

MAGISTERUPPSATS I BIBLIOTEKS- OCH INFORMATIONSVETENSKAP
VID BIBLIOTEKS- OCH INFORMATIONSVETENSKAP/BIBLIOTEKSHÖGSKOLAN
2005:97
ISSN 1404-0891

Biblioteks- och informationsvetenskap
som kunskapsområde
En bibliometrisk studie

MATS NIKLASSON



HÖGSKOLAN I BORÅS

© Författaren

Mångfaldigande och spridande av innehållet i denna uppsats
– helt eller delvis – är förbjudet utan medgivande.

Svensk titel: Biblioteks- och informationsvetenskap som kunskapsområde
– En bibliometrisk studie.

English title: The Cognitive Structure of Library and Information Science
– a Bibliometric Study.

Författare: Mats Niklasson

Kollegium: 4

Färdigställt: 2005

Handledare: Irene Wormell

Abstract: The objective of the study is to map the cognitive structure of Library and Information Science, using a quantitative, bibliometric method. This empirical study shows how co-citation analysis combined with word-frequency analysis can be used to describe the subject content of the field.

The research questions to be answered are:

Which clusters of authors can be identified?

How can these clusters be described in respect of their subject content, if the descriptors associated with the authors in the clusters are counted?

The results shows that the cognitive structure of Library and Information Science can be described by 13 clusters of authors, where each cluster is representing a specific aspect of the subject-domain.

For the description of the subject content of the clusters, descriptors (DE) of the LISA database have been used.

Nyckelord: Biblioteks- och informationsvetenskap, bibliometri, domänanalys, kognitivt innehåll, cociteringsanalys, författare, vetenskaplig kommunikation, klusteranalys

1. INLEDNING	4
1.1. Syfte och frågeställningar	5
1.2. Avgränsningar	5
1.3. Disposition	5
2. TEORETISK RAM	6
2.1. Allmänt om bibliometri	6
2.1.1. Kunskapsproduktion och vetenskaplig kommunikation	6
2.1.2. Referering och citering	8
2.2. Bibliometriska undersökningar	8
2.3. Bibliometriska metoder	8
2.3.1. Cociteringsanalys	11
2.3.1.1. Allmänt om cociteringsanalys	11
2.3.1.2. Rådata och Råmatris	12
2.3.1.3. Likhetsmått	12
2.3.1.4. Klusteranalys	17
2.3.1.5. Hierarkisk klustring	17
2.3.1.6. Multi Dimensional Scaling	18
3. TIDIGARE FORSKNING	19
4. METOD	20
4.1. Citeringsdatabaser	20
4.2. Bibexcel	23
4.3. LISAs thesaurus och deskriptorer	24
5. RESULTAT	25
6. DISKUSSION	34
7. SAMMANFATTNING	38
REFERENSER	40

1. INLEDNING

Det finns många skäl till att bestämma ett ämnes kunskapsinnehåll, eller med andra ord dess kognitiva struktur.

Ur en forskningsinstitutionens synvinkel är det viktigt att institutionen förvärvar och underhåller beståndet av de monografier och de tidskrifter som är centrala för, eller berör den forskning som bedrivs av institutionen.

Studenten eller den intresserade allmänheten kan ha ett intresse att veta vad ämnet innefattar, vilka författare och vilka verk som är centrala, vad som utgör den centrala kunskapen, och vad som är kanon, om nu någon sådan finns.

Såväl forskare som studenter har ett intresse att bestämma var den så kallade forskningsfronten befinner sig, samt för att utreda tidigare forskning.

Ur ett forskningssociologiskt perspektiv kan det vara av intresse att bestämma var forskning och undervisning i ett ämne sker, hur den är fördelad geografiskt, var olika specialiteter finns - ansatsen kan vara såväl global som lokal.

Från ett finansieringsperspektiv är det av intresse att avgöra om den forskning, som bedrivs eller önskas bedrivas, befinner sig i centrum eller i periferin av ett ämne. De sista delarna tangerar problemen kring evaluering av forskning och forskningsresultat samt forsknings- och utbildningspolitik.

Ur ett kunskapsorganisatoriskt perspektiv kan det vara av intresse att veta vilka ämnen som behandlas inom en vetenskap eftersom klassifikation av dokument bygger på idén om ”*literary warrant*” d.v.s. att klassificeringen av dokument ska bygga på vilken litteratur som finns i ämnet.

Det är teoretiskt möjligt att en person, exempelvis en student eller en bibliotekarie, som vill få en kunskap om ett ämnes kunskapsinnehåll, eller om aspekter på ett ämne av något av de angivna skälen, manuellt samlar in alla dokument som kan antas ha någon anknytning till den frågeställning som ska besvaras och sedan försöker strukturera dokumenten och tolka dess innehåll. I praktiken är det omöjligt.

Ett sätt att göra uppgiften hanterbar är att använda genvägar och schabloniseringar i form av olika bibliografiska databaser och datorprogram för databehandling av fakta. Av nödvändighet blir resultatet därmed inte uttömmande eller exakt, det kanske inte ens är önskvärt men fullt tillräckligt för att tillfredsställa det aktuella informationsbehovet.

Jag har därför valt att studera ett ämne så som detta representeras i en citeringsdatabas, strukturerat genom användning av automatiserad cociteringsanalys av författare samt förekomsten av författarna i en bibliografisk databas. Som studieobjekt har jag valt biblioteks- och informationsvetenskapen eftersom detta ligger närmast till hands då uppsatsen skrivs inom detta ämne och dels för att det inte finns något behov av kompetens utanför institutionen vid för validering av resultaten. Uppsatsen är en metodstudie och inte en studie av biblioteks- och informationsvetenskapens förändring över tid.

1.1. Syfte och frågeställningar

Syftet med den här studien är att beskriva biblioteks- och informationsvetenskapen eller Library and Information Science (LIS) som kunskapsområde, det vill säga vilka ämnen som ägnas vetenskapligt studium och, som i förlängningen kommer studenter och praktiker till del. Mer konkret är syftet att beskriva ett ämne, biblioteks- och informationsvetenskapen eller Library and Information Science (LIS), kunskapsområde, så som ämnet representeras i citeringsdatabasen *Social Science Citation Index*, strukturerat genom automatiserad co-citeringsanalys av författare och beskrivet utifrån de deskriptorer som finns i den bibliografiska databasen LISA.

Med denna metod vill man ge svar på följande frågor:

Vilka grupperingar av författare kan identifieras?

Hur kan dessa grupper beskrivas till sitt ämnesmässiga innehåll om man sammanställer de indexeringstermer som kan associeras med de i klustren ingående författarna ?

1.2. Avgränsningar

Det ämne som ska studeras betecknas i praktiken och i litteraturen som biblioteks- och informationsvetenskap, Library and Information Science. Förkortningarna B & I och LIS förekommer också. Jag kommer fortsättningsvis att kalla ämnet LIS.

Uppsatsen är en metodstudie och inte en anakron studie av biblioteks- och informationsvetenskapen. Ämnet i sig är av underordnad betydelse och kunde ha varit något annat.

Begreppet ”ämne” är inget okontroversiellt begrepp inom LIS. I den här studien ses detta begrepp utifrån ett pragmatiskt synsätt, i det att de indexeringstermer som tilldelas dokumenten av professionella indexerare och som redovisas i bibliografiska databaser och auktoritetslistor manifesterar dokumentens ämne eller *aboutness* (Rowley & Farrow 2000 s. 124f).

En begränsning i undersökningen är att ämnet LIS, kommer att beskrivas så som detta representeras i en typ av databaser som benämns citeringsdatabaser. Citeringsdatabaserna innehåller inte någon fullständig representation av alla dokument inom LIS. Eftersom citeringsdatabasen är den verklighet som ska beskrivas och därmed har en avgörande betydelse för resultaten kommer den citeringsdatabas som använts att beskrivas närmare i avsnitt 4.1.

Den frekvensberäkning av indexeringstermer som undersökningen resulterar i måste nödvändigtvis tolkas och beskrivas i klartext för att bli användbar. Ett mått av subjektivitet kommer därmed att präglade beskrivningen av ämnet.

1.3. Disposition

Avsnitt 2 innehåller en teoretisk ram där inledningsvis den bibliometriska metoden kortfattat beskrivs. Därefter sätts bibliometrin in i ett forskningskommunikativt sammanhang. Olika typer av bibliometriska undersökningar presenteras. Metoden sätts in i en kontext av andra

kvantitativa metoder inom human- och samhällsvetenskaperna. Olika typer av visualisering av analysresultaten presenteras. Avsnittet 3 innehåller exempel på tidigare forskning där bibliometriska tillämpningar använts i ett informationsförsörjningssammanhang.

I avsnitt 4 presenteras den metod som ligger till grund för ifrågavarande undersökning. Avsnittet innefattar även en presentation av citeringsdatabaser, som används vid bibliometriska undersökningar. Resultatet av undersökningen presenteras i avsnitt 5. I avsnitt 6 diskuteras resultaten i sak och metoden som sådan. I avsnitt 7 görs en sammanfattning.

2. TEORETISK RAM

2.1. Allmänt om bibliometri.

Med bibliometri menas statistisk och matematisk analys av kommunikation som sker genom dokumentation (Kärki och Kortelainen 1998). Bibliometriska metoder, särskilt cociteringsanalys, ger enligt Ding (1998) möjlighet att undersöka vetenskaplig kommunikation bland forskare baserad på skriftlig kommunikation. Metoderna kan tillhandahålla ett makroperspektiv på vetenskaplig kommunikation.

2.1.1. Kunskapsproduktion och vetenskaplig kommunikation

En grundläggande princip inom vetenskapen är att forskningsresultat måste publiceras. Vetenskap är kumulativ och framåtskridande vilket avspeglas i hur forskare förhåller sig till tidigare publicerad forskning. Varje publikation är inbäddad i ett nätverk av referenser och citeringar från många forskare. Referenser och citeringar utgör de formella länkarna mellan forskare eller andra medlemmar av det vetenskapliga kollektivet. Vetenskaplig kunskap kan därför betraktas som ett nätverk bestående av fakta, idéer och forskare ingående i det vetenskapliga nätverket. (Jarneving 2002 med referens till Ziman 1984).

Den vetenskapliga kommunikationen kan vara informell eller formell (Fjällbrant 1997).

Den informella kommunikationen är den muntliga kommunikation som kan uppkomma vid kontakt mellan kollegor och lärare på arbetsplatserna, lektioner, seminarier eller konferenser. Kommunikationen behöver inte vara personlig utan kan ofta ske via e-post, telefon etc.

Med formell kommunikation menar Fjällbrant att de vetenskapliga rönen görs publika genom publicering i någon form. Vilken form av publicering som används varierar mellan olika vetenskapliga discipliner. Inom naturvetenskap och medicin är publicering i vetenskapliga tidskrifter den avgjort viktigaste formen, i andra vetenskaper - publikation i monografier.

Fortsättningsvis kommer all uppmärksamhet i denna studie att riktas mot den formella vetenskapliga kommunikationen.

Persson (1991) åskådliggör den vetenskapliga kommunikationen som ett flöde.

Kunskapsproduktion ?	Dokument ?	Kunskapsanvändning
----------------------	------------	--------------------

Publiceringsmönster har, enligt Fjällbrant (ibid.), historiskt sett styrts av ett antal aktörer eller kunskapsanvändare. Aktörerna har bestämt hur den formella vetenskapliga kommunikationen ser ut i dag. Dessa aktörer var enligt Fjällbrant:

Författarna	Forskarna
Läsarna	Andra forskare Studenter Den intresserade allmänheten
Producenter	Utgivare, förläggare
De som tillhandahåller	Bokhandeln, Bibliotek
Konsumenter	Tillämpare av forskningsresultat, rättighetsinnehavare, akademiska institutioner

Religiösa organisationer

Förutom de religiösa organisationerna framstår de historiskt sett betydelsefulla aktörerna som viktiga kunskapsanvändare även i dag. För egen del vill jag dock lägga till olika privata och offentliga finansiärer av forskning och företagande. Sammanfattningsvis i den här delen kan man därför konstatera att den vetenskapliga kommunikationen tjäna många syften vilket då sätter sin prägel på hur kommunikationen och dokumenten ser ut i dag.

