Humboldt-Universität zu Berlin

Philosophische Fakultät I

Institut für Bibliotheks- und Informationswissenschaft

Entropie eines Forschungsgebietes – Theoretische Erarbeitung und Durchführung einer Kozitations-Clusteranalyse im Projekt „Testing the homogenisation thesis – measuring the diversity of science“

Magisterarbeit

Zur Erlangung des akademischen Grades Magistra Artium (M.A.)

Marion Schmidt

Gutachter: 1. Dr. Frank Havemann

2. Prof. Dr. Walther Umstätter

eingereicht: 08.05.2006

Inhaltsverzeichnis

1. Einleitung

Diese Magisterarbeit ist parallel zur Mitarbeit an einem Projekt, einer feasability study mit dem Titel „A Methodological Study for Measuring the Diversity of Science“ entstanden. Diese Studie hat das Ziel, die Diversität einer Forschungslandschaft mit bibliometrischen Mitteln zu messen. Neuere Entwicklungen in der Wissenschaftspolitik – die Organisation der Mittelvergabe unter der Prämisse der Exzellenzförderung – bewirken, so die Ausgangsthese, eine Abnahme an Vielfalt in der Forschung, da einerseits die Anzahl geförderter Projekte abnimmt und andererseits WissenschaftlerInnen im Vergleich zu vorher sich zunehmend dazu gezwungen sehen können, sich auf Erfolg versprechende oder sogar wissenschaftliche Modethemen zu konzentrieren und randständigere Themen zu vernachlässigen. Durch bibliometrische Methoden sollen Kriterien entwickelt werden, die unabhängig von der naturgemäß subjektiven Wahrnehmung der WissenschaftlerInnen sind. Mit Hilfe einer Clusteranalyse werden die Zitationsdaten eines Fachgebietes – der Elektrochemie – in inhaltlich kohärenten Strukturen abgebildet und diese Partitionierung anschließend mit der *Shannon-Wiener*-Entropie gemessen.

Mein Beitrag bestand vor allem in der technischen Durchführung der Studie, d. h. der Verarbeitung der Daten sowie einer partiellen konzeptuellen Mitarbeit.

Inhalt der Magisterarbeit ist neben der Darstellung der feasability study eine vertiefte Erarbeitung und Analyse der methodischen Grundlagen der Studie sowie alternativer Methoden von Zitationsstudien in Bezug auf Clustermethoden und Ähnlichkeitsmaße.

2. Die Kozitationsanalyse

2.1. Einführung in die Methodik der Kozitationsanalyse

Marshakova (1973) und Small (1973) entwickelten unabhängig voneinander das Konzept der Kozitation, in dem die Anzahl der gemeinsamen Zitationen von zwei Artikeln durch andere Artikel Maß der inhaltlichen Nähe der zitierten Artikel zueinander ist.

Diese Konzeption basiert auf der Grundannahme, dass Zitationen Indikatoren dafür sind, wie Wissenschaftler in der Entwicklung von Theorien oder Methoden einander beeinflussen und sich zueinander verhalten. Small (1978) bezeichnet Zitationen als concept symbols: Die zitierten Texte sind Objekte, die Konzepte (Ideen, Methoden, experimentelle Ergebnisse u. a. eingeschlossen) repräsentieren, welche ein Autor für seine Arbeit benutzt. Hoch zitierte Artikel sind in dieser Terminologie standard symbols. Mit der Analyse von Zitationen kann untersucht werden, welche Konzepte Autoren teilen und wie sich Konzepte in der Wissenschaftslandschaft bewegen. Das Zitieren von benutzten Artikeln ist wissenschaftlicher Standard und unumgängliche Voraussetzung dafür, dass die Urheberschaft von Erkenntnissen transparent ist. Aufgrund dieser Situation ist es wissenschaftlich legitim, Zitationen ihrerseits zum Untersuchungsgegenstand zu machen und Schlussfolgerungen aus ihren Verteilungen zu ziehen.

Nach den Angaben Smalls & Sweeneys (1985a) beschrieb De Solla Price als Erster das Konzept einer umfassenden map of science.

Darauf rekurrierend begann das Institute for Scientific Information (ISI) in den frühen 70er Jahren mit der Entwicklung eines Atlas of Science. Diese Landkarte der Wissenschaft sollte, wie es Small & Griffith (1974a) zunächst als Ziel formulierten, Fachgebiete, specialty-Strukturen als Bausteine der Wissenschaft freilegen, bzw., wie Small et al. (1985b) in einer ausführlicheren Zielvorgabe ausführten, die folgenden Fragen beantworten: was die natürlichen Einheiten der Wissenschaft seien, ob und wie stark diese Einheiten miteinander korrelieren, welche unterliegenden Kräfte die Strukturen und Relationen determinieren, welche Rolle soziale Faktoren

spielen und ob eine Karte der Wissenschaft identisch mit einer des Wissens sei.

Die Fragestellung nach natürlichen Einheiten erweist sich vor dem Hintergrund des einige Jahre später entwickelten und nachgewiesenen Theorems der fraktalen Organisation der Wissenschaft (u. a. Van Raan, 1991) als irreführend: Dieses besagt, dass auf unterschiedlichen Ebenen Strukturen existieren, die jeweils zueinander ähnlich sind (self-similarity). Es ist daher unangemessen, eine dieser Strukturebenen als natürlich zu erklären, weil dies ein dem Gegenstand externer Maßstab wäre, der sich kaum legitimieren ließe.

Tatsächlich versuchen Small und seine Kollegen durch eine hierarchische Modellierung ihrer Daten diese hierarchische Struktur der Wissenschaft nachzuvollziehen – wenn sie auch keine explizite Untersuchung von *self-similarity* unternehmen und sich sogar umgekehrt Parameter zwischen Strukturen (wie die Proportionalität der Disziplinen zueinander) verändern: wie im Folgenden gezeigt werden wird, gelingt es nicht, ab der untersten Ebene eine wirklich angemessene Repräsentation zu konstruieren.

