Wolfgang G. Stock
Karl-Franzens-Universität, Graz

Informetrische Untersuchungsmethoden auf der Grundlage der Textwortmethode

(Informetric Analyses Based on the Textword Method)

W.G. Stock: Informetric analyses based on the textword method. In: Int. Classif. 11 (1984) No. 3. p. 151–157, 14 refs (In German)

Informetrics, i.e. the quantitative consideration of (scientific) informations, has its foundation of observation in signals and messages, which are fixed within (scientific) texts. This article introduces those informetric methods working with the textword-method. This method of documentation, elaborated by *N. Henrichs,* is independent of vocabularies (thesauri) and classification systems; it allows the following informetric analyses: 1st neighbourhood and environment of terms, 2nd term clusters, and 3rd (quantitative) importance of terms. The results of "textword-informetrics" are useful for information science (e.g. thesaurus construction within normal sciences), information practice (e.g. information retrieval based on textword method), science and philosophy of science (e.g. history and systematics of concepts and terminologies), and for sciences and humanities (even informetric based scientific discoveries are possible). (Author)

1. Informetrie

Die Bemühungen um die quantitative Erfassung ihres Untersuchungsobjektes sind nahezu so alt wie die Informationsforschung selbst. Schon Mitte der 60er Jahre stellt *W. Feitscher* ein sog. „semantisches Informationsmaß" vor. Er grenzt Informationen als „Verknüpfungsmatrix aus den Klassen der Objekte und aus den Merkmalen der Objekte" (1) ab. Das semantische Maß berücksichtigt einen gewissen Stand der Verknüpfungsmatrix (Ausgangsmatrix M_0) sowie irgendwelche Texte (T_i), die einem Empfänger vorgelegt werden, so daß nunmehr die Empfangsmatrix M_1 entsteht. „Damit könnte ein *relatives* Maß für die semantische Information aufgestellt werden" (1, p. 150), insofern die Matrizen M_0 und M_1 miteinander verglichen werden und so die Semantik von T_i messen. Dies bleibt bei *Feitscher* aber ausschließlich Programm: Die Aufstellung passender Berechnungsalgorithmen bereitet nämlich „dieselben Schwierigkeiten wie sie bei der automatischen Übersetzung auftreten" (1, p. 150). Die Realisierung dieses Programms zur Messung des semantischen Aspektes von Informationen ist nach *Feitschers* Einschätzung aber für die Informationswissenschaft und -praxis wesentlich. Ohne ein semantisches Informationsmaß bleibt die „Informations‚wirtschaft' in der Situation einer Energiewirtschaft, die nicht im Besitze der physikalischen Größen Energie und Leistung wäre" (1, p. 151).

Gegenstand der Informationswissenschaft ist die Information, aber, gemäß *J. Koblitz,* nicht in allen Aspekten, sondern nur in zwei Gesichtspunkten: *erstens* interessiert die Informationswissenschaft die bedeutungstragende, sog. „semantische Information", die *Koblitz* so definiert: „Eine Information bzw. eine Nachricht ist eine Einheit aus einer Semantik und einem physikalischen Träger, die der zweckbestimmten Mitteilung von Gedanken und Gefühlen an andere Subjekte dient" (2). Die Informationsforscher und -praktiker bearbeiten solche Informationen, und diese „Informationstätigkeit" ist der *zweite* Gesichtspunkt der Informationswissenschaft. „Die Information im Sinne einer Tätigkeit ist ein Komplex zweckbestimmter Handlungen, der folgende Elemente umfaßt: Die Erzeugung (einschließlich Be- und Verarbeitung) von Informationen, die Übertragung von Informationen, die Speicherung von Informationen und die Befragung von Informationen" (2, p. 94).

Beide Aspekte von „Information" sind auch der Informetrie zugänglich. Die quantitative Betrachtung der Informationstätigkeit führt zu einer Informetrie der IuD-Systeme und der zugehörigen Arbeiten. Dies ist der Ansatz zur Informetrie von *L. Blackert* und *K. Siegel* (3) sowie von *M. Bonitz* (4). Die quantitative Betrachtung der bedeutungstragenden Information führt über die Erfassung der entsprechenden Signale etwa via Zitationsanalysen zu Darstellungen von Informationsübermittlungen sowie über die Erfassung der Nachrichten zur *Messung der Bedeutungen* selbst. Hier wird die Informetrie auch für die Wissenschaftsforschung relevant. Wenn wir unsere Betrachtungen auf wissenschaftliche Informationen einschränken, also „Wissenschaftsinformetrie" betreiben, können wir die semantischen Aspekte der im Wissenschaftssystem zirkulierenden Informationen messen und damit der Wissenschaftsforschung (Wissenschaftsgeschichte, Wissenschaftstheorie, Begriffstheorie usw.) Daten sowie Regelmäßigkeiten über wissenschaftliche Bedeutungsgehalte bereitstellen.

Während die Informetrie der Informationstätigkeiten primär Meßdisziplin der Informationswissenschaft ist, ist die Wissenschaftsinformetrie doppelt unterstellt, wie dies *H. Engelbert* zutreffend für die Wissenschaftsinformatik feststellt: „Mit der Wahl der Bezeichnung ‚Wissenschaftsinformatik' würde exakt der Platz dieser Wissenschaft definiert, deren Untersuchungsobjekt die Wissenschaft ist und die damit zu den Wissenschaften von der Wissenschaft gehört, die aber andererseits hinsichtlich der Methodologie des Herangehens zu den Strukturwissenschaften gehört, die sich mit dem Phänomen ‚Information' beschäftigen, das sämtliche Prozesse der Tätigkeit des Menschen durchdringt" (5). Die Ergebnisse der Wissenschaftsinformetrie als Meßdisziplin der Wissenschaftsinformatik (6) sind somit in der Informationswissenschaft wie in der Wissenschaftsforschung nutzbar.