En annan aspekt av den formella vetenskapliga kommunikationen formulerades av Derek J de Solla Price i artikeln *Networks of Scientific Papers* (Price 1965). Genom empiriska undersökningar kunde Price konstatera en typiskt refereringspraxis.

Although most papers produced in the year contain a near average number of bibliographic references, half of these are references to about half of all the papers that have been published in previous years. The other half of the references ties these new papers to a quite small group of earlier ones, and generates a rather tight pattern of multiple relationships. Thus each group of new papers is "knitted" to a small, select part of the existing scientific literature but connected rather weakly and randomly to a much greater part. Since only a small part of the earlier literature is knitted together by the new year's crop of papers, we may look upon this small part as a sort growing tip or epidermal layer, an active research front.

Price menar således att en forskningsfront är ett litet antal av den senast publicerade litteraturen som oftast citeras i den senaste litteraturen. Den äldre litteratur som citeras mer sällan kan anses vara den intellektuella basen i ett ämne.

2.1.2. Referering och citering

Som framgått ovan är begreppen referering och citering centrala i beskrivningen av den vetenskapliga kommunikationen och i detta arbete, begreppen referering och citering ska därför preciseras.

Antag att publikationen A åberopas i publikationen B, antingen som ett rent citat, som hänvisning i brödtexten eller som en referens i referenslistan. Publikation B ger då en referens åt publikation A. Publikation A tar emot en citering av publikation B (Diodato1994). Således: den här publikationen refererar till Diodato (1994) medan Diodato mottar en citering av den här publikationen.

2.2. Bibliometriska undersökningar

Ett sätt att beskriva och analysera vetenskaplig kommunikation och kunskapsanvändning är att göra bibliometriska undersökningar. Med bibliometri menas, som redan nämnts ovan, statistisk och matematisk analys av kommunikation som sker genom dokumentation (Kärki & Kortelainen 1998).

Närliggande begrepp är scientometri och infometri. Enligt Kärki och Kortelainen (ibid.) är infometri det övergripande begreppet som omfattar all kvantitativ analys av information. Skillnaden mellan bibliometri och scientometri är att bibliometrin inskränker sig till att studera publikationer (formell vetenskaplig kommunikation) medan scientometrin även analyserar olagrad information samt förverkligad vetenskap. (Vad författarna menar med förverkligad vetenskap är oklart. Min anmärkning.)

Vid en bibliometrisk undersökning kan, enligt Kärki och Kortelainen (ibid.), vetenskaplig formell kommunikation undersökas utifrån fyra dimensioner.

1. Producenter: enskilda författare eller gruppbildningar såsom forskningsgrupper, institutioner, forskningsområden eller länder. Producenterna undersöks i egenskap av såväl avsändare som mottagare.
2. Produkter: enskilda dokument såsom artiklar, konferensbidrag, böcker, patent eller aggregat såsom tidskrifter eller konferensbidrag.
3. Begrepp: författarnas egna termer i titlar och texter, eller indextermer och klassifikationskoder som används för att beskriva dokumenten.
4. Källciteringar samt motiv för citeringarna.

2.3. Bibliometriska metoder

Nedan följer en redogörelse för huvuddragen i den bibliometriska undersökningsmetodik och några av de bibliometriska metoderna. Syftet är att ge exempel på hur kvantitativ metod kan användas vid studiet av publikationer och publicering. Avsnittet är inte uttömmande.

De bibliometriska metoderna delas, enligt Kärki och Kortelainen (ibid.), in dels i citeringsanalys och dels i övriga analyser baserade på publikationer. Citeringsanalysen i sin tur kan vara av två slag; hänvisningsanalys och källanalys.

Vad först gäller citeringsanalysen och i synnerhet hänvisningsanalysen används denna för att analysera publikationers användning eller icke- användning. De frågor hänvisningsanalysen kan ge svar på är bland annat

- Om författare / publikationer har citerats.
- Hur många som har citerat författaren / publikationen.
- Varifrån citeringarna kommer; från hemlandet eller utlandet, inom samma ämnesområde eller från andra ämnesområden.

Mer konkret kan hänvisningsanalys i form av om författare har citerats, ske genom en publikationsanalys. Publikationsanalys kan göras genom beräkning av vissa mått. Måtten anses (Nisonger 1998) visa hur betydelsefull tidskriften är inom sitt ämnesområde.

Mått som kan redovisas är enligt Kärki och Kortelainen (ibid.), *total cite* - antal citeringar under två år, det s.k. citeringsfönstret (*citation window*), *Journal Impact Factor, JIF* - ett mått som anger hur stort inflytande eller genomslag tidskriften, d.v.s. dess artiklar, har inom forskningen inom ämnet, *Immediacy Index* - hur snabbt en artikel får genomslagskraft samt antalet artiklar under ett visst år. Dessutom redovisas måtten *Cited Half-life* och *Citing Half-life*. De senare måtten är inte mått på tidskriftens genomslagskraft och betydelse, utan snarare en angivelse av under vilka tidsintervall tidskriften är mest i bruk, d.v.s. citerad, uttryckt i termer av halveringstid.

Journal Impact Factor, Immediacy Index och *Cited Half-life* beräknas på följande sätt. Framställningen bygger på Nisonger (1998).

Journal Impact Factor för ett visst år beräknas genom att antalet citeringar det aktuella året tillställs i relation till antalet möjliga citeringar under en viss tid, det så kallade citeringsfönstret, vanligen två år.

$$\text{JIF 2002} = \frac{\text{antal citeringar 2002 till 2001 + 2000}}{\text{antal citerbara objekt 2001+2000}}$$

Immediacy Index beräknas för visst år genom att antal citeringar det aktuella året till objekt samma år ställs i relation till antalet möjliga objekt under det aktuella året

$$\text{Immediacy Index 2002} = \frac{\text{antal citeringar 2002 till objekt från 2002}}{\text{Antal citerbara objekt publicerade 2002}}$$

Cited Half Life

Antag att alla tidskrifter i ISI år 1990 citerade nedanstående årgångar av tidskriften X

1989 = 30

1988 = 20

1987 = 20

1986 = 15

1985 = 14

1984 = 1

100

Cited Half Life beräknas då genom att fördelningen delas på hälften för att bestämma halveringstiden

$$\begin{array}{r} 1989 = 30 \\ 1988 = 20 = 50 \\ 1987 = 20 \\ 1986 = 15 \\ 1985 = 14 \\ 1984 = 1 = 50 \\ \hline 100 \end{array}$$

Cited Half Life bestäms därför till 2 år.

En slutsats man kan dra av den korta halveringstiden är att tidskriften publicerar artiklar som ligger nära forskningsfronten.

En annan form av hänvisningsanalys är enligt Kärki och Kortelainen (ibid.), cociteringsanalysen. Innebär att två specifika publikationer eller författare citeras i samma senare utkomna publikation (Kärki och Kortelainens. 25). Den grundläggande tanken med cocitering är att ju större antal forskare som i sina arbeten citerar samma två forskare, desto troligare är det att det finns ett samband mellan dessa citerade författare eller publikationer. Genom att granska nätverk av publikationer eller författare kan man också åskådliggöra strukturer hos forskningsområden. En närmare beskrivning av cociteringsanalysens metod kommer att göras i avsnitt 2.3.1.1. och i avsnitt 3.

Det undersökta materialet vid källanalysen är vanligtvis litteraturförteckningarna i dokumenten. Litteraturförteckningarna utgör källorna för studier av mängden av litteratur inom ett område samt dess egenskaper och struktur. Källanalysen kan också användas för granskning av forskningskommunikation och därigenom vetenskapens kognitiva och sociala strukturer. Genom källanalys kan man också beskriva i vilken utsträckning ämnesrådet är självtillräckligt, i så motto om referenserna kommer från det egna området eller om och i så fall i hur stor andel av det som kommer från andra ämnesområden.

Bland de övriga bibliometriska metoderna finns metoder vars syfte är att beskriva frekvensen av olika företeelser. Exempel på sådana metoder är tillämpningar av bl.a. Bradfords lag och Lotkas lag.

Bradfords lag (Bradford 1936) beskriver, med stöd i empiriska studier, koncentration och spridning av hur artiklar fördelar sig på tidskrifter. I klartext och i korthet innebär lagen att ett fåtal tidskrifter innehåller en betydande mängd av artiklarna inom ett visst ämnesområde, medan resten av litteraturen sprids ut över ett stort antal tidskrifter. En praktisk tillämpning av Bradfords lag finns inom beståndsutvecklingen där lagen kan tillämpas för att bestämma kärntidskrifterna inom ett ämnesområde.

Lotkas lag (Lotka 1926) beskriver vetenskaplig produktivitet. Kortfattat innebär lagen att man empiriskt kunnat konstatera att endast ett mycket litet antal författare är högproduktiva, medan 60 % av alla författare bara producerar en publikation. Det finns ett mönster som visar att 1/4 av författarna inom ett område, har författat två publikationer, 1/9 har publicerat tre publikationer o.s.v. (1/16 4 publikationer).

Bland de övriga metoderna återfinns också *co-word*analys och samarbetsanalys.

Vid *co-word*analysen används de nyckelord som förekommer i publikationernas titlar, sammanfattningar, eller i publikationernas text. Publikationer som innehåller samma termer kan länkas till varandra och ordnas i kluster. Genom detta kan man visa på samband mellan termerna och genom kartor visualisera koncentration och spridning av termerna.

Samarbetsanalysen utgår från det problem som uppstår när flera författare samarbetar. I samarbetsanalysen undersöks bland annat hur mycket sampublikationer det produceras inom olika områden, hur produktivitet och samarbete är relaterade samt hur mycket sampublikationerna uppmärksammas i jämförelse med publikationer av en författare. Samarbetsanalysen kan också användas för att beskriva nätverk av personer vilket visar ett ämnes sociala struktur.

2.3.1. Cociteringsanalys.

2.3.1.1. Allmänt om cociteringsanalys

Som redogjorts för ovan innebär cocitering att två författare eller publikationer citeras i samma, senare utkomna publikation. Den grundläggande tanken med att tillmäta cocitering någon betydelse är att ju större andel forskare som i sina arbeten citerar samma två forskare eller publikation desto troligare är det att det finns ett samband mellan dessa citerade/citerande författare eller publikationer.

Cociteringsanalysen av författare introducerades 1981 av Howard D. White och Belver C Griffiths (White & Griffiths 1981). Författarna tillämpar den metod som tidigare använts för dokument och som tidigare grundats bl.a. av H. Small (1973). Den grundläggande principen för cocitering av dokument är, enligt White Griffiths (ibid.), att ju fler gånger två dokument citeras tillsammans, ju närmare står de varandra sett ur de citerande författarna synvinkel. Om man representerar dokumenten med en punkt och antalet cociteringar som det inverterade avståndet kan man producera kartor. Kartorna visar relationer mellan olika par av dokument.

Det som skiljer författarnas ansats från tidigare cociteringsanalyser är att de istället för dokument analyserar cociteringar av en mängd dokument representerade av ett antal författare. Det som analyseras är inte författaren som individ utan den samlade produktion av dokument som kan associeras till en författare (franskans *oeuvre*) (ibid.).

Resultatet av datainsamlingen kan presenteras och visualiseras genom att författare eller dokument eller vad det nu är man analyseras, inordnas i grupper genom att någon form av klustringsprocedur används. Som grund för gruppindelningen används då någon norm för att eliminera skalproblem samt för att bestämma likhet respektive olikhet mellan de analyserade enheterna.