Methodisch basierten die Arbeiten am Atlas of Science auf der Kozitationsanalyse von Artikeln durch ein *single-linkage*-Clusterverfahren mit festen Schwellwerten – d. h. ein Link mit einer über einem spezifizierten Schwellwert befindlichen Kozitationsstärke reicht zur Mitgliedschaft in einem Cluster aus. 1974 veröffentlichten Small und Griffith eine zweiteilige Methodenstudie (Small & Sweeney, 1974a; Small et al., 1974b) und 1981 brachte das ISI die ersten beiden Versionen des Atlas heraus. Small verfeinerte die Methoden und veröffentlichte wiederum eine zweiteilige methodologische Studie (Small & Sweeney 1985a; Small et al., 1985b).

Neben diesen Makroansätzen – d. h. Versuchen, eine Landkarte der gesamten Wissenschaft zu einem bestimmten Zeitpunkt zu schaffen – wurden zahlreiche Mikrostudien als Versuche, nur einzelne Disziplinen oder specialties im Hinblick auf ihre Binnenstruktur zu kartographieren, entwickelt und publiziert.

Zu den mittlerweile ebenfalls klassischen Methoden im Bereich des science mappings gehören agglomerativ-hierarchisches Clustern, *Faktorenanalyse*

und andere *Eigenvektor-Analysen*, *Multidimensionale Skalierung (MDS)*, gefolgt von *K-Means Clustering*, *Self-Organizing Maps (SOM)* und *Latent Semantic Indexing* sowie graphanalytische Analysen und Visualisierungen (Small, 2003; Liu 2005; Boyack et al., 2002; Chakrabarti, 2003; Leydesdorff, 1993; Leydesdorff, 2004a).

Das Kozitationsprinzip kann auf Artikel, Autoren und Journale angewandt werden und ist ein Spezialfall der *co-occurence*. *Co-occurence* umfasst auch Bibliographische Kopplung, die zur Kozitation direkt komplementäre Methode: Während bei der Kozitation aus der cited-Perspektive die Anzahl der gemeinsamen Zitationen für zwei Objekte – Artikel, Autoren, Journale – gezählt wird, wird bei der Bibliographischen Kopplung umgekehrt aus der citing-Perspektive die Anzahl der gemeinsam referenzierten Artikel (oder Autoren bzw. Journale) für zwei zitierende Artikel (oder Autoren bzw. Journale) gezählt.

Co-occurence muss sich jedoch nicht auf unterschiedliche Objekte, die durch Zitationsbeziehungen miteinander verbunden werden, beziehen, sondern umfasst auch die Analyseebene des gemeinsamen Vorkommens von Wörtern oder Phrasen in Titeln; aufgrund der semantischen Beziehungen innerhalb von Objekten, die in anderen Objekten ebenfalls vorhanden sind, werden inhaltliche Zuordnungen der Objekte zueinander vorgenommen. Natürlich sind auch Kombinationen der Verfahren möglich, so schlugen Braam, Moed und van Raan (1991) eine Kombination von Kozitation und Ko-Wort-Analyse vor. Journal-Kozitationsanalysen sowie insbesondere Autor-Kozitationsanalysen (ACA), die durch White & Griffith 1981 eingeführt und durch weitere Studien der Drexel University bekannt wurden, sind sehr populär, was z. T. daran liegen mag, dass beide Typen gegenüber Artikel-Studien den Vorteil haben, eine sehr viel kleinere Datenbasis notwendig zu machen.

Gegen autorzentrierte Ansätze mit dem Ziel des science mappings spricht, dass Autoren im Rahmen ihrer Disziplin häufiger zu unterschiedlichen Themen publizieren und damit verschiedenen inhaltlichen communities angehören. Es ist außerdem als problematisch anzusehen, dass im überwiegend als Datenbasis verwendeten Science Citation Index des ISI in

den Zitationsstrings nur der erste Autor genannt wird, so dass daraus resultierende Zitationsstudien zwangsläufig unexakt sind, worauf auch Persson (2001) und Gmür (2003) hinweisen. Zudem ist die Tendenz zu Kollaborationen seit der Entwicklung des Konzeptes stark angestiegen: In einer Situation, in der Mehrautorenstudien die Regel sind, verliert das gemeinsame Zitieren von Autoren an Aussagekraft, da dies immer auch gleichzeitig auf alle anderen Autoren der jeweiligen Artikel zutrifft, die Anzahl der Kozitationen insgesamt also immer größer wird.

Bei Journal-Studien ergibt sich in Bezug auf die großen interdisziplinären Journals, die mit einer großen Anzahl von Journalen Zitationsbeziehungen unterhalten, eine ähnlich problematische Situation: Wie Autoren sind auch viele Journale a priori nicht community-bezogen. Zudem spielt das Renommee einer Zeitschrift bei ihrer Rezeption und Zitation eine immense Rolle. Small et al. (1985b) resümieren deshalb in Bezug auf die journal influence maps von Narin, diese intendierten primär die Repräsentation von Qualitätshierarchien und erst sekundär die von Ähnlichkeiten.

Kozitationsanalysen wurden durch Interviews, *mental maps* der jeweiligen Fachexperten u.ä. evaluiert. Wie Leydesdorff (1993) jedoch feststellte, bedeutet die Tatsache, dass maps von ExpertInnen als mit ihren eigenen Vorstellungen kongruent oder nützlich identifiziert werden, nicht zwangsläufig, dass sie korrekt sind bzw. die bestmögliche Repräsentation bieten.

Eine generelle Kritik an Kozitationsanalysen richtet sich darauf, dass theoretische Artikel überrepräsentiert würden (vgl. Braam et al. 1991). Verzerrungen im Gesamtbild können auch aus anderen Phänomen resultieren, wie fachspezifisch unterschiedlichen Publikations- und Zitationsmodi[1] oder extrem übergreifend referenzierten wissenschaftlichen Konzepten; Probleme, die Small und seine Kollegen durch unterschiedliche Modulationen der Schwellen auszugleichen versuchten und die m. E. zwar nicht generell Kozitationen als aussagekräftige Untersuchungsgegenstände der Szientometrie in Frage stellen, wohl aber in der Wahl der Methoden

1 Nach den ISI-Erkenntnissen werden in der Mathematik durchschnittlich weniger Referenzen pro Artikel aufgeführt als in der Biologie

berücksichtigt werden sollten. Während mit der Verkettungstendenz hoch zitierter Artikel generell zu rechnen ist, spielen fachbezogene Unterschiede natürlich insbesondere bei disziplinenübergreifenden Studien eine Rolle.