Wissenschaftliche Informationen sind in wissenschaftlichen Texten fixiert. Zur Beschreibung solcher Texte dienen Dokumentationsmethoden, etwa Klassifikation, Thesaurus oder Textwortmethode. Diese Dokumentationsmethoden benutzen wir als Datenerhebungsmethoden für informetrische Analysen. Solch ein Verfahren ist günstiger als das eingangserwähnte von *Feitscher,* das sich als praktisch undurchführbar erwiesen hat. Dokumentationsmethoden haben ein (relativ) sicheres theoretisches Fundament, und es gibt bereits große Datenmengen, die

auf der Basis der angesprochenen Dokumentationsmethoden erstellt worden sind. Die Wissenschaftsinformetrie braucht quasi nur noch auf die Datenbanken zuzugreifen. – Wir wollen im Folgenden diejenigen informetrischen Untersuchungsmethoden skizzieren, die auf der Grundlage von Textbeschreibungen mittels der *Henrichs*schen Textwortmethode durchgeführt werden können.

2. Textwortmethode

Die Textwortmethode wurde Mitte bis Ende der 60er Jahre von *N. Henrichs* zum Zwecke der philosophischen Dokumentation ausgearbeitet (7). Die Erschließung von Texten nach dieser Dokumentationsmethode ist völlig unabhängig von Wörterbüchern (Thesauri) und Systemen (Klassifikationen). Die Beschreibungsbasis ist ausschließlich der (individuelle) Text. Bei der Indexierung werden gewisse Terme markiert, die für solche Themen stehen, die im Text (in den Ausführungen wie im Titel) besprochen werden. Es dürfen grundsätzlich nur Textwörter zur Markierung von Themen Verwendung finden, textfremde Terme sind verboten (deswegen „Textwortmethode"). Die markierten Textwörter werden dergestalt miteinander verknüpft, daß die entstehenden Textwortketten thematische Komplexe ausdrücken, die so im Text abgehandelt sind. Z.B.: Ein gegebener Text möge sieben Themen A, B, C, D, E, F und G bei vier thematischen Komplexen 1, 2, 3 und 4 umfassen. Das Indexat habe folgenden Aufbau:

A (1–4); B (2,4); C (4); D (1, 3–4); E (1–2, 4); F (3); G (4),

was Folgendes besagt: zunächst werden im Text A, D und E gemeinsam abgehandelt (Komplex 1), dann A, B und E (Komplex 2), gefolgt von A, D und F und letztlich A, B, C, D, E und G (Komplex 4).

Was sind die Vorteile der Textwortmethode gegenüber „klassischen" Dokumentationsmethoden wie Klassifikation und Thesaurus? Klassifikation wie Thesaurus bauen nach vorgegebenen Aspekten Netze bzw. Hierarchien von als wichtig angesehenen wissenschaftlichen Begriffen bzw. Wissensgebieten auf. Es sind zwar jederzeit neue Terme in die Begriffssysteme bzw. Wörterbücher aufnehmbar, aber diese müssen stets im gegebenen System verortet werden. Gibt es aber überhaupt solche, von den betreffenden Fachwissenschaftlern allgemein akzeptierbaren Begriffssysteme? Auf diese Frage gibt der Wissenschaftshistoriker *Th.S. Kuhn* eine (eingeschränkt) positive Antwort. „Solange die von einem Paradigma gelieferten Hilfsmittel" – also wohl auch entsprechende Thesauri oder Klassifikationssysteme – „sich als fähig erweisen, die von ihm definierten Probleme zu lösen, schreitet die Wissenschaft dann am schnellsten voran und dringt am tiefsten ein, wenn diese Hilfsmittel voll Überzeugung gebraucht werden" (8). Aber dies gilt ausschließlich im Rahmen der „Normalwissenschaft", in einer als Ganzheit aufzufassenden theoretischen Tradition. Gibt es so etwas überhaupt? Die Geschichte der Wissenschaft zeigt, daß es solche Ganzheiten durchaus gibt, sie zeigt aber auch, daß diese Ganzheiten verdrängt und neue kreiert und vertreten werden. Jede einigermaßen umfassende wissenschaftliche Theorie gibt letztlich eine gewisse Systematisierung ihre Vokabulars vor. Aber *ein* Begriffssystem einer gegebenen Fachdisziplin, das alle Vertreter dieser Disziplin akzeptieren könnten, dürfte es bei kaum einer Wissenschaftsdisziplin geben. So zwingen Thesaurus und Klassifikation das disziplinspezifische Begriffsrepertoire in ein gewisses „normalwissenschaftliches Korsett" hinein, das – im Sinne *Kuhns* – zwar dieser Normalwissenschaft nützt, einer Theoriendynamik und einem wissenschaftlichen Pluralismus aber prinzipiell schadet. „Es kann nicht die Aufgabe der Dokumentation sein, Ideologien, die sich herausgebildet haben, weiterhin zu verfestigen" (9). *Henrichs* versucht vielmehr, „möglichst ‚objektiv' zu arbeiten und nicht von unserer Seite eine Ideologie in die Verarbeitung der Dokumente mit einfließen zu lassen" (9, p. 234). Dieser pluralistische Ansatz, der allen theoretischen Ganzheiten gerecht wird, ist der Vorteil der Textwortmethode. Ihr Nachteil wäre, daß Terme aus allen angebotenen Theorien zusammenfallen und zu einem Begriffswirrwarr entarteten, so daß ein Information Retrieval praktisch unmöglich würde.

