Tefko, Kantor, Paul, Chamis Alice Y., Trivison, Donna (1987). Experiments on the Cognitive Aspects of Information Seeking and Information Retrieving: Final Report for National Science Foundation Grants IST-8505411. Washington, D.C.: National Techical Information Service; Educational Research Information Center.

Saracevic, Tefko, Kantor, Paul, Chamis Alice Y., Trivison, Donna (1988). A Study of Information Seeking and Retrieving. I. Background and Methodology. - Journal of the American Society for Information Science 39 (3): 161-176.

Saracevic, Tefko, Kantor, Paul (1988). A Study of Information Seeking and Retrieving. III. Searchers, Searches, and Overlap. - Journal of the American Society for Information Science 39 (3): 197-216.

Sievert, Mary Ellen, Verbeck, Alison (1987). The Indexing of the Literature of Online Searching: a Comparison of ERIC and LISA. - Online Review 11 (2): 95-104.

Suomenkielisen tesauruksen laatimis- ja ylläpito-ohjeet SFS 5471-1987.