Wie im Folgenden analysiert wird, resultierten die Schwellen der ISI-Studien aus Heuristiken, d. h. sie sind Resultate vorhergehender Untersuchungen und Thesenbildungen und wurden im Lauf der Zeit experimentell weiterentwickelt.

Obwohl im Fall der Kozitationsmethode weder Gesetze noch eine weitergehende Theorie, die der Feinmodellierung der Daten zugrunde gelegt werden könnte, existieren und deshalb eine eindeutige analytische Methodenbewertung schwer fällt, sollen die Parameter einer Kozitationsanalyse, die Clustermethode und gegebenenfalls das Ähnlichkeitsmaß, im Hinblick auf die Probleme des Datenmaterials und die Erfordernisse unserer Studie untersucht werden. Mögliche Kriterien dazu sind: Cluster sollen diskret definiert sein, d.h. exakt definierte Grenzen haben und nicht kontinuierlich ineinander übergehen; die coverage, d. h. der Anteil des ursprünglichen Datensets, der Teil der Clusterdistribution ist, soll möglichst hoch sein; die Modellierungsfaktoren sollten möglichst wenig willkürlich, sondern in Bezug auf die Eigenschaften der Datenbasis begründbar sein und insbesondere bekannte bzw. identifizierbare Faktoren, die eine authentische Repräsentation eines Fachgebietes verfälschen, wie der Einfluss hoch zitierter Artikel, die typischerweise zu einem Makrocluster oder mehreren sehr großen Clustern führen, sollen möglichst gut nivelliert werden. Des weiteren muss das Verfahren auch bei größeren Datenmengen praktizierbar sein (*scalability*).

Dabei soll *MDS*, da es nur eine räumliche Darstellung ohne diskrete Gruppenzuteilungen vornimmt, im Folgenden nicht näher analysiert werden, ebenso wenig *LSI*, das eine Repräsentation aufgrund der Analyse von Wörtern und Wortkontexten und der Extraktion von latent terms (Börner et al., 2003) konstruiert.

2.2. Kozitation und Bibliographische Kopplung

Bibliographische Kopplung ist ein der Kozitation exakt entgegen gesetztes Verfahren: bei der Bibliographischen Kopplung gilt, dass für zwei Artikel die Anzahl der gemeinsamen Referenzen Maß der inhaltlichen Nähe ist, mit ihr werden also auf einer aktuellen Ebene Forschungscluster ermittelt, während die Kozitation durch die retrospektive Perspektive die Basis aktueller Forschung analysiert.

Jarneving (2005) vergleicht die beiden Methoden in ihrer Funktion der Konstitution von research fronts: Als research fronts gelten Gruppierungen von Artikeln eines aktuellen Jahrgangs, die in ihrem Zitierverhalten ähnlich sind. Während die Bibliographische Kopplung wie erläutert gemeinsame Referenzen direkt misst, müssen die Kozitationscluster auf den aktuellen Vergleichsjahrgang rückprojiziert werden, um research fronts zu konstituieren.

Jarneving lädt aus den Journal Citation Reports des SCI 2000 Artikel der 50 meistzitierten Journale im Fachgebiet Environmental Science und kreiert durch Segmentation und Zufallsauswahl ein repräsentatives Datenset.

Für die Schwellwerte für die Kozitation und Bibliographische Kopplung wurde der Kosinus-Koeffizient als Ähnlichkeitsmaß verwendet und jeweils das dritte Quartil der nach dem Kosinus-Wert geordneten Paare einer single linkage-Clusterprozedur zugeführt. Jarneving definiert für die Konstitution der research fronts aus den Kozitationsclustern die einfache Bedingung, dass Artikel des aktuellen Jahrganges mindestens einen Artikel aus einem Kozitationscluster zitieren müssen.

Im finalen Untersuchungssetting werden nur Cluster ab einer Größe von 10 Artikeln verwendet.

Jarneving untersucht die Ähnlichkeit der durch die beiden Methoden entstandenen research fronts über eine Wortprofil-Analyse. Hierfür werden Titelwörter durch den Ausschluss aller nicht-bedeutungstragender Wörter wie Präpositionen; Pronomen, Artikel (sog. Stoppwörter) und teilweise durch den Ausschluss von Verben extrahiert, eine Häufigkeitstabelle angelegt und ähnliche Wörter manuell standardisiert. Es wird jede Kozitations-*research*

front mit jedem bibliographischem Kopplungscluster anhand der normalisierten Wortähnlichkeiten – die gemeinsamen Titelwörter zweier Cluster werden mit dem Kosinus-Maß durch die Anzahl der einzelnen Titelwörter je Cluster normalisiert[2] – verglichen. Die Entscheidung, ab welchem Level die Gruppen als definitiv ähnlich zu betrachten sind, erfolgt heuristisch; die höchste vorkommende Übereinstimmung beträgt 0.67 und Jarneving wählt alle Gruppierungen ab dem Übereinstimmungsgrad 0.3 aus, was in 36 Kopplungsclustern (41%), die von den Kozitations-*research fronts* reflektiert werden, resultiert und in umgekehrt 32 bzw. 33% der Kozitations-research fronts.

36% der 1691 Artikel aus den 88 bibliographisch gekoppelten *research fronts* zitieren auch Kozitationscluster, während 29% der 2094 Artikel, die den 96 Kozitations-research fronts entstammen, auch Teil der Kopplungscluster sind.

Beide Methoden zum Vergleich von Kozitationsanalyse und bibliographischer Kopplung, die Wortanalyse und die Analyse der Zitationsbeziehungen, ergeben übereinstimmend, dass die beiden research fronts-Distributionen eher unterschiedlich sind, während die Mengen der research fronts, die jeweils durch die beiden alternativen Methoden generiert wurden recht nahe beieinander liegen: Die bibliographische Kopplung führt zu 88 *research fronts* und die Kozitationsanalyse führt zu 96 research fronts. Die coverage, die die Kozitationsmethode in der Generierung der research fronts erreicht, ist größer als die der Kopplung. Es lässt sich also keine eindeutige Überlegenheit einer der beiden Methoden schlussfolgern.