Ett sätt att visualisera resultaten är att presentera kartor genom *Multi Dimensional Scaling (MDS)*. Nedan följer en redogörelse för begreppen rådata och råmatris, normering och likhetsmått samt klusteranalys och *Multi Dimensional Scaling (MDS)*. Kännedom om begreppen är centrala för förståelse av cociteringsanalysen som verktyg.

2.3.1.2. Rådata och Råmatris

För datainsamling vid cociteringsanalys används citeringsdatabaser d.v.s. databaser där citeringsdata från innehållsförteckningar insamlats och strukturerats. En närmare presentation av citeringsdatabaser finns nedan i avsnitt 4.1. Citeringsdatabasen tillhandahåller kvantitativa rådata av den eller de parametrar som ska studeras, exempelvis de 100 mest citerade författarna och vilka som citerar dem.

Efter datainsamlingen redovisas antalet cociteringar i en matris (råmatris).

Nedan ses en råmatris som visar cocitering av tre författare (eller dokument) A, B, och C.

	A	B	C
A	0	2	7
B	2	0	10
C	7	10	0

Av matrisen, som har formatet 3 x 3, går att utlösa att författare A och B cociteras i 2 publikationer; A och C i 7 st. samt C och B i 10 publikationer.

Att presentera cociteringar i en råmatris innehållande råpoäng medför vissa svårigheter, exempelvis skalproblem. Någon form av normering (standardisering) med avseende på var gränserna mellan lika fall och olika fall ska dras av värdena är nödvändig inför den fortsatta behandlingen av data.

2.3.1.3. Likhetsmått

Kvantitativ uppskattning av likhet bygger på tanken att likheten och olikheten mellan observationerna avgörs av avståndet mellan observationerna så som dessa placerats i ett koordinatsystem. (Aldenderfer & Blashfield 1984). Koordinatsystemets dimensioner begränsas inte till två (x och y) utan bestäms av antalet variabler som mäts.

Så vitt jag kan förstå är den bild av ett koordinat system som författarna använder en adekvat bild av problemet kring begreppen likhet och olikhet. Då en observation av en kort och lätt individ och en av en lång och tung placeras i ett koordinat-system kan graden av olikhet såväl som gränserna mellan likartade och olikartade beskrivas och avgöras med ett värde.

Likhetsmått har betydelse vid tillämpningen av klusteranalys och multidimensionell skalning som är metoder att klassificera och presentera observationer. Metoderna kommer att översiktligt beskrivas nedan i avsnitt 2.3.1.3. och 2.3.1.4.

Enligt (Aldenderfer & Blashfield 1984) används olika likhetsmått inom samhälls- och humanvetenskaperna. De vanligaste är

- Korrelationskoefficienter
- Distansmått
- Associationskoefficienter

Valet av vilken normeringsskala man använder är inte oväsentligt. Aldenderfer & Blashfield menar att valet av likhetskoefficient kan påverka resultatet av en följande klusteranalys. Valet av likhetsmått är en central fråga när forskningen ska planeras och det är nödvändigt att noga beakta de teoretiska grundvalarna för de olika likhetsmåttin innan valet görs.

Korrelationskoefficienter

Pearson-r produktmomentkorrelationskoefficient (Rudberg 1993)

$$r_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n * s_x * s_y}$$

där

X = variabel X

\bar{X} = medelvärdet av värdet på variabel X

Y = variabel Y

\bar{Y} = medelvärdet av värdet på variabel Y

n = antal observationer

$$s_x = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} \text{ vilket är standardavvikelsen för } x$$

$$s_y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n}} \text{ vilket är standardavvikelsen för } y.$$

r_{XY} anges i ett intervall om -1 till 1 , där -1 och 1 ger full korrelation medan 0 visar att någon korrelation inte föreligger.

Distansmått

Distansmåttin är egentligen ett mått på olikhet.

Det vanligaste måttet är det Euclidiska avståndsmåttet definierat som

$$d_{ij} = \sum_{k=1}^p (x_{ik} - x_{jk})^2$$

där

d_{ij} = avståndet mellan fall i och j (*case i and j*)

x_{ik} = är värdet på den k -onde variabeln för den i -onde fallet (*the value of the k^{th} variable for the i^{th} case*).

Associationskoefficienter

Associationsmåtten bygger på förekomsten av en matris 2 x 2 med följande värden.

$$\begin{array}{cc} 1 & 0 \\ 1 & a \quad b \\ 0 & c \quad d \end{array}$$

1 och 0 är binära värden på likhet, med andra ord likhet och ingen likhet.

a, b, c, d är värden på olika observationer. Placeringen i matrisen visar deras inbördes relation till varandra. I matrisen utgör observation a ena extreman (1,1) medan observation d är den andra (0,0.) Någon likhet föreligger inte mellan a och d . Observationerna b och c förhåller sig däremellan.

En enkel matchning är

$$S = \frac{(a+b)}{(a+b+c+d)}$$

där

a, b, c, d är värden på olika observationer.

och

S är ett mått på likhet mellan observationerna som spänner mellan 1 och 0. Där 1 är fullständig likhet och 0 ingen likhet.

En annan form av associationsmått är Jaccards koefficient

$$S = \frac{a}{(a+b+c)}$$

där

S är ett mått på likhet som spänner mellan 1 och 0.

Det som skiljer Jaccards koefficient från den enkla matchningen är att observationen d som saknar likhet med de övriga inte ska påverka värdet av S

Inom bibliometri och cociteringsanalys kan, som exempel följande tillämpningar av likhetsmått noteras.

White & Griffiths (1981) tillämpar normalisering av rådata genom att använda Pearsons produktmomentkorrelationskoefficient, Pearson- r . Författarna skriver att

(Briefly) the procedure eliminates scale effects due to the relative citedness of documents, or in this case authors and instead measures the degree to which cocitations with documents or authors follow a similar profile or pattern.

McCain (1990) konstaterar att man inom författarcociteringsanalyser ofta använder Pearsons produktmomentkorrelationskoefficient, Pearson-r. McCain menar att bruket av en korrelationsmatris åtminstone har två fördelar. För varje givet par av författare fungerar korrelationsmålet som ett mått på hur ofta det paret författare är cociterade utan också hur lika deras cociteringsprofil är. Två författare som alltid citeras många gånger tillsammans med vissa tredje författare men sällan med andra, kommer att ha en hög positiv korrelation och kan därför sägas vara betraktade som relaterade eller lika varandra i något avseende av de citerande.

Korrelationskoefficienten tar också bort skilnader mellan författare som ofta är citerade och de som är mindre citerade men som har samma profil.

Även andra likhetsmått används inom bibliometrin. Syftet med redogörelsen är inte att i detalj förklara eller granska de använda måtten utan att visa att det inte är givet vilket likhetsmått som ska tillämpas.

Ytterligare exempel på hur likhetsmått används i den bibliometriska kontexten ses i Jarneving (2002) där flera likhetsmått används. I sammanhanget ska påpekas att likhet i den här kontexten uttrycks som styrka, *strength*.

Saltons cosinusmått används som likhetsmått mellan cociterade dokument (*co-citation strength between journals*)

$$S_s(i, j) = \frac{coc(i, j)}{(cit(i) * cit(j))^{1/2}}$$

där

$S_s(i, j)$ = cociteringsstyrkan mellan dokument i och j .

$coc(i, j)$ = antal cociteringar mellan dokument i och j .

$(coc(i) * cit(j))^{1/2}$ = antalet citeringar för i multiplicerat med antalet citeringar för j upphöjt till $1/2$

Saltons cosinusmått används också vid beräkning av cociteringsstyrkan mellan tidskrifter

$$S_s(i, j) = \frac{coc(i, j)}{(cit(i) * cit(j))^{1/2}}$$

där

$S_s(i, j)$ = cociteringsstyrkan mellan tidskrift i och j .

$coc(i, j)$ = antal cociteringar mellan tidskrift i och j .

$(coc(i) * cit(j))^{1/2}$ = antalet dokument som citerar dokument i multiplicerat med antalet dokument som citerar dokument j upphöjt till $1/2$.

Samma mått går också att använd för att mäta cociteringsstyrkan mellan författare

där

$S_s(i, j)$ = cociteringsstyrkan mellan författare i och j .

$coc(i, j)$ = antal cociteringar mellan författare i och j .

$(coc(i) * cit(j))^{1/2}$ = antalet dokument som citerar författare i multiplicerat med antalet dokument som citerar författare j upphöjt till $1/2$.

Likheten vid bibliografisk koppling (*the strength of bibliographic coupling*) bestäms som

$$T_{ij} = \frac{r_{ij}}{\sqrt{n_i X n_j}}$$

där

T_{ij} = kopplingsstyrka mellan dokument i och j .

r_{ij} = antal gemensamma referenser hos i och j .

n_i = antal referenser i referenslistan i dokument i .

n_j = antal referenser i referenslistan i dokument j .

Medelcociteringsstyrka mellan kluster (*mean co-citation strength between clusters*) anges genom

$$\frac{coc(d_1, d_s)}{n_1 X n_2} \quad \text{där}$$

$coc(d_1, d_s)$ = antalet cociteringar mellan kluster d_1 och d_2 .

n_1 = antal dokument i kluster 1

n_2 = antal dokument i kluster 2

Likheten mellan samförekomsten av titelord (*the strength of co-citation*) beräknas (ibid.) med tillämpning av associationsmättet Jaccards koefficient.

$$J_{ij} = \frac{C_{ij}}{(C_i + C_j - C_{ij})}$$

där

J_{ij} = styrkan av samförekomsten mellan ord i och j .

C_{ij} = antalet samförekomster mellan orden i och j .

C_i = frekvensen av ordet i .

C_j = frekvensen av ordet j .

Ovanstående exempel visar den mångfald av likhetsmått som förekommer i litteraturen och att samma likhetsmått kan användas för att normera olika företeelser. Det gemensamma är dock bristen på transparens när det gäller motiven för valet av likhetsmått i det enskilda fallet.

2.3.1.4. Klusteranalys

Klusteranalys är den övergripande beteckningen på en mängd procedurer som kan användas för att skapa en klassifikation. Den grundläggande anledningen till att använda klusteranalys är att hitta grupper av likartade företeelser i ett urval av data. Dessa grupper kallas passande nog för kluster (Aldenderfer & Blashfield 1984).

Klusterproblemet är att bestämma m antal kluster av individer i en mängd I så att I_j tillhör en och endast en delmängd och att de individer som inordnas i samma delmängd är lika samt att individer i andra delmängder är olika (Duran & Odell 1974).

För att lösa klusterproblemet är det önskvärt att bestämma begreppen likhet och olikhet på ett kvantitativt sätt (ibid.).(se avsnitt 2.3.1.3 ovan)

Andra aspekter på klusterproblematiken är hur många kluster som skall skapas.

Enligt Aldenderfer & Blashfield (1984) har det utvecklats sju större grupper av kustringsmetoder.

hierarchical agglomerative

hierarchical divisive

iterative partitioning

density search

factor analysis

clumping

graph theoretics.

I den här undersökningen används klustringsmetoden till att identifiera författare som genom citeringar antas tillhöra samma ämnesområde. I det sammanhanget är det bara klustringsmetoderna *hierarchical agglomerative* och *hierarchical divisive* som är av intresse, de övriga lämnas härmed.

2.3.1.5. Hierarkisk klustring

Hierarchical agglomerative clustering kan översättas som sammanläggande hierarkisk klustring och medan *hierarchical divisive* dess logiska motsats (Aldenderfer & Blashfield 1984) d.v.s. nedbrytande hierarkisk klustring.