Dies muß aber nicht sein. Über die Beschreibung der Texte via Textwortmethode und ihre informetrische Auswertung wird es nämlich möglich, die entsprechenden theoretischen Ganzheiten empirisch zu erfassen, etwa als Cluster gewisser Begriffe. Sollten sich abgrenzbare Ganzheiten und somit eine Indikation auf eine „Normalwissenschaft" ergeben, so könnten auf dieser Grundlage sogar Thesauri bzw. Klassifikationssysteme für die Zwecke dieser Normalwissenschaft konstruiert werden. Hierdurch erhalten die informetrischen Untersuchungsmethoden auf der Basis der Textwortmethode eine wesentliche Relevanz für informationswissenschaftliche Forschungen.

Die Erfassung solcher Ganzheiten hat aber auch eine zentrale Bedeutung für die Wissenschaftsforschung, können doch so Paradigmen, Theorien, Systeme, überhaupt wissenschaftliche Bedeutungsträger, empirisch erfaßt werden. Klassifikation wie Thesaurus eignen sich hierfür nur bedingt, denn sie setzen ja eine Normalwissenschaft voraus und betrachten aus diesem Blickwinkel die wissenschaftlichen Informationen. Durch das Faktum, ihre Beobachtungsbasis ausschließlich im wissenschaftlichen Text zu haben, ist die Textwortmethode geradezu prädestiniert, Hilfsmittel der Wissenschaftsforschung zu werden.

N. Henrichs, der Entwickler der Textwortmethode, hat selbst darauf hingewiesen, daß die Nachweise von Informationssystemen, die Dokumente nach seiner Methode ausgewertet haben, als Grundlage für quantitative Analysen von wissenschaftlichen Thematiken herangezogen werden können. Neben die Anwendung solcher Informationssysteme für das Information Retrieval und für die Erstellung von Bibliographien oder Registern „läßt sich noch eine weitere Nutzungsmöglichkeit stellen, und zwar im Zusammenhang mit heuristischen Verfahren, die auf Datenbankinhalte angesetzt werden können, um sie nach verschiedenen Aspekten hin zu durchforschen. Beispiele dafür sind etwa ideengeschichtliche Untersuchungen" (10).

Wenden wir uns nun den Untersuchungsmöglichkeiten und Kennwerten zu, die auf der Grundlage der Textwortmethode angewandt werden können. Wir werden drei Gruppen von informetrischen Methoden vorstellen: *erstens* Untersuchungen zum Umfeld wissenschaftlicher Begriffe, *zweitens* Analysen zu den Relationen zwischen

wissenschaftlichen Begriffen und *drittens* Untersuchungen zur Wichtigkeit wissenschaftlicher Begriffe.

3. Begriffsumfelder

Zunächst kommen wir zu den Untersuchungen zum Umfeld eines wissenschaftlichen Begriffs. Wir gehen hierbei aus von einer Liste aller markierten Textwörter („Deskriptoren" im Sinne der Textwortmethode), die in einem Informationssystem gespeichert sind, und wählen die für unseren Forschungsgesichtspunkt relevante Teilmenge von Dokumentennachweisen aus. Diese Teilmengenauswahl kann beispielsweise Dokumente bestimmter Zeitschriften, Autoren, Schulen, Zeiten, Räume, Sprachen usw. beinhalten, deren Terminologie wissenschaftsinformetrisch analysiert werden soll.

Nach *Henrichs* (11) erhalten wir das Umfeld eines Terms durch *vier* Listen. Die *erste* und einfachste Liste besteht aus einer alphabetisch sortierten Wortfolge, in die auch Begriffsderivate, Komposita oder stehende Wendungen eingehen, z.B.

Freiheit
Freiheit, geschöpfliche
Freiheit, göttliche
Freiheit, mathematische
Freiheit, politische
Freiheit, soziale
Freiheit, transzendentale
⋮
Freiheitsantinomie
Freiheitsbewußtsein
Freiheitsdetermination
Freiheitsfortschritt
⋮
Freiheitszweck
(11, p. 227)

Als *zweite* Liste nennt *Henrichs* eine rückläufig sortierte alphabetische Liste. Hier erhalten wir diejenigen Begriffsderivate, bei denen der letzte Wortbestandteil das Ausgangswort ist, z.B.

Freiheit
⋮
Ideologiefreiheit
Wahlfreiheit
Widerspruchsfreiheit
Werturteilsfreiheit
Gewissensfreiheit
Wertfreiheit
(11, p. 228)

Die *dritte* Liste gibt eine Abfolge der permutiert geordneten mehrgliedrigen Terme, die vom Ausgangswort abgeleitete Teile besitzen, z.B.

freie Arbeit
freie Forschung
freie Konkurrenz
freie Lehre
freie Ursache
freier Geist
freier Wettbewerb
freies Handeln
Freiheitsbewußtsein Entwicklungsstufen-des-
(11, p. 228)

Als *vierte* Liste wird von *Henrichs* das sog. „Wortfragmentwörterbuch" vorgestellt. Wir erhalten hier eine Liste von Termen, wo das Ausgangswort eine Teilmenge der Buchstaben ausmacht. Zur Erzeugung dieser Liste werden alle Deskriptoren in Pentagramme (Buchstabenfolgen von fünf Zeichen) zerlegt, z.B.

Wider		
iders	freih	
dersp	reihe	
erspr	eihei	
rspru	iheit	
spruc	heits	
pruch	eitsb	
ruchs	itsbe	
uchsf	tsbew	
chsfr	sbewe	
hsfre	bewei	
sfrei	eweis	(11, p. 229)

Durch das Pentagramm „freih" wird der Term „Widerspruchsfreiheitsbeweis" auch als Umfeld von „Freiheit" erkannt.