Klavans & Boyack (2005) vergleichen im Unterschied zu Jarneving in ihrer Visualisierungsstudie research fronts-Distributionen, die aus Bibliographischer Kopplung entstanden, mit Distributionen, die aus Kozitationsclustern bestehen. Sie vermerken als Resultat ihres Untersuchungssettings weniger Kozitationscluster als Bibliographische Kopplungscluster, treffen in der Bestimmung der Schwellwerte aber offenbar die methodisch fragwürdige Entscheidung, die Schwellwerte nach der identisch gesetzten Anzahl der durch sie selektierten Artikel und Referenzen

2 Diese Normalisierung ist sinnvoll, da sonst unterschiedliche lange Titel und unterschiedlich große Cluster den Vergleich verfälschen würden

zu justieren. So ist der Kopplungsschwellwert viel niedriger als der Kozitationsschwellwert[3]; während andererseits, da die Anzahl der gesamten Referenzen eines Artikeljahrganges per se wesentlich größer als die Artikelmenge selbst ist, die Relation zwischen den zur Clusterprozedur zugelassenen Objekten und der jeweiligen Gesamtobjektmenge unterschiedlich ist. Ihr Befund, dass die coverage bei Kopplungsclustern wesentlich größer als die coverage bei Kozitationsclustern sei, ist daher zwingend schon im Untersuchungssetting angelegt, da von ungleichen Grundbedingungen ausgegangen wird. Jarnevings Methode der Schwellwertermittlung erscheint dagegen für den Vergleich von Bibliographischer Kopplung und Kozitation wesentlich sinnvoller. Zudem ist es adäquater, die beiden Methoden in Bezug auf die Konstitution von research fronts zu vergleichen, da hier in beiden Fällen die gleiche Objektmenge im gleichen zeitlichen Segment geclustert wird, wenn auch mit unterschiedlicher Gewichtung, wie die relative Unähnlichkeit als Resultat der Untersuchung Jarnevings belegt.

3 Bei der Kozitation ist die absolute Kozitationsstärke > 3 und mindestens zwei Kozitationspartner sind erforderlich, bei der bibliographischen Kopplung ist die absolute Kopplungsstärke > 1 und mindestens zwei Kopplungspartner sind notwendig

3. Entwicklung der Kozitationsanalyse in den ISI-Studien

3.1. Anfänge

Erste Experimente von Small und Griffith (1974a) zur Schaffung einer map of science schließen ein Viertel des SCI-Jahrgangs von 1972 als Datenbasis und die Verwendung von single linkage als Clusterverfahren auf der Basis von verschiedenen festen Kozitations-Levels (d. h., mit einfacher Kozitationsschwelle, die dem simple matching coefficient entspricht) ein. Als weiterer Schwellwert wird eine absolute Zitationsschwelle von 10 Zitationen (in weiteren Studien variiert im Bereich von 4-20 Zitationen, abhängig davon, ob die ganze Datenbank oder ein Teil geclustert wurde) benutzt, um hoch zitierte Artikel zu selektieren.

Die Kozitationsschwelle wird auf verschieden Levels variiert: während das erste Level alle Kozitationspaare einschließt, umfasst z.B. das dritte Level nur noch die Paare, die mindestens dreimal kozitiert wurden; als höchstes Level experimentieren Small und Griffith mit sechs Kozitationen. Die Distribution auf dem Kozitationslevel 3 erweist sich als tauglichste Basis für die weitere Arbeit und Evaluation.

Die Tatsache, dass Cluster überhaupt entstehen, sehen sie als prinzipielle Bestätigung der specialty hypothesis. Sie beobachteten, dass große Gruppierungen durch Erhöhung der Kozitationsschwelle aufgespalten werden können und schließen auf eine potenzielle hierarchische Binnenstruktur der Cluster.

Das biomedizinische Cluster fiel durch eine abnormale Größe auf[4]. Es wird nach anderen Versuchen durch ein – zurecht als „rude“ bezeichnetes – brute force-Verfahren aufgebrochen, indem sowohl einzelne, besonders hochzitierte Methodenartikel völlig entfernt werden als auch die verbliebenen Artikel auf dem vierten Kozitationslevel einer Re-Clusterung unterzogen

4 es enthält 801 Dokumente auf Kozitations-Level drei, als nächstes folgt Chemie mit 92 Papers und mit jeweils 41 und 32 Nuklear- und Teilchenphysik

wurden. Es zerfiel dadurch in 74 Subcluster, von denen das größte 65 Dokumente enthielt.

Small und Griffith begründen den sog. biomedizinische Bias, den überproportionalen Anteil der Biomedizin, sowohl mit einer starken Repräsentanz biomedizinischer Literatur im ISI-Datenset als auch mit der beherrschenden Stellung einiger Standardmethoden in der Biomedizin, die stark zitiert wurden und damit zu einer großflächigen Verlinkung führten.

3.2. Fraktionale Zählweise und variable Kozitationsschwellen

Die einfache Kozitationsschwelle impliziert, da keine Normalisierung der Vektorlängen erfolgt (vgl. 5.1), dass Artikel mit sehr vielen Zitationen, die eine größere Wahrscheinlichkeit vieler Kozitationen haben, die Verteilung dominieren.

Small und Sweeney (1985a) ersetzen deshalb die einfache Kozitationsschwelle durch den Kosinus-Koeffizienten, nachdem vorher mit dem Jaccard-Koeffizienten experimentiert wurde, der aber hoch und niedrig zitierte Artikel weniger gut ausgleiche.

Trotz der Normalisierung werden die Ergebnisse noch nicht als repräsentativ angesehen; als Reaktion auf den biomedizinische Bias schlagen Thompson und Dean 1976 (nach Small & Sweeney, 1985a) unabhängig voneinander fractional citation counting statt der einfachen Zitationsschwelle als Methode zum Ausgleich unterschiedlichen Publikations- und Zitieraufkommens vor: das Gewicht jeder Zitation wird proportional zur Gesamtzahl der Referenzen des zitierenden Artikels berechnet und die fraktionalen Werte werden dann für jeden zitierten Artikel aufsummiert.