Sammanläggande hierarkisk klustring kan beskrivas så att man med utgångspunkt från en mängd observerade enskilda fall bildar överordnade grupper (kluster) med utgångspunkt i de likhetsmått som man valt att tillämpa. Lika fall hänförs således till ett kluster. De överordnade grupperna bildar i sin tur nya överordnade grupper. Ju högre upp man kommer i hierarkin desto färre grupper får man och ju mindre liket finns mellan de underordnade grupperna i hierarkin. Slutresultatet är en grupp som omfattar samtliga observerade fall.

Nedbrytande hierarkisk klustring innebär då å andra sidan att man utifrån ett kluster där de ingående fallen har liten likhet successivt och bildar underordnade grupper med ökande likhet.

Att i detalj redogöra för hur en klustring teoretiskt går till faller utanför ramen för den här studien. Den klustringsalgoritm som är *hierarchical divisive* och som används i BIBEXCEL och därmed i den här studien, kommer att presenteras närmare i avsnitt 4.2.

2.3.1.6. Multi Dimensional Scaling

För att presentera resultaten av en cociteringsanalys kan multidimensionell skalning (MDS, *Multi Dimensional Scaling*) användas. MDS är en matematisk - statistisk metod att visualisera vissa typer av relationer spatialt. De relationer som visas är närhetsrelationer beräknade som likhetsmått (*proximity*). Relationerna visualiseras ofta i en karta.

Utgångspunkten för MDS är en matris där graden av närhet mellan objekten i matrisen åsatts ett numeriskt värde (jmf. avsnitt 2.3.1.2.ovan.) I sammanhanget ska nämnas att MDS är möjligt även om värdena i matrisen inte är närhetsvärden. Det krävs då någon matematisk operation för att omvandla värdena till mått på likhet eller olikhet, exempelvis genom att beräkna korrelationer. Inom bibliometrin kan värdet vara hur många gånger två författare citeras tillsammans (cociteringsanalys av författare) eller hur många gånger två artiklar citeras tillsammans. (cociteringsanalys av artiklar)

Analogin med en karta kan vara vilseledande eftersom begreppet karta kan tolkas som två-dimensionellt. MDS-kartorna begränsas inte av två dimensioner utan kan representera fler dimensioner. Genom den grafiska presentationen går det att utläsa ”dolda strukturer” och relationer mellan de presenterade objekten.

Kruskal & Wish (1978) redogör för följande, pedagogiska, exempel. Antag att man ska upprätta en tabell över avståndet mellan olika städer. Detta kan göras genom att mäta avståndet mellan exempelvis New York - Los Angeles, New York- Miami, New York – Boston. Om avståndsmätningen sker mellan många städer kan dessa avstånd ordnas i en matris. Den här typen av matriser förekommer ofta i vägatlaser. Kartan föregår i detta fall matrisen. MDS innebär att man har en matris och utifrån den ska skapa en karta.

Inom ramen för det här arbetet är det inte möjligt att i detalj redogöra för de komplexa matematiska och statistiska metoder som möjliggör MDS. De beräkningar som krävs och den grafiska framställningen görs med datorer.

Den grundläggande tanken är att två företeelser som karaktäriseras som närstående (mer lika) grafiskt placeras närmare varandra medan företeelser som inte är närstående (är olika) placeras långt ifrån varandra. Mer formellt kan det uttryckas på följande sätt (Jarneving 2002, min översättning)

Låt A, B, C, D representera fyra objekt i en matris. Låt X vara ett värde på likheten mellan A och B. Låt Y vara ett värde på likheten mellan C och D.

- 1) Om $X = Y$, så ska avståndet mellan de båda punkterna i konfigurationen som representerar A och B, vara samma som avståndet på de punkter som representerar C och D.
- 2) Om $X < Y$, så ska avståndet mellan de båda punkterna i konfigurationen som representerar A och B, vara större än avståndet på de punkter som representerar C och D.

3. TIDIGARE FORSKNING

Biblioteks och informationsvetenskapen har varit föremål för bibliometriska studier. Infallsvinklarna och metoderna har varierat; såväl hela ämnesområdet som enskildheter såsom enskilda tidskrifter, forskningsområden, land har studerats. Sammantaget ger studierna en bild över ämnet

Studier av tidskrifter har företagits av Bonnevie (2003). Bonnevie har studerat *Journal of Information Science (JIS)* utifrån ett flertal aspekter och metoder. Studien omfattar tiden från 1979 till 2001. Tidskriften studerades utifrån synlighet d.v.s. tidskriftens förekomst i databaser, en studie av författarskap innefattande bl.a. en analys med tillämpning av Lotkas lag, studier av självciteringsmönster, innefattande ”*self-citing*” och ”*self-cited*”. Vidare studerades hur internationell tidskriften var med hänsyn tagen till författarnas härkomst och varifrån tidskriften citeras. Studien innefattade också en analys av tidskriftens grad av inflytande, *scientific impact* genom cociteringar och antal citeringar i andra tidskrifter. I korthet visar studien vad anbelangar tidskriftens synlighet att *JIS* förekommer i 24 databaser och förekommer flest gånger i LISA och i SSCI. Författarproduktiviteten överensstämmer i stort sett med Lotkas lag. Något fler författare skriver dock fler artiklar än vad som kunde förväntats, vilket författaren tolkat som att tidskriften har ett antal trogna författare. Omfattningen av *self-citing* överensstämmer med genomsnittet medan *self-cited* nivån uppfattas som hög. Bonnevie tolkar detta som att tidskriften kan ha ett något mindre inflytande inom det vetenskapliga kollektivet eller att tidskriften har en annan inriktning än andra tidskrifter inom ämnesområdet. Bonnevie konstaterar vidare att de flesta författarna och citeringarna kommer från Europa. 40 % av citeringarna kommer dock ifrån nordamerikanska tidskrifter vilket tolkas som att tidskriften används internationellt. Studien visar också att *JIS* cociteras med sådana högt ansedda tidskrifter som *Journal of documentation*, *Journal of the American Society for Information Science and Technology*, (*JASIST*) samt *Scientometrics*.

Wormell (1998) studerar med hjälp av infometrisk metod sju LIS- tidskrifter med syfte att bestämma hur stort inflytande dessa har inom ämnesområdet. I syfte att få en mer belysande bild av tidskrifternas genomslagskraft än den som *Journal Impact Factor* ger, studerades tidskrifterna med avseende på de kombinerade relationerna mellan lokaliteten av författare, läsare och prenumeranter. Som ett annat mått på genomslagskraft föreslås i vilken utsträckning tidskriften påverkar ämnen utanför de traditionella ämnena som tidskriften representerar s.k. *knowledge export*. I artikeln redovisas undersökningen explicit för tidskrifterna *LIBRI* och *JASIS*.

En cociteringsanalys av tidskrifter inom *Information Retrieval (IR)* området gjordes av Ding et al (2000) Studien omfattade tiden 1987-1997, uppdelade på hela tidsintervallet samt på perioderna 1987-1991 och 1992-1997. Studien utfördes utifrån två synvinklar. Den ena synvinkeln var från tidskrifter i olika ämnesområden som ofta citerades av IR-forskare. Den andra synvinkeln var från LIS-tidskrifter som ofta citerades av IR-forskare. Resultatet korrelerades genom bruk av *MDS*. Studien jämfördes med en cociteringsstudie av författare inom IR-området som tidigare utförts av Ding (1998).

Denna komparativa studie visade att kartorna producerade med utgångspunkt i cocitering av tidskrifter hade stora likheter med de kartor som producerades utifrån cociteringsanalys av

författare. Studien visade också att IR som ämne var stabilt under den studerade perioden i relation till ämnets sammansättning och dess relation till omgivande ämnen. Att IR är ett tvärvetenskapligt ämne kunde klart konstateras. Cociteringskartorna över LIS-tidskrifter visar att IR har en stark koppling till datologi och att IR-teori haft stor betydelse för att hantera kemisk, medicinsk och rättslig information.

White & McCain (1998) studerar LIS med utgångspunkt att visa vad cociteringsanalys kan användas till. Studien, som har karaktären av metodpresentation, omfattar tidsperioden 1972 - 1995 och anger värden för intervallen 1972-1979, 1980-1987, 1988-1998 samt för hela perioden 1972- 1995, vilket medför att trender kan urskiljas. Författarna anser att cociteringsanalysen kan användas till att lista författare, ange författarnas specialiteter, bestämma författarnas anseende, påvisa förändringar i författarens påverkan, åskådliggöra relationer mellan enskilda författare och grupper av författare, förändringar i författarens "karriär" samt bestämma kanoniska författare. De kluster av författare som cociteringsanalysen genererar bestäms till sitt innehåll av författarna själva på grundval av deras personliga kunskaper om ämnet.

Tidskriften *Journal of the American Society for information Science (JASIS, numera Journal of the American Society for Information Science and Technology, JASIST)* studerades av Persson (1994). Studien gjordes genom en analys av artiklar publicerade under tidsperioden 1986 – 1990. Författaren gjorde en cociteringsanalys av författarna och en på cociteringar baserad analys av bibliografisk koppling. Syftet med studien var att utreda vilka forskningsfronter som kommit till uttryck genom artiklarna i JASIS. Resultatet redovisades i form av 5 kluster av författare och artiklarnas titlar visualiserade i kartor skapade genom *MDS*. De kluster av författare som utredningen presenterade bestämdes till sitt innehåll av titlarna på 51 artiklar.

Särskilt White & McCain och Perssons undersökningar visar att metoden att bestämma ett ämnes ämnesmässiga innehåll inte är given. Man kan rikta kritik mot att deras metoder grundar sig dels på utredarnas subjektiva uppfattning dels på att titlarna lämnar alltför lite information. Däremot saknas studier där det i praktiken viktigaste instrumentet för informations- och dokumentsökning, nämligen indexeringstermer/deskriptorer, används för att bestämma det kognitiva innehållet i ett ämne. Förevarande utredning är ett försök att fylla den luckan och visa på en mer objektiv och uttömmande metod.

4. METOD

De data som ligger till grund för den här undersökningen är inhämtade från en citeringsdatabas nämligen *Social Science Citation Index (SSCI)*. Avsnitt 4.1 innehåller en presentation av citeringsdatabasen som fenomen.

4.1. Citeringsdatabaser

Mot bakgrund av vad som sagts ovan i avsnitt 2.1.1 och 2.1.2, angående referensers och citeringars betydelse inom den vetenskapliga kommunikationen, samt publikationens förhållande till den intellektuella basen och forskningsfronten, är det ett rimligt antagande att kunskap om hur publikationer och författare citeras är viktig för kunskapsanvändarna. Detta

har uppmärksamats och därför tillhandahålls på marknaden tillgång till olika citeringsdatabaser som omfattar större eller mindre delar av vetenskapliga discipliner. Representationen av ett ämne i en citeringsdatabas, *Social Science Citation Index (SSCI)*, är föremålet för denna studie. Nedan följer därför en beskrivning av citeringsdatabasen som fenomen och vilka data som kan återfinnas där.

Exempel på citeringsdatabaser är *Science Citation Index (SCI)*, *Social Science Citation Index (SSCI)* och *Arts and Humanities Citation Index (AHCI)*.

Science Citation Index (grundat 1963) innehåller uppgifter om referenser till naturvetenskaplig, medicinsk och teknisk litteratur, *Social Science Citation Index* (grundat 1972) bevakar samhällsvetenskaplig litteratur och *Arts and Humanities Citation Index 1980* humanistisk litteratur (Kärkki & Kortelainen 1998)

Citeringsdatabasen *Social Science Citation Index (SSCI)* produceras av *Institute of Scientific Information (ISI)* men administreras och tillhandahålls genom databasvärden The Thomson Corporation. Databasen tillhandahålls (<http://www.isinet.com/products/citations/ssci>) på marknaden i ett antal olika format, genom internet eller intranet eller genom den webbaserade *Web of Science*, On-line via *Social SciSearch*, en *SSCI* finns också som CD-ROM. *SSCI* distribueras också av databasvärdar som DIALOG, DIMI och STI. (<http://www.isinet.com/products/citation/ssci> 2004-08-04).