Als *Maß* für Begriffsumfelder werden die vier dargestellten Kennwerte jedoch erst dann interessant, wenn wir die Ausprägungen, etwa relative Häufigkeiten, der einzelnen Umfeldterme wissen. – Es muß bei Untersuchungen dieser Art darauf geachtet werden, daß wir die Themen stets nur in Form ihrer materiellen Träger, der Signale, messen können. Die Betrachtung der Signale kann aber in Einzelfällen zu einer „semantischen Falle" werden. So würde z.B. der Term „Widerspruchsfreiheitsbeweis" auch zum begrifflichen Umfeld von „Reihe" gezählt, was ein offensichtlicher semantischer Fehler wäre.

4. Begriffsnetze

Kommen wir nun zur zweiten Gruppe von Untersuchungsmöglichkeiten wissenschaftlicher Terminologie, zu den Analysen der Relationen zwischen wissenschaftlichen Begriffen! Hier können wir zwei Varianten vorstellen, von denen die eine ein Vorläufer der zweiten, eleganteren Methode ist.

Die ursprüngliche Methode wurde von *Henrichs* (11, p. 230–232) erarbeitet und bietet zwei Kennwerte. Die „thematischen Bezüge" eines Begriffs erhält man, indem man dem Begriff alle diejenigen Terme zuordnet, die mit ihm in mindestens einer Textwortkette vorkommen, z.B.

Freiheit
  Arbeit
  Aristoteles
  Bewußtsein
  Gesellschaftsordnung
  ⋮
  Sozialismus
  Theokratie
  Unterdrückung
  Wille
  Zweck (11, p. 230)

Die „thematischen Invarianten" sind Terme, die zwischen zwei Begriffen stehen. Man erhält sie durch eine Schnittmengenbildung der zugehörigen Textwortketten, z.B.

Freiheit – Wille
  Absolute, das
  Arbeit

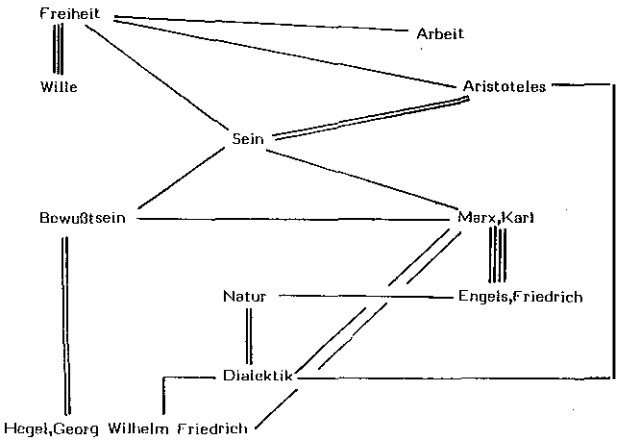
Gesellschaftsordnung
⋮
Sozialismus
Zweck (11, p. 231)

Wir wollen ein zweites Verfahren zur Analyse von Begriffsrelationen vorschlagen, das – ebenso wie das *Henrichs*sche Verfahren – auf der Basis der Textwortketten aufbaut. Es dies eine Variante des *Jaccard-Sneath-Koeffizienten* der Clusteranalyse.

Sei a die Anzahl der Dokumente aus der selektierten Dokumentenmenge DokM, wo der Deskriptor A auftritt, b die Anzahl der Dokumente aus derselben Dokumentenmenge, wo B auftritt, und g die Anzahl der Dokumente aus DokM, wo A und B gemeinsam auftreten und dort auch innerhalb (mindestens) einer Textwortkette verknüpft sind, so ist der Koinzidenzkoeffizient Ψ zwischen den zwei Themen A und B bzl. DokM:

$$\Psi_{A-B}^{DokM} = \frac{g}{a+b-g}.$$

Ψ_{A-B}^{DokM} nimmt Werte von 0 bis 1 an; 0, wenn keine Abhängigkeit vorliegt, 1, wenn maximale Abhängigkeit vorliegt. Der Koinzidenzkoeffizient ist nur sinnvoll anzuwenden bei häufig auftretenden Themen (etwa $n > 5$), da ansonsten der Zufall eine zu große Rolle spielt. Der Koinzidenzkoeffizient ist Grundlage für eine Cluster-Bildung von Thematiken. Wir erhalten hier ein „Soziogramm" wissenschaftlicher Themen. Alle Werte der Koinzidenzkoeffizienten werden mit einem festzulegenden Schwellenwert verglichen und diejenigen ausgesondert, die den Schwellenwert nicht erreichen. Die Liste der verbleibenden Paare ist nun graphisch umzusetzen in ein Netz, wobei das Ausmaß der Abhängigkeit zwischen zwei Themen durch die Stärke der Verbindung angedeutet wird, z.B.



5. Wichtigkeit von Begriffen

Die dritte Gruppe von informetrischen Analysen im Rahmen der Textwortmethode betrifft die Wichtigkeit wissenschaftlicher Begriffe. Ausgangspunkt für weitere Untersuchungsschritte ist die Berechnung der „Gewichtung" eines Deskriptors (Themas) in einem Dokument. Der entsprechende Algorithmus wurde von *Henrichs* (12) kreiert; er berücksichtigt drei Aspekte: *erstens* die Häufigkeit eines Deskriptors (die Anzahl der Textwortketten, in denen der betreffende Term vorkommt), *zweitens* die Mächtigkeit der Textwortketten (die Summe der Häufigkeitswerte der Deskriptoren der jeweiligen Kette) und *drittens* die Komplexität des Dokuments (die Summe der Mächtigkeiten aller Textwortketten, normiert auf 100). Der Berechnungsalgorithmus durchläuft sieben Schritte, die wir an unserem Beispiel-Indexat vorführen wollen.