Small und Sweeney setzen außerdem an der Kozitationsschwelle an: Trotz Normalisierungseffekt benachteilige diese Gebiete mit weniger hohem Publikationsaufkommen und langen Referenzenlisten noch zu sehr, so dass die Schlussfolgerung nahe liegt, dass die optimale Kozitationsschwelle von Gebiet zu Gebiet differiere. Small und Sweeney entwickeln das Konzept des variable level clustering: Startend von einem singulären Item wird versucht, von der niedrigsten möglichen Kozitationsschwelle aus ein Cluster zu bilden.

Wenn dieses eine definierte Maximalgröße übersteigt, wird die Schwelle schrittweise für dieses Cluster erhöht.

Durch die anfänglich sehr niedrigen Schwellen werden kleine Areale relativ vergrößert und große Gruppenbildungen durch die Anwendung der Maximalgrößen in kleinere Fragmente gespalten.

Die Anwendung einer Größenbeschränkung für Cluster erscheint willkürlich – Small & Sweeney schreiben, dass auch eine maximale Dichte denkbar wäre; die Maximalgröße wird von ihnen durch die Analogie auf den von Price geprägten Begriff der invisible colleges plausibilisiert. Sie argumentieren, auch die präzise Clustergröße sei nicht von entscheidender Bedeutung, solange die Entstehung von Makroclustern verhindert werde.

Nach Small & Sweeney (1985a) kann Price für seinen Begriff der invisible colleges, einer informellen wissenschaftlichen community, keine exakte Maximalgröße definieren, begründet aber eine Anzahl von 100 Mitgliedern damit, dass darüber hinaus die interpersonale Kommunikation schwierig bis unmöglich sei. Eine größere Anzahl würde daher eine weitere Spezialisierung nach sich ziehen.

Small und Sweeney postulieren, in empirischen Studien sei diese Zahl als mittlere Anzahl zitierender Artikel einer specialty bestätigt worden, die in etwa sechs Artikel enthalte, die im Durchschnitt 17mal zitiert würden[5].

Die Zahl von 17 Zitationen wird durch Rekurs auf diese Ergebnisse heuristisch als Integerschwellwert definiert und ein äquivalenter Fraktionalschwellwert bestimmt (S. ANHANG A. HEURISTISCHE BESTIMMUNG DER SCHWELLWERTE VON SMALL & SWEENEY UND RESULTATE DER VERGLEICHSSTUDIE).

Die empirische Evaluation der Methoden und Schwellen, in der variable level und constant level clustering jeweils mit der Anwendung eines Integer- und eines fraktionalen Eingangs-Schwellwertes kombiniert und die Resultate der vier Kombinationen verglichen werden, zeigt, dass variable level-Cluster im Durchschnitt größer als constant level-Cluster sind (kleinere Cluster wachsen

5 „In particular the clustering of annual SCI files consistently gave mean cluster sizes in the size of 100 citing authors per specialty. (This roughly translates to about six highly cited works per cluster cited an average of 17 times each, and assumes that the number of citing papers assigned to a cluster is a rough measure of the size of an invisible college„ Small & Sweeney, 1985a

durch die niedrigen starting levels stärker an) und Integer-Cluster, egal ob variable oder constant, etwas größer und dichter als Cluster, die auf fraktionalen Schwellwerten basieren. In der absoluten Anzahl der zitierten Dokumente, die in Clustern gebunden sind, führt jedoch die fraktionale Zählweise zu jeweils höheren Werten als die Integer-Zählweise, sie generiert mehr Cluster.

Während sich als Vorteil des variable level clusterings wie erwartet die Vergrößerung des recalls bei gleichzeitiger Kontrolle der Clustergrößen bestätigt, wird der Anteil der Biomedizin in der Repräsentation aber durch das Verfahren nur um 2% verringert. Der Anteil der fraktionalen Schwellwerte an der Lösung des biomedizinischen Bias ist dagegen größer: Die fraktionale Eingangsschwelle sorgt für eine Reduktion der Biomedizin um 10%, der Chemie um 2% und der Physik um 5%. Small und Sweeney argumentieren, dass der Effekt des variable level clustering auf disziplinäre Biase bei niedrigeren starting levels größer wäre, wenn der Bereich niedrig zitierter Kozitations-Paare mit größerer Diversität erschlossen würde.

Als potenzielles Problem der fraktionalen Schwellen begreifen Small und Sweeney eine zu große Gewichtung für kurze Referenzenlisten und schlagen deshalb einen Integer-cut off vor, der zu einer Trunkierung der Referenzenlisten führt. Bevor die fraktionalen Werte berechnet werden, werden die niedrig zitierten Referenzen aus den Referenzenlisten gestrichen, was u.a. auch mit der verbesserten Performance begründet wird (zu den Schwellwertempfehlungen im Detail S. ANHANG B. SCHWELLWERTEMPFEHLUNGEN VON SMALL & SWEENEY).

3.3. Clusterung von Clustern

Auf der Suche nach einer verschachtelten Struktur, die einerseits der unterstellten hierarchischen Struktur der Cluster entspricht und andererseits die Komplexität der map reduziert, entwickelte das ISI (Small et al, 1985b) eine Prozedur der Clusterung von Clustern. Aufsetzend auf den Resultaten des variable level clustering als erster Stufe werden in den nächsten Schritten jeweils iterativ die Cluster des vorherigen Schrittes im variable