Av licensskäl har jag inte haft tillgång till *Social Science Citation Index* på annat sätt än genom *Web of Science*. Det kan inte uteslutas att de olika formaten tillhandahåller olika typer av sökmöjligheter och sorteringsfunktioner. Principerna för databasernas uppbyggnad torde dock vara den samma.

Nedan följer en beskrivning av *SSCI* så som det tillhandahålls och beskrivs i *Web of Science*. (http://wos3.isiknowledge.com/help/h_database.htm. 2004-07-29)

“The Social Science Citation Index is a multidisciplinary index to the journal literature of the social sciences. It fully indexes more than 1,725 journals across 50 social sciences disciplines, and it indexes individually selected, relevant items from over 3,300 of the worlds leading scientific and technical journals.

In addition, the Social Science Citation Index:

- *Provides access to current information and retrospective data from 1956 forward. Note that you may not have access to all data years.*
- *Averages 2,900 new records per week.*
- *Includes approximately 60,000 new cited references per week.*
- *As of January 1992, contains searchable full length, English-language author abstracts for approximately 60 % of the articles in the indexes”*

Information Science & Library Science är ett av de ämnen som täcks av *Social Science Citation Index*.

I indexet presenteras en post ”full record” (<http://wos3.isiknowledge.com/CIW.cgi>) med uppgifter om

- Titel
- Författare
- Källa (*Source*)
- Dokumenttyp
- Citerade referenser (*Cited References*)
- Citeringar (*Times Cited*)
- *Find Related Records*
- Språk
- Adresser (innefattande författarens organisationstillhörighet, *affiliation*)
- Utgivare
- Ämnes kategori
- IDS –nummer
- ISSN

För sökning i databasen tillhandahålls tre sökalternativ. Avancerad sökning , *General search* och *Cited Referens Search*.

Avancerad sökning görs genom formulering av *queries* (sökfrågor) som byggs upp av s.k. *fieldtags* och Boolska operatorer.

De sökbara *fieldtags* som kan användas är :

TS = Topics, TI = Title, AU= Author, GP= Group Author, SO= Source, AD=Adress, OG= Organization, SG= Sub organization, SA= Street Address, CI=City, PS= Provins /State CU=Country ZP= Zip Code /Postal code

Sökningarna kan kombineras genom de Boolska operatorerna *AND, OR, NOT, SAME*.

Cited reference med ger sökning på *Cited author, Cited Work* samt *Cited Years*.
Kärnan i citeringsdatabasen är således uppgifterna om citeringar.

Cited references visar de bibliografiska referenserna i den referenslista som ingår i artikeln. Eftersom alla referenser i de artiklar som indexerats finns registrerade i databasen, är det möjligt att ta fram vilka författare och artiklar som citerar den ifrågakvarande artikeln. Detta sker genom *Citing Articles*.

Citeringsdatabasernas betydelse vid bibliometriska undersökningar gör det nödvändigt att undersöka de urvalsprinciper som används när tidskrifter väljs ut för att indexerats i databasen. Enligt ISI (<http://www.isinet.com.lib.costello.pub.hb.se/essays/selectionofmaterialforcoverage/199701.html/> 2004-06-15) ligger följande principer till grund för urvalet.

ISI begränsar sig till att bevaka kärntidskrifter och inte samtliga tidskrifter inom ett visst område. Att bevaka samtliga tidskrifter vore oekonomiskt och opraktiskt. Begreppet kärntidskrifter bygger på förekomsten av den s.k. Bradfordfordelningen, formulerad i Bradfords lag. (Fördelningen är ett mått på koncentration och spridning. Förenklat innebär denna lag att en liten mängd tidskrifter (s.k. kärntidskrifter) innehåller en betydande mängd av artiklarna inom ett bestämt ämnesområde medan resten av artiklarna sprider sig till ett stort antal tidskrifter, se avsnitt 2.3 ovan).

Tidskriften måste vidare, enligt ISI, komma ut i enlighet med fastlagd utgivningsplan. ISI accepterar inte att tidskriften inte följer angiven utgivningsplan. ISI menar att förmågan att komma ut vid rätt tid visar att det ligger artiklar i lager för publicering och därmed en önskan från många författare att bli publicerade i tidskriften. Detta visar i sin tur på att tidskriften har ett högt anseende bland forskarna.

Det är viktigt att tidskriften följer vedertagna publiceringskonventioner, vilket gör det möjligt att återvinna källdokumentet. Titlar på artiklar, *abstracts* och *keywords* måste finnas på engelska. ISI anser också att förekomst av *peer-reviewing* process (att inlämnat material fackgranskas av experter) är viktigt, vilket visar på kvalitet på tidskriften och i den forskning som publiceras i den. ISI strävar efter att täcka tidskrifter med en geografisk mångfald av såväl prenumeranter, författare och citerade artiklar.

Citationsdata används för att avgöra hur inflytelserik och använd tidskriften är. Hänsyn tas till *impact factor*, antal citeringar och *immediacy index*. Begreppen *impact factor* och *immediacy index* har förklarats ovan i avsnitt 2.3.

Även ett sådant mått som antal självciteringar (att artiklar i tidskriften citerar andra artiklar i samma tidskrift) tas i beaktande, eftersom det antas att en hög andel självciteringar visar att ett forskningsfält är litet eller isolerat, även om detta varierar mellan olika forskningsområden.

De citeringsdata, som den här undersökningen har baserats på, har inhämtats från CD-ROM - utgåvan av *Social Science Citation Index*. Urvalet består av 22,973 bibliografiska poster inom ämnesområdet *Library and Information science*, publicerade 1995-2000, och fördelade över 54, av totalt 55, tidskrifter listade i 2002 års upplaga av *Social Science - utgåvan av Journal Citation Reports (JCR)*. Urvalet är nedladdat från CD-ROM - utgåvan av *Social Science Citation Index*. Därur har först de 100 mest citerade författarna inhämtats.

Författarna har sedan inordnats i en råmatris, 100 x 100, varefter råmatrisen har normaliserats med Saltons cosinusformel (se ovan, avsnitt 2.3.1.2.)

Efter det att citeringsdata inhämtats och ordnats i råmatris, har dessa sedan bearbetats med det bibliometriska dataprogrammet BIBEXCEL, varvid författarna har ordnats i kluster med hjälp av BIBEXCELS klusteranalysalgoritm.

4.2. Bibexcel

Idén bakom BIBEXCEL beskrivs i programmets hjälppil på följande sätt:

“Bibexcel is designed as a tool box for manipulating bibliographic data. The result of your manipulations will be saved in files that can be opened with Excel or any other software reading text-files tabbed into columns. Bibexcel let's you combine information from several fields of a document record, count frequencies, co-occurrences and shared units (bibliographic coupling). Among other things there is also a procedure for finding citation links among the documents within a given set. Above all, the tools can be combined - the result of using them depends far more on your own imagination than the tools themselves.”

Den klustringsalgoritm som programmet tillämpar och som kan beskrivas som *hierarchical divisive* (se avsnitt 2.3.1.4) beskrivs metaforiskt men pedagogiskt på följande sätt i hjälppilen.

”A quite nice home-made routine for clustering of pairs, sometimes called Persson's Party Clustering.

Imagine you have the following list of pairs:

10	A	B
9	D	F
8	B	C
7	A	C
6	F	G
5	H	I
4	A	H

Note that the pairs are sorted by a co-occurrence frequency in column 1, and that the pair-list also must have tabs separating the two units of a pair.

*The clustering routine will have the following sequence of events:
(the pairs are invited to a party):*

A-B comes first, have to wait in hall

D-F comes next, have to wait in hall.

B-C comes next, forms a cluster-table with A-B in the room

A-C comes next, will be deleted since A-C already in room

F-G comes next, will not find a friend in room

Goes to the hall and finds D-F and

Then D-F-G will form a cluster in the room

H-I comes next, have to wait in hall

A-H comes next, H will cluster with A-B-C

then searches the hall and finds H-I

I will cluster with A-B-C-H

Cluster 1 will hold: A-B-C-H-I

Cluster 2 will hold D-F-G”

BIBEXCEL kan inte representera mer än 75 författare vid MDS, någon karta över ämnesområdet kan därför inte presenteras. I stället redovisas de kluster som BIBEXCEL genererar. En MDS –karta hade varit att föredra eftersom kartan kan visa dolda relationer mellan olika författare och kluster. En ren presentation av klustren ger likväl en väl beskrivande bild av ämnets kognitiva innehåll.

4.3. LISAs thesaurus och deskriptorer

De författarnamn som ingår i klustren ger inga indikationer på vilka ämnen som författarna och därmed klustren representerar. I syfte att bestämma klustrens ämnesmässiga innehåll har bibliografiska poster avseende de i klustren ingående författarna hämtats från databasen *LISA: Library and Information Sciences* som tillhandahålls av databasvärden *Cambridge Scientific Abstracts*. De bibliografiska posterna innehåller indexeringstermer som beskriver dokumentets ämnesmässiga innehåll. Deskriptorer som tilldelats dokumenten i LISA har tilldelats dokumenten manuellt av indexerarna med ledning av LISAs thesaurus.

Bibliografiska poster har hämtats från 1969 till 2000, totalt 2254 bibliografiska poster. För varje författare har de deskriptorer (DE) som förekommer i posten sedan brutits ut, bearbetats och räknats i kalkylprogrammet Excel. Därefter har indexeringstermerna för varje kluster ordnats efter frekvens. Slutligen har indexeringstermerna tolkats för att få en uppfattning om de enskilda klustrens ämnesmässiga innehåll. Ett visst mått av subjektivitet, dock grundad på min förförståelse av LIS, har därför inte gått att utesluta.

5. RESULTAT

Efter datainsamling och databehandling av data har följande framkommit.

Resultaten redovisas i form av kluster av författare. Antalet författare i klustren redovisas samt namnen på de ingående författarna. Vid varje författare anges antalet kopplingar till författaren som BIBEXCEL funnit. Antalet bibliografiska poster och de indexeringstermer som tilldelats de bibliografiska posterna anges därefter. I de fall det inte har varit möjligt att ta fram bibliografiska poster för någon författare i klustret redovisas detta. Det kan bero på att författarens namnform har kunnat bestämmas men inte förekommit i LISA. Andra fall har namnformen inte kunnat bestämmas med säkerhet då olika förnamns- och mellannamnformer har påträffats i LISA. Namnen på dessa författare redovisas. Avslutningsvis görs en tolkning i klartext av klustrens ämnesmässiga innehåll med utgångspunkt i de redovisade indexeringstermerna.

Kluster 1

9 författare

Ives, B. 3; Davis, F.D. 3; Doll, W.J. 2; Delone, W.H. 2; Nunnally, J.C. 2;
Igbaria, M. 1; Cohen, J. 1; Lucas, H.C. 1; Bandura, A.1.

3 författare har inte återfunnits i LISA: Bandura, A.; Delone, W.H.; Nunnally, J.C.
En namnform har inte kunnat bestämmas: Cohen, J

24 bibliografiska poster.

Termer:

Computer applications 6, Business management 4, Information technology 3,
Information work 3, Services 3, User services 3, Book review abstracts 2,
Evaluation 2, Planning 2, Use 2, User satisfaction 2, Virtual workplace 2.

Tolkning:

Klustret innehåller relativt få dokument och därmed indexeringstermer. Författarna synes publicera sig i inriktning på praktiska tillämpningar av informationsteknologi, med särskild inriktning på företagsledning. En viss betoning på användarna kan utläsas.

Kluster 2

3 författare

Schubert, A. 2; Glanzel, W. 2; Eur com. 1; Braun, T. 1

Eur com = Europeiska Unionens kommission har uteslutits eftersom kommissionen inte är en författare i egentlig mening och därför inte kan representera något *oeuvre*.