Schritt 1: Errechnung des Themenwertes

Themen:	*Anzahl der Indices:*
A (1–4)	4
B (2,4)	2
C (4)	1
D (1,3–4)	3
E (1–2,4)	3
F (3)	1
G (4)	1

Schritt 2: Zuordnung der Themen zu den einzelnen Textwortketten

Schritt 3: Errechnung des Kettenwertes

Kette / Thema	Index 1	Index 2	Index 3	Index 4
A	4	4	4	4
B		2		2
C				1
D	3		3	3
E	3	3		3
F			1	
G				1
Kettenwert	10	9	8	14

Schritt 4: Errechnung des Dokumentwertes

10 + 9 + 8 + 14 = 41
41 ≙ 100

Schritt 5: Zuordnung der Kettenwerte zu den einzelnen Themen

Schritt 6: Summierung der Kettenwerte der einzelnen Themen

Schritt 7: Relativierung der in (6) erhaltenen Werte auf den Dokumentwert

Thema:	*Summenbildung:*	*Gewichtung:*
A	10+9+8+14 = 41	100,0
B	9 +14 = 23	56,1
C	14 = 14	34,1
D	10 +8+14 = 32	78,0
E	10+9 +14 = 33	80,5
F	8 = 8	19,5
G	14 = 14	34,1

Mit der „Gewichtung" können wir abschätzen, wie wichtig ein Thema im Rahmen eines gegebenen Dokuments ist. Uns interessiert nun aber die Wichtigkeit von Themen in bestimmten Dokumentenmengen. Wir wollen hierzu *drei* Kennwerte vorstellen, die diesen Sachverhalt quantitativ erfassen.

Erstens können wir mit der durchschnittlichen Wichtigkeit eines Themas arbeiten. Wir sehen die einzelnen Gewichtungswerte zu einem bestimmten Deskriptor (Des) innerhalb der in Betracht kommenden Dokumentenmenge DokM als Meßreihe an und bestimmen das D-Gewicht von Des in DokM als arithmetisches Mittel der Meßwerte nach der Formel

$$\text{D-Gewicht}_{Des}^{DokM} = \frac{1}{n} \cdot \sum_{i=1}^{n} (\text{Gewicht}_{Des})_i ,$$

wobei „n" die Anzahl der Dokumente in DokM zählt. Der Wertebereich von D-Gewicht$_{Des}^{DokM}$ liegt zwischen 0 und 100.

Verdeutlichen wir dies durch ein Beispiel! DokM habe 100 Dokumente (also n = 100), in deren fünf der Deskriptor A mit jeweils $Gew_A = 50$ vorkommt. In sechs Dokumenten von DokM komme B vor, und zwar fünfmal mit $Gew_B = 50$ und einmal mit $Gew_B = 10$. Wir errechnen folgende D-Gewichte.

$$\text{D-Gewicht}_A^{DokM} = \frac{1}{100} \cdot (50 + 50 + 50 + 50 + 50) = 2{,}5$$

$$\text{D-Gewicht}_B^{DokM} = \frac{1}{100} \cdot (50 + 50 + 50 + 50 + 50 + 10) = 2{,}6$$

Das durchschnittliche Gewicht (D-Gewicht) relativiert die Wichtigkeit eines Themas ausschließlich auf die unterstellte Dokumentenmenge.

Anders verfährt das *zweite* Maß zur Charakterisierung von Themen. Hier interessiert uns die Zentralität des Themas innerhalb einer Dokumentenmenge. Die Anzahl der Dokumente in DokM sei wiederum n. Die Anzahl derjenigen Dokumente aus DokM, in denen das zu beschreibende Thema Des vorkommt, sei durch „m" gezählt. Wir errechnen nun die durchschnittliche Häufigkeit von Des in DokM nach

$$P_{Des}^{DokM} = \frac{m}{n} \cdot 100$$

und die durchschnittliche Komplexität von Des nach

$$Q_{Des}^{DokM} = \frac{1}{m} \cdot \sum_{i=1}^{m} (\text{Gewicht}_{Des})_i \,.$$

Die durchschnittliche Zentralität von Des in DokM ergibt sich nach der Formel

$$\text{Z-Gewicht}_{Des}^{DokM} = \frac{P_{Des}^{DokM} + Q_{Des}^{DokM}}{2}$$

Der Wertebereich liegt auch beim Z-Gewicht$_{Des}^{DokM}$ zwischen 0 und 100.

Zur Veranschaulichung führen wir unser obiges Beispiel fort. In der beschriebenen Dokumentenmenge wollen wir die Themen A und B näher bestimmen, indem wir uns hier mit der Zentralität beider Deskriptoren befassen. Wir errechnen folgende Werte.

$$P_A^{DokM} = \frac{5}{100} \cdot 100 = 5$$

$$Q_A^{DokM} = \frac{1}{5} \cdot (50 + 50 + 50 + 50 + 50) = 50$$

$$\text{Z-Gewicht}_A^{DokM} = \frac{5 + 50}{2} = 27{,}50$$

$$P_B^{DokM} = \frac{6}{100} \cdot 100 = 6$$

$$Q_B^{DokM} = \frac{1}{6} \cdot (50 + 50 + 50 + 50 + 50 + 10) = 43{,}3$$

$$\text{Z-Gewicht}_B^{DokM} = \frac{6 + 43{,}3}{2} = 24{,}65$$

Das Maß relativiert die Einzelgewichtungen hier nicht auf die gesamte DokM, sondern durch den Koeffizienten Q_{Des}^{DokM} auf diejenige Teilmenge von DokM, wo Des auftritt. So kommt es, daß in unserem Beispiel das Z-Gewicht von A größer ist als das Z-Gewicht von B, wiewohl sich die beiden D-Gewichte genau andersherum verhalten. Die durchschnittliche Stellung von A ist nämlich zentraler als die von B, da B im Gegensatz zu A einmal nur am Rande abgehandelt wird.