level-Modus, also anhand einer Kosinus-normalisierten Koziationsschwelle, einem Inkrement-Wert und einer konstanten Maximalgröße von 60 geclustert. Diese Struktur enthält insgesamt vier Hierarchiestufen, von denen die vierte nur noch ein Cluster enthält und die dritte (C3) 57 Cluster.
Bezüglich des biomedizinischen Bias führt die Clusterung von Clustern zu einer Reduktion des Anteils biomedizinischer und biochemischer Cluster von 61% auf der ersten Stufe (C1, identisch mit dem Output der fractional/variable level-Prozedur) zu nur noch 47.3% auf der dritten Iterationsstufe (C3), ebenso sinkt der Physikanteil von 20.8% auf 17.6% - soweit es möglich war, den Makrostrukturen auf dieser Stufe Disziplinen zuzuordnen, dies wurde z.T. aufgrund von Dominanzen entschieden. Die Reduktion der Biochemie und Physik hängt zwar damit zusammen, dass einige Makrocluster auf dieser Ebene eine große Anzahl kleiner Cluster dieser Disziplinen aus vorhergehenden Iterationen in sich vereinigen, doch empfinden Small & Sweeney die finalen Proportionen der Clusteranzahl je Disziplin als die ausgewogenste Repräsentation aller bisherigen cluster- und mapping-Experimente, da im Originaldatenset des SCI-Jahrgangs Biochemie auf etwa 50% der Artikel geschätzt wurde.
Abschließend resümieren Small und Sweeney, dass durch die Methodenkombination fraktionaler Schwellwert, variable level clustering und der Clusterung von Clustern die Fähigkeit des umfassenden mappings durch Ausbalancierung der Repräsentation stark verbessert worden sei. Als persistierendes Problem, das in Zukunft gelöst werden sollte, wird die Anzahl der auf den jeweiligen Stufen unverbunden gebliebenen Items genannt. Sie schlagen u.a. niedrigere starting levels und größere Maximalgrößen vor.

3.4. Kritik

Small (1993) resümiert, dass single linkage lose, instabile Cluster generiere, complete linkage jedoch nur maximal verlinkte, isolierte Muster und dass die soziologische Theorie nahe liege, die linkage-Methode von Fachgebiet zu Fachgebiet zu variieren. Ein generelles methodologisches Problem der Zitations-Clusteranalyse besteht, wie schon erwähnt, darin, dass, abgesehen

vom grundlegenden Axiom, dass Wissenschaftler benutzte Arbeiten zitieren müssen, um überhaupt wissenschaftlichen Standards zu genügen und publizieren zu können, keine echten Gesetzmäßigkeiten für Zitationsgraphen erwiesen sind. Die Modellierung einer Clusterrepräsentation ist daher zwangsläufig nicht völlig objektivierbar. Paradigmatisch für problematische Modellierungsaspekte ist die gesetzte Maximalgröße, die ab Anfang der 80er Jahre als Korrektiv des single linkage-Verfahrens mit dessen Verkettungstendenz und großen, losen Clusterbildungen verwendet und tendenziell zirkulär durch die invisible colleges-Analogie begründet wird, während andererseits erst herauszufinden ist, welches die natürlichen Einheiten der Wissenschaft sind. Small (1993) leitet später ihre Rolle über die These her, dass Wissenschaftsfelder in Konkurrenz zueinander stehen und die Maximalgrößen diese Konkurrenz modellieren, indem die schwächer verlinkten Areale eines zu großen Clusters über die erhöhte Kozitationsschwelle ausgeschlossen werden. Dieser Begründungsversuch ist m. E. nicht befriedigend, da es sich bei ihm nur um eine metatheoretische Analogie handelt, die den massiven methodischen Eingriff – wieso kann es keine größeren Cluster als solche mit 60 Mitgliedern geben bzw. wieso ist die konkrete Maximalgröße andererseits sogar egal – nicht legitimieren kann.

Aus der Perspektive jährlicher Vergleiche beobachtet Small (1993) eine strukturelle Oszillation zwischen Expansion und Kontraktion der Disziplinen, gibt aber andererseits zu, dass dieses Phänomen zumindest teilweise aus der Sensitivität der Methode in Bezug auf initiale Zustände resultieren könnte – eine immanente Schwäche hierarchischer Clusterverfahren – was insbesondere aus dem Grund, dass Strukturen nach einem Wechsel häufig wieder in den vorherigen Zustand zurückkehren, nahe liegend wirkt. Er hält das Ziel des science mappings mit den bisherigen Arbeiten deshalb nur für sehr partiell erreicht.

Es ist m. E. zudem als eine Schwäche zu sehen, dass erst die finalen Proportionen nach der cluster of cluster-Prozedur als annähernd repräsentativ gelten können – wenn auch Disproportionalitäten vor allem im Verhältnis der Disziplinen zueinander auftreten und dies für die Binnenstruktur deshalb keine Auswirkungen haben muss. Der

disziplinenübergreifende Vergleich auf den vorherigen Levels ist aber durch die Disproportionalitäten tangiert. Die coverage ist – bezogen auf die Größe des SCI-Jahrgangs – sehr gering.

Die Praxis der Größenbeschränkung für Cluster ist, wenn Small und Sweeney (1985a) auch diese Weise der Modellierung zu begründen versuchen, ein massiver Eingriff und scheint ungeeignet, insb. wenn die quantitativen Eigenschaften der Clusterdistribution im Fokus stehen. Zudem ist die Deduktion der Schwellwerte im Allgemeinen nicht durchgehend transparent.

4. Weitere Methoden

4.1. Agglomerativ-hierarchische Clusterverfahren

Bei agglomerativ-hierarchischen Verfahren gehen alle Objekte als einzelne Partitionen in den Algorithmus ein und werden anhand eines kontinuierlich ansteigenden Distanzmaßes in unterschiedlicher Reihenfolge zusammengeführt, bis am Ende nur noch ein Cluster existiert.

Der Verlauf der Clusterprozedur wird in einem Dendrogramm dargestellt. Dendrogramme sind, obwohl häufiger als finale Resultate präsentiert, uneindeutig durch den Mangel an einer definitiven Distribution. Um diese herzustellen, also die finale Clusteranzahl zu bestimmen, wird in den meisten Fällen anhand des Dendrogramms eine sinnvolle, stabile Gruppenbildung ausgewählt; es kann auch die Heterogenitätsentwicklung (Fehlerquadratsumme) der einzelnen Clusterdistributionen zu Hilfe genommen werden. Heterogenitätsentwicklung und Clusteranzahl können in einem Diagramm aufgetragen werden, um einen elbow, der einen sinnvollen cutting point anzeigt, auszumachen (Backhaus et al., 2003) oder analytische stopping rules anzuwenden (Bortz, 2005). Dieses Struktogramm ähnelt dem *scree test* bei der *Faktorenanalyse*. Daneben existieren noch weitere analytische Verfahren (Bortz, 2005).