89 bibliografiska poster

Termer:

Bibliometrics 67; Scientometrics 48; Science 47; Library materials 44; Citation analysis 34; Periodicals 9; Articles 17; Science and technology 14; Physics 13; Productivity 12; Evaluation 11; Co authorship 9; Indicators 9; Information work 9; International co authorship 9; International conferences 9; Chemistry 8; Countries 8; International comparisons 8; Research 8; Social sciences 8; Services 7; Models 6; Publication output 6; Gatekeepers 5; National productivity 5;

Tolkning

Författarna inom klustret ägnar sig åt bibliometri och scientometri sett ur ett forskningssociologiskt och forskningskommunikativt perspektiv. De företeelser som studeras synes vara olika författarskap, publikationer, publicering inom olika ämnesområden (domäner) och inom olika länder samt produktivitet och publiceringsmönster och samarbete mellan olika länders forskare.

Kluster 3

5 författare

Dervin, B. 3; Hernon, P. 2; Kuhlthau, C.C. 1; McClure, C.R. 1; Taylor, R.R. 1.

314 bibliografiska poster.

Termer:

USA 99; Information work 60; Book review abstracts 57; Government publications 45; Services 44; Federal government 31; Library materials 31; User training 30; Research 26; Evaluation 23; User services 23; Management 19; Depository libraries 19; Academic libraries 18; Computerised information retrieval 17; Surveys 17; Information storage and retrieval 15;

Central government 14; Libraries 14; Online information retrieval 14; Organization 14;

Library management 13; Subject indexing 13; Access to information 12; Federal government information 12; Information seeking behaviour 12; Public libraries 12; Technical services 12; Reference work 11; Use 11; Computerized information storage and retrieval 10; Internet 10; National policies 10 ;

Government information 8; Information communication 8; Performance measures 8; Searching 8; Acquisitions 7; Networks 7; Reference work 7; Federal information policies in the 1990s 6; Law 6; Librarianship 6; Library and information science 6; Policies 6; Publishing 6; Seeking meaning 6; University libraries 6; User needs 6; Teaching the library research process 6; Information services 5; Library and Information science publications 5; Library technology 5; New Zealand 5; Research methods 5; Research misconduct 5; Selection 5; Social sciences 5

Tolkning:

Klustrets författare ägnar sig åt s.k. soft information retrieval dvs. informationsåtervinning sett ur ett psykologiskt och pedagogiskt perspektiv där användaren står i centrum för studierna (User training 30; Evaluation 23; User services 23; Information seeking behaviour 12; Seeking meaning 6; User needs 6)

Även dokument som innehåller studier av offentliga myndigheters tillhandahållande och spridning av information, främst då i USA.

Kluster 4

21 författare

Bates, M.J. 4; Ingwersen, P. 4; Schamber, L. 3; Buckland, M.K. 3; Saracevic, T. 3; Marchionini, G. 3; Belkin, N. J. 2; Fidel, R. 2; Gorman, M. 2; Harter, S.P. 2; Shneiderman, B. 2; Lancaster, F.W. 1; Line, M.B. 1; Ellis, D. 1; Nielsen, J. 1; Crawford, W. 1; Cronin, B. 1; Borgman, C.L. 1; Spink, A. 1; Tenopir, C. 1; Wilson, P. 1.

En författare har en namnform som inte kunnat bestämmas: Nielsen, J;

1 487 bibliografiska poster.

Termer:

Searching 272; Online information retrieval 250; Subject indexing 247; Information storage and retrieval 242; Computerized information storage and retrieval 232; Computerised information retrieval 172; Book review abstracts 136; User services 101; Services 92; Information services 78; Professional education 74; Research 67; Librarianship 65; Evaluation 65; Databases 58; Computerized subject indexing 55; Online databases 54; Periodicals 50;

Citation analysis 29; Library schools 29; User interface 29; External magnetic tape information services 27; User behaviour 26; Electronic media 25;

Hypertext 24; Electronic publishing23; Information seeking behaviour 23; Library management 23; Profession 23; World Wide Web 23; Reference work22; User training21; State of the art reviews20; User surveys21; Australia20; Data processing20; End users20; Eyewitness Encyclopaedia of Nature 20; Eyewitness Encyclopaedia of Science 20; Medicine20;

Articles19; CD-ROMs19; Intermediaries19; Microcomputers19; Developing countries18; Library and information science periodicals18; Library technology 18; Social sciences18; Acquisitions17; Full text searching17; If you want to evaluate your library 17; Interfaces17; Public libraries17; Selection17; UK17; Cooperation16; Innovations16; Libraries16; CD-ROM databases15; Cataloguing 15; Search strategies15;

History14; Information industry14; Library staff14; Personal computers 14; Promotion 14; Scholarly publications14; Full text databases13; Libraries and the future Libraries and the future13; Optical discs13; Publishing13; Storage13; Theories13; Vendors13;

Academic libraries12; Children12; Computerized information work 12; Digital libraries12; Discs 12; Internet12; Library and information science12; Models12; Published materials12; Acknowledgements11; Compact discs11; Relevance11; Science11; Telecommunications 11; Choice of terms10; Feedback10; Library students10; Policies10;

Tolkning:

Klustret innehåller det största antalet dokument. Författarna ägnar sig åt *soft information retrieval*, d.v.s. informationsåtervinning satt i en kognitiv kontext.

Även frågor kring kunskapsorganisation finns inom detta kluster.

Aspekten studeras också utifrån användarnas perspektiv såväl utifrån ”slutanvändarnas perspektiv som mellanhändernas (informationsspecialisternas perspektiv). Inom klustret behandlas också LIS som ämne, informationsspecialisters professionella utveckling och biblioteksutbildningarnas utveckling.

Kluster 5

9 författare.

Robertson, S.E. 4; Salton, G. 4; Vanrijsbergen, C.J. 2; Harman, D. 2;
Lindberg, D.A.B. 2; Jones, K.S. 1; Croft, W.B. 1; Cooper, W.S. 1; Cimino, J.J. 1

331 bibliografiska poster

Termer:

Information work 167; Subject indexing 126; Information storage and retrieval 121;
Searching 119; Technical services 115; Evaluation 57; Computerised information retrieval 42;
Online information retrieval 38; Computerized information storage and retrieval 36; Terms 32;
Computerized subject indexing 30;

Relevance 25; Retrieval performance measures 25; Automatic subject indexing 24 ;
Performance measures 23; Mathematical models 22; Services 20; User services 20; Models

18; Research 18; Automatic indexing 16; Strategies 16; Weighting 16; Conferences 13; Clustering 11; Information services 10; SMART Project 9;

Artificial intelligence 8; Associative indexing 8; Computerized information work 8; Computers 8; Experimental retrieval collections 8; Medicine 8; Ranking 8; Statistical techniques 8; Boolean strategies 7; Future developments 7; Information retrieval 7; Information science 6; Library staff 6; Online cataloguing 6; Expert systems 5; Feedback 5; Information professionals 5; Interactive systems 5; Output 5;

Tolkning:

Inriktningen i kluster 5 är mot inriktning på matematiska modeller, statistik och algoritmisk IR, s.k. *hard IR*. Olika aspekter på indexering såsom automatisk indexering, klustring, associativ indexering studeras också. Vidare studeras olika sökstrategier samt vägning och rankning av söktermer och återvunna dokument.

Kluster 6

11 författare.

Callon, M. 4; Small, H. 3; White, H.D. 2; Latour, B. 2 ; Leydesdorff, L. 2; Garfield, E 2; Merton, R.K 1; Moed, H.F. 1; Narin, F.1; McCain, K.W. 1; Kuhn, T.S. 1

En författare saknas i LISA: Kuhn, T.S.

300 bibliografiska poster.

Termer:

Bibliometrics 66; Citation analysis 61; Library materials 49; Periodicals 48; Information work 31; Articles 20; Science 24; Cocitation 23; Scientometrics 23; Science Citation Index 22; User services 19; Computerized information storage and retrieval 18; Services 18; USA 18; Computerised information retrieval 17; Citation indexes 17; Information services 17; Subject indexing 16; Technical services 16;

Librarianship 15; Book review abstracts 13 ; Information storage and retrieval 13; Online information retrieval 13; Software 13; Clustering 12; Manual systems 10; Medicine 10; Optical character recognition 14; Research 12; Science and technology 12;

Databases Evaluation 9; Library staff 9; Patents 8 ; Chemistry 7; Citations 7; Conferences 7; Library management 7; Searching 7; State of the art reviews 7; Computerized information services 6; Profession 6; Publishing 6; Reference work 6; Selection 6; Computerized information work 5; Impact factors 5; Librarians 5; Management 5;

Tolkning:

Författarna inom klustret ägnar sig åt bibliometri, scientometri, och citeringsdata. Inom klustret finns bl. a. författare som ägnat sig åt metodutveckling av cociteringsanalyser och som citerats i den här studien. Inom klustret återfinns bibliometrins klassiska metodutvecklare och citeringsanalysens uppfinnare.

Kluster 7

9 författare.

Orlikowski ,W.J. 5; Markus, M.L. 2; Foucault, M. 2; Giddens, A. 2; Kling, R.1
Checkland, P.1; Budd, J.M. 1; Robey, D. 1; Zuboff, S. 1.

Två författare har inte kunnat återfinnas i LISA: Foucault, M; Zuboff, S.

93 bibliografiska poster.

Termer:

Book review abstracts 13; Computer applications 9; Surveys 9; Technical services 9;
Acquisitions 8; Articles 7; Citation analysis 7; Librarianship 7; Bibliometrics 6;
Periodicals 6; Academic libraries 5; Faculty 5; Library materials 5; Selection 5;
Social aspects 5;

Communications technology 4; Finance 4 ; Information communication 4; Information
technology 4; Models 4; Professional education 4; Searching 4; Universities 4;
Computer science 3; Digital libraries 3; Library and information science periodicals 3;
Research 3; Social informatics 3; Use 3; Design 3;

Context 2; Cultural aspects 2 ; Electronic mail 2; Electronic publishing 2; Information
retrieval 2; Information, systems and information systems 2; Library staff 2;
Management information systems 2; Organizational culture 2; Publishing activities 2
Systems development 2; Theories 2; USA 2 ; User services 2; Voice mail 2

Tolkning:

Flertalet av indexeringstermerna är vanligt förekommande i de övriga klustren och därmed inte speciellt särskiljande. Dock förekommer termer som hänvisar till sociala och kulturella aspekter, organisationskultur etc.

Kluster 8

4 författare.

Davenport, T.H. 2; Nonaka, I. 2; Hammer, M. 1; Drucker, P.F. 1.

En författare har inte kunnat påträffas i LISA: Nonaka, I.

13 bibliografiska poster.

Termer:

Management 6; Organization 6; Information work 5; Library management 5;
Book review abstracts 2; Knowledge management 2;

Tolkning:

Även kluster 8 innehåller termerna *management*, *organization*.

Därtill kommer *Information work* och *Library management*. Innehållet kan tolkas så att det handlar om ledning och styrning av bibliotek och biblioteksarbete.

Klustret innefattar också deldisiplinen *Information Management*, d.v.s. information sedd som resurser inom organisationer och som objekt för strategisk planering och förvaltning inom organisationer.

Kluster 9

Tre författare.

Rice, R.E. 2; Hiltz, S.R. 1; Daft, R.L. 1.

En författare har inte återfunnits i LISA: Daft, R.L.

51 bibliografiska poster.