Ein *drittes* Maß betrifft die durchschnittliche Gleichartigkeit bzw. Differenz der Stellung (Zentralität) von Themen in einer Dokumentenmenge. Zur Errechnung verwenden wir die Standardabweichung

$$s_{Des}^{DokM} = \sqrt{\frac{1}{m-1} \cdot \sum_{i=1}^{m} \langle (\text{Gewicht}_{Des})_i - Q_{Des}^{DokM} \rangle^2} \,,$$

die über die Summe der Abweichungsquadrate vom Mittelwert bestimmt wird.

Bezogen auf unsere Beispielthemen A und B erhalten wir die Werte

$$s_A^{DokM} = \sqrt{\frac{1}{5-1} \cdot 0^2} = 0 \,,$$

$$s_B^{DokM} = \sqrt{\frac{1}{6-1} \cdot \langle 5 \cdot (50 - 43{,}3)^2 + (10 - 43{,}3)^2 \rangle} = 16{,}33 \,.$$

Mit beiden Kennwerten Z-Gewicht$_{Des}^{DokM}$ und s_{Des}^{DokM} können wir prüfen, ob ein Thema ein durchschnittliches „Außenseiterthema" oder ein „Zentralthema" ist. Außenseiter ist ein Thema bei niedrigem s_{Des}^{DokM} und niedrigem Z-Gewicht$_{Des}^{DokM}$, was bedeutet, daß das Thema gleichmäßig am Rande abgehandelt wird. Zentralthema ist ein Thema bei niedrigem s_{Des}^{DokM} und hohem Z-Gewicht$_{Des}^{DokM}$, was bedeutet, daß das Thema gleichmäßig zentral besprochen wird. Bei hohem s_{Des}^{DokM} ist die Stellung des Themas in der entsprechenden Dokumentenmenge nicht einheitlich.

Soweit zu den Meßmethoden, die ausschließlich auf dem *Henrichs*schen Gewichtungsalgorithmus aufbauen. Wir wollen nun abschließend ein Verfahren vorstellen, daß die Ausführungen von Texten mit deren Titel in eine quantitative Beziehung bringt. Hier geht es um die Frage: Welche Terme tauchen bevorzugt in Titeln auf? Oder auch anders gefragt: Wie kann man die „Titelwürdigkeit" von Themen messen? Als Grundlage benötigen wir ein Kriterium, das uns unwichtige und wichtige Themen in einem Dokument trennt. In einer Untersuchung wissenschaftlicher Dokumente (13) konnte eine entsprechende Grenze festgestellt werden. Wichtige Dokumente zu einem Thema sind solche, die den entsprechenden Deskriptor mit einem Gewichtungswert auf der *Henrichs*metrik von $Gew_{Des} \geqslant 60$ enthalten. Sei t die Anzahl der Dokumente aus DokM, wo Des im Titel vorkommt, und w die Anzahl der Dokumente aus der gleichen Dokumentenmenge, wo Des wichtig, d.h. mit $Gew_{Des} \geqslant 60$ vorkommt, so ist die Titelwürdigkeit von Des in DokM

$$\text{TitW}_{Des}^{DokM} = \frac{t}{w} \,.$$

Ist TitW$_{Des}^{DokM}$ kleiner als eins, so wird das Thema „Des" zwar im Schnitt wichtig besprochen, aber nicht häufig in den Titeln genannt. Das entsprechende Thema ist nicht titelwürdig. Ist TitW$_{Des}^{DokM}$ größer als eins, so ist das Thema zwar häufig im Titel genannt, in den Ausführungen aber nicht entsprechend wichtig abgehandelt. Das Thema ist in diesem Fall ein „Modewort" bzl. seiner Nennung in Titeln. Ist letztlich TitW$_{Des}^{DokM}$ gleich eins, so be-

deutet dies, daß das Thema in der gleichen Anzahl von Dokumenten, wo Des im Titel genannt wird, auch wichtig abgehandelt wird.

Bei systematischen informetrischen Untersuchungen können die Selektionskriterien, nach denen die entsprechende Dokumentenmenge DokM aus der Datenbank ausgewählt wird, frei gewählt werden. Bei historischen informetrischen Analysen ist zusätzlich nach dem Kriterium „Zeit" (etwa jahrgangsweise) zu selektieren.

6. Anwendung der Informetrie in Informationswissenschaft und -praxis

Die Anwendung wissenschaftsinformetrischer Ergebnisse in der Informationswissenschaft und -praxis betrifft primär den Aufbau von Thesauri bzw. Klassifikationssystemen. Hier erhält man für solche Begriffsnetze eine empirische Basis. Sollten sich abgrenzbare Cluster von wichtigen Termen ergeben, so liegt es nahe, die im Netz zentralen Terme als Oberbegriffe anzusehen. Ebenso ist es möglich, Terme aus dem Umfeld eines Begriffs entweder als Nichtdeskriptoren oder als Unterbegriffe zum gegebenen Term zu definieren. Durch diese Informationsverdichtung entsteht nun eine Übersichtlichkeit und Strukturiertheit des Termmaterials derjenigen Bereiche, die überhaupt eine Strukturierung zulassen (Normalwissenschaften), so daß das bei der Textwortmethode befürchtete Begriffsdurcheinander nicht zutreffen muß. (In nichtnormalwissenschaftlichen Bereichen herrscht auch in der Wissenschaftssprache Konfusion – und damit in der IuD.)

Zu beachten ist hierbei, daß die entstehenden Wörterbücher bzw. Begriffssysteme ausschließlich beim Information Retrieval Relevanz haben, d.h. die Nachfrage erleichtern. Ein einmal entstandenes Begriffsnetz auch beim Indexieren vorauszusetzen, hieße, die empirische Basis des Textes zu verlassen und ist daher unsinnig. Jeder neue Text ändert letztlich die Struktur des Netzes und hält damit den Aufbau der Thesauri bzw. Klassifikationssysteme in ständiger Bewegung.