Die Verfahren unterscheiden sich nach der Verlinkungsart; die wichtigsten werden im Folgenden vorgestellt. Beim *complete linkage*-Verfahren wird die größte Einzeldistanz zwischen Mitgliedern zweier Cluster zur Distanz zwischen den beiden Clustern, d. h. es werden nacheinander diejenigen Cluster fusioniert, deren größte Einzeldistanz im Vergleich am geringsten ist. Die Formel lautet (*r* und *s* sind Cluster, n_r die Anzahl der Objekte in Cluster *r*, x_{ri} das *i*-Objekt in Cluster *r*):

$$d(r,s) = \max(dis(x_{ri}, x_{sj}))$$

Beim *single linkage* entscheidet umgekehrt die kleinste Einzeldistanz über die Fusionierung:

$$d(r,s) = \min(dis(x_{ri}, x_{sj}))$$

Bei beiden ist die Wahl des Distanzmaßes freigestellt. Das Grundverfahren Smalls ist ein nicht-hierarchisches single linkage-Verfahren. Im Unterschied zum hierarchischen Verfahren, wo das Distanzmaß von null ansteigt und sich auf unterschiedlicher Höhe Cluster bilden, wird die Höhe der Schwellwertes bei Small vor Beginn der Clusterung auf einen heuristisch ermittelten Wert festgelegt.

Während *complete linkage* in der Literatur (Backhaus et al., 2003) eine dilatierende Clusterbildung, d. h. die Neigung zu gleich großen, kompakten, tendenziell kleinen Clustern zugeschrieben wird, tendiert das hierarchische single linkage-Verfahren umgekehrt zu einer kontrahierenden Clusterbildung, das impliziert die Neigung zur Kettenbildung und als Resultat wenige große und viele kleine Cluster.

Weitere agglomerativ-hierarchische Verfahren sind *median linkage*, *centroid linkage, average linkage* und als Sonderform der *Ward*-Algorithmus, der in vielen Grundlagenstudien gut bewertet wurde (Bortz, 2005).

Median, *centroid*, *average* und der *Ward*-Algorithmus liegen in der Clusterbildung zwischen den beiden genannten Polen, wobei der *Ward*-Algorithmus *complete linkage* nahe kommt und zu gleichmäßig großen, eher kleinen Clustern neigt.

Beim *median linkage* werden die Cluster fusioniert, deren (quadrierter) euklidischer Zentroidabstand am geringsten ist:

$$d(r,s) = d(\bar{x}_r, \bar{x}_s)$$

Eine Abwandlung dessen ist centroid linkage, bei dem die euklidischen Abstände gewichtet werden. Clusterzentroiden werden aus den arithmetischen Mittelwerten der Merkmalsausprägungen aller Objekte eines Clusters gebildet, sie befinden sich (graphisch) also in der Mitte des multidimensionalen Raums.

Der *Ward*-Algorithmus fusioniert diejenigen Paare bzw. Cluster, deren Zusammenlegung zum geringsten Anstieg der Fehlerquadratsumme, eines Heterogenitätsmaßes, führt.

Für die Fehlerquadratsumme wird die Differenz zwischen Beobachtungswerten und Mittelwerten berechnet und im Quadrat für alle Cluster aufsummiert.

In der Prozedur werden die Cluster mit den kleinsten (gewichteten) quadrierten euklidischen Distanzen zwischen Clusterschwerpunkten fusioniert, was gleichbedeutend ist mit der Minimierung der Fehlerquadratsummen-Zuwächse (Backhaus et al., 2003, Bortz, 2005). Aufgrund des konkreten Prozesses – wenn zwei Clusterpaare identische Distanzen haben, werden die Cluster fusioniert, deren Besetzungszahlen die größere Varianz aufweisen und wenn dieses Verhältnis konstant ist, werden diejenigen Cluster fusioniert, deren Gesamtumfang kleiner ist – bildet der Algorithmus bevorzugt kleine Cluster in Regionen mit hoher Objektdichte. Fortschreitend wird dies ausgeglichen, was bedeutet, dass tendenziell gleich große Cluster gebildet werden (Bortz, 2005). Die positive Evaluierung des *Ward*-Algorithmus in Monte-Carlo-Studien bezieht sich nur auf euklidische Distanzen als Distanzmaß.

Average linking gilt für Korrelationen als Distanzmaß als ähnlich erfolgreich wie *Ward* für euklidische Distanzen (Bortz, 2005), es ist aber im Unterschied zu den drei vorigen freier in der Wahl des Distanzmaßes. Es wird die durchschnittliche Objektdistanz des gewählten Distanzmaßes zwischen allen Paaren zweier zu fusionierender Cluster berechnet und dann die Cluster fusioniert, bei denen dieser Wert am geringsten ist.

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dis(x_{ri}, x_{sj})$$

Newman und Girvan (2004b) wenden sich prinzipiell gegen agglomerativ-hierarchische Verfahren mit dem Argument, dass diese in der Praxis häufig daran scheiterten, in Netzwerken, deren Strukturen bekannt sind, die korrekten Gruppierungen zu finden. Sie argumentieren auch, dass agglomerativ-hierarchische Verfahren zwar die Kerne von communities identifizierten, aber die peripheren Objekte, die oft nur schwach – graphanalytisch ausgedrückt: mit einem oder sehr wenigen links – mit einer community verbunden sind, vernachlässigt würden – ein Befund, der m. E.

plausibel ist. In der Tat ist dies im *Girvan-Newman-Algorithmus*, dem edge removal anhand von betweenness-Werten, anders gelöst. Bei agglomerativ-hierarchischen Verfahren liegt die Lösung für peripherere Items, geht man davon aus, dass die Schnittlinie zur Bestimmung der Clusteranzahl in der Regel eher in der Mitte des Dendrogramms angesetzt wird, tendenziell eher darin, dass sie nur teilweise größeren Clustern zugeordnet werden, sondern statt dessen Mikro-Cluster bilden oder unverbunden bleiben.