Termer:

Information work 16; Telecommunications 13 ; Teleconferencing 10 ; Data transmission 10
Services 9 ; User services 8 ; Computerised information retrieval 6; Computerized
information storage and retrieval 6; Conferences 6 ; Information communication 6 ;
Information storage and retrieval 6; Communications 5; Computers 5; Online information
retrieval 5 ;Searching 5; Subject indexing 5; Technical services 5 ;

Electronic mail 4; Online computers 4; Publishing 4; Computerized information work 3;
Electronic Information Exchange System 3; Electronic media 3; Information
communication 3; Innovations 3; Library and information science periodicals 3; Mail 3
Periodicals 3; Published materials 3; Technological innovations 3; Bibliometrics 3

Book review abstracts 2; Communications technology 2; Computer supported cooperative work 2; Evaluation 2 ; Information science 2; Librarianship 2; Library materials 2; Managers 2; Research 2 ; Securities analysts 2 ; USA 2 ; Use 2

Tolkning:

Klustret synes kunna karaktäriseras av den tekniska aspekten av kommunikation i form av telekommunikation, telefonkonferenser, dataöverföring, kommunikationsteknologi.

Kluster 10

9 författare

Zmud, R.W. 3; Henderson, J.C. 3; Earl, M.J. 2; Keen, P.G.W. 2; Brancheau, J.C. 2; Rockart, J.F. 1 ; Rogers, E.M. 1; Venkatraman N. 1; Mintzberg, H. 1.

En författare saknas i LISA: Rockart, J.F.

En har en oidentifierbar namnform: Henderson, J.C.

22 bibliografiska poster.

Computer applications 4, Management 3, Economic aspects 2, Book review abstracts 2
Business management 2, India 2, Information communication 2,
Information management:the organizational dimension 2, Information work 3
Models 2, Networks 2, Organization2, Reengineering 2, Reorganization 2
Research 2, Telecommunications 2

Tolkning:

Klustret synes innehålla dokument med information i organisationen information management. Klustret tangerar de företagsekonomiska disciplinerna management och organisationsteori.

Kluster 11

5 författare

Clemons, E.K. 3; McFarlan, F.W. 2; Malone, T.W. 1; Brynjolfsson, E. 1; Porter, M.E. 1.

Två författare har inte kunnat återfinnas i LISA: McFarlan, F.W; Porter, M.E.

4 bibliografiska poster.

Termer:

Database management systems 2; Information work 2 ; Management information systems 2
Services 2; User services 2; CODASYL 2

Computer supported cooperative work 1; Computers 1 ; Operating systems ; Oval ; Systems
management 1 ; Computerized information work 1

Tolkning:

Klustret innehåller få dokument och därmed indexeringstermer. Innehållsmässigt synes
klustret tangera datalogin och systemvetenskapen med inslag av administration av datasystem.

Kluster 12

8 författare.

Benbasat, I. 2; Desanctis, G. 2; Glaser, B.G. 2; Huber, G.G. 2; Yin, R.K. 2; Smaglik, P. 2;
Simon, H.A. 1; Lewis, R. 1.

Två författare saknas i LISA: Yin, R.K.; Smaglik, P.

Två har en namnform som inte har kunnat identifieras. Huber, G.G., Glaser, B.G

5 bibliografiska poster.

Termer:

Bias; Cognitive science; Computer supported cooperative work; Computerized models;
Computers; Decision support systems; Expert systems; Groupware; Information technology;
Planning; Software engineering; Telecommuting;

Tolkning :

Termerna tyder på att författarna behandlar informationsvetenskapen ur ett *knowledge
management* - perspektiv innefattande effektivisering av informationshantering genom
expertsystem och tekniska hjälpmedel.

Kluster 13

3 författare.

Egghe, L. 2; Price, D.D. 1; Rousseau, R. 1.

153 bibliografiska poster

Termer:

Bibliometrics 56; Library materials 31; Citation analysis 23; Scientometrics 19; Frequency distributions 15; Informetrics 14; Mathematical models 13; Periodicals 12; Obsolescence 9; Impact factors 9; Bradford's Law of Scatter 9; Scatter 8; Online information retrieval 8; Lotka's Law 8; Science 7; Authors 7; Growth of literature 6; Science and technology 4; Research 4; Ranking 4; Productivity 4; Mathematical models 4; Information science 4; Concentration measures 4; Co authorship 4; Belgium 4; Articles 4;

Tolkning:

Även kluster 13 består av författare som ägnar sig åt bibliometri, scientometri och infometri. De här författarna är dock mer inriktade på kvantitativa mått samt matematiska och statistiska modeller för domänanalyser och beståndsutveckling.

6. DISKUSSION

Den företagna undersökningen har haft till syfte att genom cociteringsanalys och hierarkisk klustring ordna författare i grupper. Tanken är att de författare som inordnas i en grupp har en ämnesmässig gemenskap i sitt författarskap. Detta gemensamma ämne kan uttydas genom att de indexeringstermer som åsätts den bibliografiska posten i en bibliografisk databas räknas och tolkas. Det övergripande syftet har varit att undersöka om metoderna kan användas i praktisk verksamhet för att tillfredsställa ett informationsbehov som t.ex. en student eller en bibliotekarie har. I denna undersökning har LIS valts som ämne, citeringsdatabasen *SSCI* har använts för att inhämta citeringsdata och databasen *LISA* har valts som bibliografisk databas.

Undersökningen visar att det är möjligt att utmönstra grupper av författare ur den mängd av de 100 mest citerade författare som legat till grund för utredningen, vilket inte är givet på förhand. Det är fullt tänkbart att det inte går att urskilja grupper i ett undersökt material. De här författarna har ordnats i 13 kluster. Genom frekvensberäkning av indexeringstermerna är det möjligt att bestämma det ämnesmässiga innehållet i klustren. Sammanfattningsvis kan de 13 klustren och dess kognitiva innehåll beskrivas på följande sätt.

Kluster 1	(9)	Praktiska tillämpningar av IT med inriktning på användarna
Kluster 2	(4)	Scientometri, Forskningssociologi
Kluster 3	(5)	IR med inriktning på användarna och pedagogik.
Kluster 4	(21)	IR med inriktning på datalogiska tillämpningar, "soft IR"
Kluster 5	(9)	Algoritmisk IR "hard IR"
Kluster 6	(11)	Bibliometri, domänanalyser
Kluster 7	(9)	Sociologiska och kulturella aspekter på information
Kluster 8	(4)	<i>Management, Library Management, Information Management.</i>
Kluster 9	(3)	Kommunikation, Kommunikationsteknologi
Kluster 10	(9)	Organisationen och information
Kluster 11	(5)	Informationssystem
Kluster 12	(8)	Knowledge management
Kluster 13	(3)	Matematisk och statistisk bibliometri

Ett sätt att använda den här informationen är att ordna den efter storleken och då efter antalet författare i klustren. Ordningen kan sägas utvisa forskningsintensiteten inom ämnet.

Kluster 4	(21)	IR med inriktning på datologiska tillämpningar ” <i>soft IR</i> ”
Kluster 6	(11)	Bibliometri, domänanalyser
Kluster 7	(9)	Sociologiska och kulturella aspekter på information
Kluster 10	(9)	Organisationen och information
Kluster 1	(9)	Praktiska tillämpningar av IT med inriktning på användarna
Kluster 5	(9)	Algoritmisk IR, ” <i>hard IR</i> ”
Kluster 12	(8)	<i>Knowledge management</i>
Kluster 3	(5)	IR med inriktning på användarna och pedagogik
Kluster 11	(5)	Informationssystem
Kluster 2	(4)	Scientometri, Forskningssociologi
Kluster 8	(4)	<i>Management, Library Management, Information Management</i>
Kluster 9	(3)	Kommunikation, Kommunikationsteknologi
Kluster 13	(3)	Matematisk och statistisk bibliometri

Vissa kärnområden såsom ”pedagogisk” IR, Soft IR, och Hard IR urskiljs klart. Klustren 2, 6 och 13 som innefattar olika aspekter på bibliometri och scientometri är också tydliga. De klustren innehåller också flest författare. Svårast att avgränsa och beskriva är klustren som befinner sig i periferin, d.v.s. klustren som gränsar mot datologi, företagsekonomi och sociologi.

Det kan noteras att en subdisciplin som kulturpolitik som anses ingå i biblioteks- och informationsvetenskapen i Sverige inte finns representerat som kluster. *Management, Library management*, som jag anser vara representerat som kluster, återfinns inte i grundutbildningen vid Bibliotekshögskolan i Borås, med undantag av de inslag av undervisning inom beståndsutveckling som bedrivs inom kollegium 4.

Även om behoven kan skilja sig något mellan en students eller en bibliotekaries informationsbehov bedömer jag att metoden väl tjänar sitt syfte. Bibliotekarien kan använda resultatet för att bygga upp ett bestånd koncentrerat på kärnan i ett ämne. Studenten kan snabbt få en överblick av centrala författare och snabbt överblick över ett ämnes innehåll.

En förutsättning är då att hanteringen av indexeringstermerna automatiseras. Såvitt jag kan bedöma är detta inte tekniskt komplicerat utan att termerna formateras på ett sådant sätt att de enkelt kan inhämtas för att sedan sorteras och räknas automatiskt.

Även om det finns ett värde i att beskriva ett ämne vid en viss tidpunkt ligger metodens verkliga värde i att regelbundet upprepa den. Resultatet blir då att det går att utläsa ett ämnes utveckling. Det går att utläsa förändringar i forskningens frontlinje, samtidigt som man kan se hur deldiscipliner och frågesällningar minskar i intresse medan andra vinner terräng. Det går vidare, bland annat, att utläsa generationsskiften inom forskarsamfundet eller eventuellt geografiska förändringar var de mest excellenta forskarna inom ett ämne befinner sig. Detta har dock inte varit studiens syfte. Värdet av en sådan studie hade kanske också varit av mindre värde eftersom jag inte har kunnat finna någon tidigare studie där exakt samma metod som jag använt har använts. Resultaten är därför inte jämförbara.

Det ska framhållas att den här utredningens resultat inte är en ”exakt avbildning av verkligheten”. Olika faktorer och val påverkar resultatet. Det antal författare och det tidsintervall som väljs påverkar sannolikt resultaten. Val av likhetsmått vid cociteringsanalysen torde också ge utslag, likaså påverkas resultatet av den klustringsalgoritm som används.

Att författaren förekommer i den bibliografiska databasen man hämtar indexeringstermerna är givetvis också viktigt för hur klustren tolkas. I den praktiska och vardagliga tillämpningen är det rimligt att man söker efter författarna i andra bibliografiska databaser.

Att tilldela varje kluster ett ämne är i hög grad beroende av mängden indexeringstermer som kan associeras till klustret. Som framgår ovan varierar klustrens storlek med avseende på antalet författare i klustret. Inom klustren varierar antalet bibliografiska poster för de olika författarna.

Inom klustren men och också mellan klustren varierar författarnas produktivitet eller, mer korrekt, hur många dokument som indexerats i LISA. Ju fler dokument som indexerats desto fler indexeringstermer. En produktiv författares ämnesinriktning kommer därför att påverka frekvensen av indexeringstermer i klustret. Detta kan innebära att en eller ett fåtal författares *oeuvre* får genomslag vid frekvensberäkningen. Författaren kan också publicera artiklar inom ämnen som inte är typiska för klustrets övriga författare. Eftersom frekvensanalysen baseras på samtliga dokument och inte bara de dokument som cociterats (cocitering av författare innefattar även en citering av dokument även om dessa dokument inte individualiseras) kommer frekvensberäkningen att påverkas av indexeringstermer som inte finns i de citerade dokumenten. Det kan också få till följd av att indexeringstermerna späds ut.

Indexeringspraxis kan förändras över tiden. Nya termer kan infogas i den thesaurus som ligger till grund för indexeringen. Detta torde inte påverka frekvensanalysen eftersom nya ämnesområden tillkommer i klustret. Termer blir också omoderna eller obsoleta. Har dessa termer används ofta när de var adekvata, påverkar också detta hur klustrets ämnesinnehåll kan uppfattas.