7. Anwendung der Informetrie in der Wissenschaftsforschung

Die Relevanz informetrischer Ergebnisse für historische Untersuchungen ist direkt einsichtig. Der Wandel von Begriffsumfeldern, Begriffsbeziehungen und Wichtigkeiten von Begriffen ist unerläßliches Ausgangsmaterial für jede Art begriffsgeschichtlicher Untersuchungen.

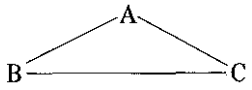
Die Wissenschaftsinformatik, also diejenige metawissenschaftliche Disziplin, die die Übermittlungen wissenschaftlicher Informationen analysiert, bekommt ihren „Rohstoff" durch informetrische Erhebungen der Themen bestimmter Autoren, Schulen, Zeitschriften, Disziplinen, Epochen, Ländern, Sprachen usw. Auch ist es möglich, von den einzelnen wissenschaftlichen Thematiken zu abstrahieren, um allgemeine Aussagen über Themen zu machen, etwa: Wie lange wird ein durchschnittliches Thema in der Wissenschaft besprochen? Wie lange bleiben Cluster stabil? Wie breiten sich Themen (über Schulen, Länder, Sprachen o. ä.) aus? Über derartige Informationsverdichtungen kann es durchaus möglich werden, Gesetzmäßigkeiten der Wissenschaftsinformatik zu entdecken.

Für die Wissenschaftstheorie ist die Informetrie insofern wichtig, als sie die „historische Relevanz" wissenschaftlicher Aktivitäten ausdrückt. Gemäß *E. Oeser* entscheidet „in der Rekonstruktion historischer Erkenntnisprozesse ... nicht der Wissenschaftstheoretiker, was wichtig oder unwichtig ist, sondern die historische Relevanz einer wissenschaftlichen Entdeckung steht bereits fest" (14). Die Aufgabe des Wissenschaftsinformetrikers, der mit der Textwortmethode arbeitet, ist, die Relevanz wissenschaftlicher Thematiken (etwa via den Kennwerten der Wichtigkeit) quantitativ zu erfassen, die des Wissenschaftstheoretikers, „die Anatomie des Erkenntnisprozesses zu leisten", also „die Freilegung des Argumentationsskeletts" (14, p. 29).

8. Anwendung der Informetrie in den Einzelwissenschaften

In denjenigen Einzelwissenschaften, deren Gegenstand Texte sind (also Sprachwissenschaft, Literaturwissenschaft, Geschichte usw.), hilft die Informetrie, die Themen der Texte im Sinne begriffsgeschichtlicher bzw. begriffstheoretischer Untersuchungen quantitativ zu erfassen.

Aber auch in allen anderen Einzelwissenschaften gibt es *vier* Aspekte, wo informetrische Analysen Bedeutung haben. Nehmen wir an, ein Forscher möge ein bestimmtes Thema bearbeiten wollen. Dieses Thema sei aber im betreffenden Cluster schon mit großer Ausprägung vorhanden. Das informetrische Ergebnis wirkt nun *erstens* als „Verbot", etwas Altbekanntes noch einmal zu bearbeiten. Nun sei das gestellte Thema nicht im Cluster enthalten. Dieses Ergebnis zeigt *zweitens,* daß die intendierten Forschungen Neuland betreffen, ein Faktum, das bei Projektförderungen oder bei der Vergabe von Hochschulschriften große Bedeutung erlangen kann. Nehmen wir jetzt an, daß ein Forscher von den informetrischen Ergebnissen ausgeht. Er kann hier relativ zu seinen Vorstellungen Lücken erkennen, oder er kann bislang am Rande liegende Themen in den Mittelpunkt rücken. Dieses heuristische Moment ist der *dritte* für alle Einzelwissenschaften relevante Aspekt. Gehen wir nun letztlich davon aus, daß bislang Arbeiten (möglicherweise aus verschiedenen Disziplinen) zu den Themen A und B, zu A und C sowie zu B und C, aber keine zu A, B und C gemeinsam ausgeführt worden sind. Im Cluster der Begriffsbeziehungen ergibt sich jedoch



und dies ist eine *neue wissenschaftliche Entdeckung,* die nunmehr theoretisch oder experimentell zu überprüfen ist. Informetrische Ergebnisse können also *viertens* Einzelwissenschaften durch neue Entdeckungen bereichern.

Quellen:

(1) Feitscher, W.: Ein Beitrag zur semantischen Informationstheorie, *ZIID-Zs.* 14 (1967), 147–153, hier: 148
(2) Koblitz, J.: Zum Informationsbegriff, *ZIID-Zs.* 14 (1967) p. 92–94, hier: 93.
(3) Blackert, L./Siegel, K.: Ist in der wissenschaftlich-technischen Information Platz für die INFORMETRIE?, *Wiss. Zs. d. TH Ilmenau* 25 (1979) p. 187–199

(4) Bonitz, M.: Scientometrie, Bibliometrie, Informetrie, *Zentralblatt für Bibliothekswesen* 96 (1982) p. 19–24
(5) Engelbert, H.: Wissenschaftliche Informationstätigkeit und wissenschaftlich-technischer Fortschritt, *Informatik* 25 (1978) 5, p. 41–47, hier: 46
(6) Stock, W.G.: Wissenschaftsinformatik, Fundierung, Gegenstand und Methoden, *Ratio* 22 (1980) p. 155–164
(7) Henrichs, N.: Philosophische Dokumentation. Zweite Mitteilung, *Z.f. philos. Forsch.* 23 (1969) p. 122–131; ders.: Philosophische Dokumentation, *Nachr. Dokum.* 21 (1970) p. 20–25; ders.: Philosophie-Datenbank, *Conceptus* 4 (1970) p. 133–144
(8) Kuhn, Th.S.: *Die Struktur wissenschaftlicher Revolutionen,* Frankfurt 4. Aufl. 1979, 89
(9) Henrichs, N.: (Diskussionsbeitrag zu seinem Vortrag) Sprachprobleme beim Einsatz von Dialog-Retrieval-Systemen. In: *Deutscher Dokumentartag 1974,* Bd. 2, München 1975 p. 233–235, hier: 234
(10) Henrichs, N.: Dokumentenspezifische Kennzeichnung von Deskriptorenbeziehungen. Funktion und Bedeutung. In: *Deutscher Dokumentartag 1974,* Bd. 1, München 1975 p. 343–353, hier: 351
(11) Henrichs, N.: Sprachprobleme beim Einsatz von Dialog-Retrieval-Systemen, in: *Deutscher Dokumentartag 1974,* Bd. 2, München 1975, p. 219–232
(12) Henrichs, N.: Benutzungshilfen für das Retrieval bei wörterbuchunabhängig indexiertem Textmaterial, in: R. Kuhlen (Hrg.): *Datenbasen – Datenbanken – Netzwerke.* Praxis des Information Retrieval, Bd. 3: *Erfahrungen mit Retrievalsystemen,* München/New York/London/Paris 1980, p. 157–168
(13) Stock, W.G.: Die Wichtigkeit wissenschaftlicher Dokumente retrieval zu gegebenen Thematiken, *Nachrichten für Dokumentation* 32 (1981) p. 162–164
(14) Oeser, E.: *Wissenschaftstheorie als Rekonstruktion der Wissenschaftsgeschichte,* Bd. 1, Wien/München 1979, p. 29

Dr. Wolfgang G. Stock
Langegg-Ort 25
A-8302 Nestelbach bei Graz

British Classification Society / Systematics Association Joint Meeting Nov. 1984

A joint one-day meeting is to take place on Friday, Nov.16, 1984 in the Meeting Room of the Linnean Society, Burlington House, London on the topic "Graphics in Classification". The program lists the following six papers with introductory remarks by J.C.Gower, the president of the British Classification Society and closing remarks by R.G.Davies, the president of the Systematics Association.

F.A.BISBY: Diagrammatic representation of biological taxonomies through the ages. – M.J.CLARK: Land surface classification – images and spatial information systems. – F.CRITCHLEY: Statistical classification and its presentation. – D.WISHART: Traversing the dendogram – in colour. – R.J.WHITE: Image analysis and intro-specific variation. – G.C.S.CLARKE: Presenting classifications to the public.

Organizers of the meeting are R.Alkin, Biology Department, Bldg.44, University, Southampton, S09 5NH and R.W.Payne, Statistics Department, Rothamsted, Harpenden den, AL5 2JQ, England.

Secretary of the British Classification Society is Dr.A.J. BOYCE, Dept.of Biological Anthropology, 58 Banbury Road, Oxford OX2 6QS, England.

Reports and Communications

Classification Society of North America Conference 1984

The annual conference (the 15th in the series of conferences of the former North American Branch of The Classification Society) took place in Santa Barbara, California, June 25–27, 1984 as a joint meeting together with the Psychometric Society. Altogether some 83 papers were presented in 16 sessions on 9 topics and during 3 symposia. The local arrangements were taken care of by the president of the Classification Society, Prof.Lawrence Hubert, University of California, he was also the program chairman for his Society and Prof. Bruce Bloxom, Vanderbilt University, for the Psychometric Society.

The following 9 topical areas were chosen, they are listed here together with the number of papers in brackets: Test Theory (13), Mathematical Models (5), Statistical Methods (10), Covariance Structures (12), Multidimensional Scaling (10), Measurement Theory (4), Classification (5), Classification Methodology (9), Classification Application (5).

The 3 Symposia were devoted to the following topics. New Developments in Item Factor Analysis (4), Consensus Analysis (3) and Statistical Analysis of Growth and Learning (3).

The 25 abstracts submitted through the Classification Society can be had by writing to the Secretary of the Society, Dr.George W.Furnas, Morris Research and Engineering Center. Bell Communications Research, Inc., 445 South Street, Morristown, NJ 07960, USA. Their titles are as follows: BOZDOGAN, H.: AIC – Replacements for multivariate multi-sample conventional tests of homogeneity models. – BOZDOGAN, H.. Multi-sample cluster analysis as an alternative to multiple comparison procedures. – REHNER, N.F.: Sampling distributions of consensus indices when all classifications are equally likely. – BAILEY, K.D.: Entropy clustering: from categorical to continuous data. – CORTER, J.E.: Extended similarity trees. – MARCUS, R.J.: Computer generation of structure-effect relationships from text data. – FULLER, V.. Classification and science for women and girls. – EDGINGTON, E.S.: Where do you draw the line? – RAVI KUMAR, C.N., CHIDANANDA GOWDA, K.: Generation and recognition of Kannada characters. – GHOLSTON, L.R., SCHEETZ, J.P.: Classifying males and females by mandibular and maxillary arch form and shape. – STOUT, R.L.: Multidimensional scalogram models for changes in psychotic symptoms over time. – LORR, M., KENDALL, A.J.: A comparison of IICA with two cluster methods using MMPI profiles. – GREEN, P.E., KRIEGER, A.M.: Buyer similarity measures in conjoint analysis: some alternative proposals. – HOLMAN, E.W.: Evolutionary effects in pre-evolutionary classifications. – CIAMPI, A., HOGG,