I LISA har jag kunnat notera att äldre bibliografiska poster tilldelas fler indexeringstermer än yngre bibliografiska poster. Detta påverkar naturligtvis frekvensberäkningen. Jag har också kunnat notera att vissa termer som t.ex. *information work*, *library materials* förekommer i många bibliografiska poster utan att vara så distinkta att de ger en klar bild av ämnesinnehållet.

Hierarkisk klustring är en god metod att visualisera ett ämnes kognitiva innehåll och ämnets subdiscipliner. Trots de metodologiska brister som finns när det gäller att bestämma klustrens kognitiva innehåll med frekvensberäkningar av indexeringstermerna kan man få en god bild av vad som avhandlas inom ämnet. Vad gäller LIS går det att urskilja ett huvudsakligt ämnesinnehåll med subdiscipliner.

Metoden skulle kunna utvecklas genom att någon typ av vägning av termer gjordes så att termer med hög frekvens tilldelas ett lägre viktat värde medan en term med låg frekvens tilldelas ett högre värde. Ett annat sätt skulle kunna vara att beräkna någon form av normaliserat värde så att skillnaderna mellan högfrekventa termer och lågfrekventa termer jämnas ut. Fördelen är att med sistnämnda metod är att högfrekventa termer inte får genomslagskraft på bekostnad av lågfrekventa men mer särskiljande termer.

Andra sätt att komplettera utredningen är att jämföra olika bibliografiska databasers indexeringspraxis. Som framgått ovan i avsnitt 4.1 tilldelas de dokument som indexerats i citeringsdatabasen *SSCI* sökbara *topics*-fält. Det visar sig dock att alltför få dokument tilldelas några termer för att en sammanställning av dem ska ge en rättvisande bild av innehållet.

Som framgått av metodavsnittet finns det flera olika typer av likhetsmått. Så vitt jag har kunnat finna saknas någon systematisk och jämförande analys om, och i vilken omfattning de olika måtten ger olika resultat och vilket likhetsmått som är mest adekvat för bibliometriska tillämpningar.

Klustren har visualiserats genom hierarkisk klustring. Det hade varit önskvärt att kunna visualisera klustren och dess relationer till varandra med en karta baserad på MDS. Skälet till att detta inte har kunnat ske är att den programvara som använts inte klarar av att grafiskt representera mer än 75 författare på en karta. I studien har de 100 mest citerade författare valts ut. 75 författare är dock för lågt för att man ska kunna uttala sig med säkerhet om ämnets kognitiva innehåll. Problemet är inte ett metodologiskt problem utan ett praktiskt, som kommer att lösas med tiden om programvaruutvecklarna vill anpassa programmet till personatorer med större arbetsminne och snabbare processorer.

Hur förhåller sig då metoden att frekvensberäkna indexeringstermer till de metoder som ex.vis Persson (Persson 94) och White & McCain (White & McCain 98) använder för att bestämma av klustrens ämnesmässiga innehåll i sina respektive utredningar? Bägge utredningarna använder sig av författar co-citerigs data, samt inordnar författarna kluster.

Persons ansats skiljer sig avsevärt från förevarande utredning. Bortsett från tekniska aspekter i fråga om val av likhetsmått m.m. skiljer sig utredningarna åt dels vad gäller syftet med utredningen dels vad avser det källmaterial som avses. Person har för avsikt att bestämma forskningsfronten i de artiklar som publicerats i en viss tidskrift under en tidsperiod på fem år. Redan här är Perssons utredning avsevärt mycket snävare. Det som trots detta anknyter till min utredning är att Person har ett behov av att bestämma innehållet i de kluster som han har kunnat identifiera. Detta görs genom titlarna på 51 artiklar från författarna som ingår i de fem slutligt bestämda klustren.

Det är inte givet att titeln på dokument avspeglar dokumentets innehåll, alla aspekter på innehållet eller förmedlar något djup. Det finns inte heller någon standardisering av de termer som titeln ska bestå av. Detta finns å andra sidan när ett dokument ska indexerats och auktoritetslistor i form av thesaurier används. Vidare kan man, i vart fall vid manuell indexering, fånga upp och beskriva aspekter på dokumentet som inte framgår av titeln. Det saknas inte skäl till att bibliografiska databaser har andra sökbegrepp än dokumentets författare och titel.

Bestämningen av klustrens innehåll i är än mer problematisk hos White & McCain eftersom etiketteringen av klustren helt och hållet baserar sig på utredarnas personliga kunskaper om ämnet. Författarna ifrågasätter själva i många fall rubriceringen och fyller ut och förklarar dem. Av naturliga skäl är det inte möjligt för en praktiskt verksam informationsspecialist att besitta så djup och specialiserad kunskap om författarnas *oeuvre* i alla ämnen. Det är inte heller praktiskt möjligt att remittera kluster till en eller flera fackspecialister för ämnesbeskrivning.

Givet de brister som åtminstone finns i Perssons och White & McCains utredningar, utgör en co-citeringsanalys kombinerat med en frekvensberäkning av indexeringstermerna på ett sådant sätt som i förevarande utredning en i vart fall mer objektiv, mer komplex och nyanserad beskrivning av klustrens ämnesmässiga innehåll. Är man medveten om de brister som framgår ovan, kan ämnesbestämning genom frekvensbestämning av indexeringstermerna bli ett adekvat och snabb metod för praktiskt verksamma informationsspecialister eller andra som har behov av att snabbt få en överblick av ett ämnesområde förutsatt att frekvensberäkningen automatiseras.

7. SAMMANFATTNING

Syftet med studien har varit att tillämpa och pröva bibliometrisk metod i ett informationsförsörjningssammanhang. De metoder som valts är cociteringsanalys av författare, automatiserad klusteranalys samt frekvensberäkning av indexeringstermer associerade till bibliografiska poster för de i klustren ingående författarna.

Den teoretiska grundvalen är att författare som citerar samma författare har en ämnesgemenskap. samt att summan av indexeringstermerna för varje kluster ger en bild av klustrets ämnesmässiga innehåll.

De frågeställningar som besvaras är

Vilka grupperingar av författare kan identifieras?

Hur kan dessa grupper beskrivas till sitt ämnesmässiga innehåll om man sammanställer de indexeringstermer som kan associeras med de i klustren ingående författarna?

Som ämne för studien har Biblioteks- och informationsvetenskapen valts (LIS). Den företagna utredningen har gått till så att citeringsuppgifter hos de mest 100 citerade författarna under tiden 1995- 2000 har inhämtats från citeringsdatabasen *SSCI*. Cociteringsdata har ordnats i en råmatris och närhetsmått har beräknats. Därefter har data bearbetats med *BIBEXCELS* klustringsalgoritm. Därmed har tretton kluster med författare kunnat utmönstras. För varje författare i klustren har sedan bibliometriska data i form av indexeringstermer inhämtats från *LISA* och räknats. Utifrån frekvensen av de olika indexeringstermerna i klustren har dessa kluster tolkats och ett ämnesinnehåll har tilldelats varje kluster. Tretton kluster har kunnat utmönstras och ett ämnesinnehåll har tilldelats varje kluster.

Kluster 1	(9)	Praktiska tillämpningar av IT med inriktning på användarna
Kluster 2	(4)	Scientometri, Forskningssociologi
Kluster 3	(5)	IR med inriktning på användarna och pedagogik.
Kluster 4	(21)	IR med inriktning på datologiska tillämpningar, ”soft IR”
Kluster 5	(9)	Algoritmisk IR ”hard IR”
Kluster 6	(11)	Bibliometri, domänanalyser
Kluster 7	(9)	Sociologiska och kulturella aspekter på information
Kluster 8	(4)	<i>Management, Library Management, Information Management.</i>
Kluster 9	(3)	Kommunikation, Kommunikationsteknologi
Kluster 10	(9)	Organisationen och information
Kluster 11	(5)	Informationssystem
Kluster 12	(8)	Knowledge management
Kluster 13	(3)	Matematisk och statistisk bibliometri

Trots att det går att anföra metodologiska invändningar särskilt när det gäller frekvensberäkningen av indexeringstermerna, ger metoden en god och användbar bild av ett ämnes innehåll.

Tack till fil.mag. Bo Jarneving som hjälpt mig att ta fram data från *Social Science Citation Index*.

REFERENSER

- Aldenderfer, M. S., Blashfield, R K. (1984). *Cluster Analysis*. (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-044.) Beverly Hills and London: Sage Publications.
- Baeza-Yates, R., Ribeiro- Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.
- Bayer, A E., Smart, J C., Mc Laughlin, G W. (1990). Mapping Intellectual Structure of a Scientific Subfield Through Author Cocitations. *Journal of the American Society for Information Science*, vol 41, nr 6, s 444 – 452.
- Bonnevie, E. (2003). A Multifaceted Portrait of a Library And Information Science Journal: the case of the Journal of Information Science. *Journal of Information Science*, vol 29 nr1, s 11 – 23.
- Bradford, S.C. (1934) Sources of Information on Specific Subjects. *Engineering*, vol. 137, s 85-86.
- Ding, Y., Chowdury, G.D., Foo, S. (2000). Journals as markers of intellectual space: Journal co-citation analysis of information retrieval area, vol 1987-1997. *Scientometrics*, vol. 47, nr 1, s 55 – 73.
- Ding, Y. (1998). Visualization of Intellectual Structure in Information Retrieval: Author Cocitation Analysis. *International forum on Information and Documentation*, Vol 23 nr 1 s 25 – 36.
- Diodato, V. (1994) *Dictionary of Bibliometrics*. Binghamton, NY: The Haworth Press Inc.
- Duran, B S., Odell, P L. (1974) *Cluster Analysis- A Survey*. (Lecture Notes in Economics and Mathematical Systems.) Berlin: Springer-Verlag.
- Fjällbrant, N. (1997). *Scholarly Communication – Historical Development and New Possibilities*. <http://iatual.org/conference/proceedings/vol07/papers/nfpapers.html> [2004-08-25]
- Jarneving, B. (2002). The Enhancement of Information-seeking Through Visualization of Cognitive Structures. *The New Review of Information Behavior Research*, 2002.
- Kruskal, J B., Wish, M. (1978). *Multidimensional scaling*. (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-011.) Beverly Hills and London: Sage Publications.
- Kärkki R., Kortelainen, T. (1998). *Introduktion till bibliometrin*. Helsingfors: NORDINFO
- Lotka, A J. (1926). Statistics – the Frequency Distribution of Scientific Productivity *Journal of the Washington Academy of Science*, 17: s 317-325.
- Nisonger T. (1998) *Management of Serials in Libraries*. Englewood: Libraries Unlimited Inc.

- Persson, O (1994) The Intellectual Base and Research Front of JASIS 1986-1990. *Journal of The American Society for Information Science*, vol 45 nr 1, s 31- 38.
- Price, Derek J de Solla, (1965) Networks of Scientific Papers. *Science*, vol. 149, no 3683, s. 510-515.
- Rowley, J., Farrow, J. (2000). *Organizing Knowledge, An Introduction to Managing Access to information*. Adlershot England: Ashgate.
- Rudberg, B. (1993). *Statistik*. Lund: Studentlitteratur.
- Small, H. (1973). Co-citation in the Scientific Literature: a New Measure of the Relationship Between Two Documents. *Journal of The American Society for Information Science*, vol 24, s 265 – 269.
- White, H. D.,Griffiths B.C (1981) Author Cocitation: A Literature Mesasure of Intellectual Structure. *Journal of The American Society for Information Science*, vol 32, s163- 172.
- White, H. D., McCain K.W (1988) Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995. *Journal of The American Society for Information Science*, vol 49, s 327-355.
- Wormell, I. (1998). Infometric Analysis of the International Impact of Scientific Journals: How International are the International Journals? *Journal of Documentation*, vol 55 nr5, s. 584 – 605.
- Ziman. J. *An introduction to science studies: the philosophical and social aspects of science and technology*. Cambridge: Cambridge University press.