4.2. Partitive Clusterverfahren

Bei partitiven Verfahren werden die gesamten Daten in eine Startgruppierung, d. h. eine Distribution von Objekten zu einer festgelegten Zahl von Clustern, gebracht, die durch schritt weises Verschieben der Objekte zu Clustern verbessert werden soll.
Es ist daher ein Vorteil partitionierender Verfahren, dass sie keine Reduktion der Objektmenge, auch nicht durch singuläre Items, implizieren. Nachteilig ist dagegen, dass die Anzahl der Cluster a priori vom Benutzer vorgegeben werden muss (vgl. auch Wang et al., 2005). Insbesondere für den Fall, dass die quantitativen Eigenschaften der Distribution selbst – wie Anzahl und Größen der Cluster – im Fokus der Untersuchung stehen, wie es bei der vorliegenden der Fall ist, sind partitionierende Verfahren daher a priori weniger geeignet. Die beiden Verfahren *K-Means* und *Self-Organizing Maps* sind Beispiele partitionierender Verfahren, die im Hinblick auf die erörterten Eigenschaften nur in relativer Kürze theoretisch umrissen werden sollen.

4.3. K-Means-Algorithmus

Die Anfangspartition eines K-Means-Algorithmus besteht aus einem Vektorraum mit einer nutzerspezifisch oder zufällig festgelegten Clusterdistribution. Für alle Objekte (Datenvektoren) werden die euklidischen Distanzen zu allen Clusterschwerpunkten berechnet. Ein Objekt, das zu seinem bisherigen Clusterzentroiden eine größere Distanz aufweist als zu

einem anderen, wird daraufhin in Letzteres verschoben. Die Schwerpunkte der Cluster werden nun neu berechnet und dieser Prozess iteriert, bis sich wiederholt nur noch wenige Objekte bewegen oder sich nur noch sehr kleine Veränderungen der Zentroiden ergeben (Bortz, 2005; Chakrabarti, 2003).
Trotz des theoretischen Vorteils partitiver Verfahren gegenüber hierarchischen, dass vorgenommene Gruppenbildungen immer wieder reversibel sind, hängt das Resultat eines *K-Means*-Verfahrens häufig von der Ausgangsdistribution bzw. der Reihenfolge der verschobenen Objekte ab; d. h. die Prozedur erreicht lokale Optima, aber wegen des enormen Rechenaufwandes keine global optimale Verteilung. In der Praxis werden deshalb häufig mehrere Initialisierungen probiert und das beste Ergebnis verwendet (Bortz, 2005; Boyack et al. 2002).

4.4. Self-Organizing Maps (nach Kohonen)

Bei einer *Self-Organizing Map*[6]-Prozedur werden hochdimensionale Datenvektoren auf ein zweidimensionales Netz, das aus regelmäßig angeordneten Clustervektoren (nodes, Neurons) besteht, gemappt. K-Means prinzipiell recht ähnlich, werden für die Datenvektoren jeweils die nach der euklidischen Distanz ähnlichsten Clustervektoren berechnet und die Objekte diesen attribuiert. Es wird daraufhin ein neuer Zentroid für die Datenmenge, die dem gleichen Clustervektor assoziiert ist, kalkuliert und der ursprüngliche Clustervektor dorthin verschoben (Chakrabarti 2003; Amano, 2003). Die *neighbourhood function* – der wohl wichtigste Unterschied zu *K-Means* – wählt für jeden Datenvektor nicht nur den ähnlichsten Clustervektor, sondern alle seine Nachbarn innerhalb eines festzulegenden Radius aus, so dass, wenn ein Clustervektor verschoben wird, ebenfalls die Clustervektoren in seiner Nachbarschaft zu einem schwächeren Grad, gemäß einer Funktion, der *learning rate*, aktualisiert werden. Die Nachbarschaft wird neu kalkuliert und der Prozeß iteriert. Ein Cluster wird daher im Unterschied zu *K-Means*

[6] *SOM* ist in einigen Software Packages enthalten (*SOM Toolbox* bei *MATLAP*, *SOM Package* bei *R*, *SOM_PAC*)

nicht durch einen einzigen Clusterschwerpunkt, sondern durch mehrere Vektoren beschrieben.
Als nachteilig gilt der Mangel an generellen Nachweisen für Konvergenz und einer theoretischen Basis für die Auswahl von *learning rates* und Nachbarschaftsparameter (Börner et al., 2003), beide müssen für jedes Datenset experimentell gefunden werden.
SOM sind benutzt worden, um *nodes* im Millionenbereich zu verarbeiten (Boyack et al. 2002; Börner et al., 2003).

4.5. Faktorenanalyse

Die *Faktorenanalyse* ist ein dimensionenreduzierendes multivariates Analyseverfahren, mit dem eine Datenstruktur anhand der in ihr enthaltenen Korrelationen auf möglichst wenige hypothetische Größen – sog. Faktoren – reduziert wird. Die Faktoren repräsentieren Gruppen hoch inter korrelierender Variablen und werden so positioniert, dass die Varianz der Beobachtungswerte auf dem ersten Faktor maximal ist und die weiteren jeweils sukzessiv maximale Anteile der Restvarianz auf sich vereinen. Mit der *Faktorenanalyse* kann für ein großes Datenset eine ordnende Struktur (Bortz, 2005) bzw. eine Klassifikation konstruiert werden.
Voraussetzungen für die Faktorenanalyse sind metrische Daten (Intervall- oder Rationalskalierung, intervallskalierte Daten werden i. d. R. standardisiert), eine homogene Struktur der Daten (i. d. R. Prüfung auf Normalverteilung) und die Existenz von linearen Zusammenhängen.
Eine Modellannahme der *Faktorenanalyse* ist es, Korrelationen zwischen den Variablen kausal durch die Faktoren begründet zu begreifen (Backhaus et al., 2003, Eckey et al., 2002). Werden Faktoren als synthetische Variablen heraus partialisiert, ergeben sich Restkorrelationen, die durch die sukzessive Bestimmung weiterer Faktoren aufgeklärt werden können, aber auch durch Messfehlervarianz oder Einzelrestvarianzen bedingt sein können[7]. Die Bestimmung des Anteils an der Varianzerfassung, die tatsächlich von

7 Varianz, die nicht durch gemeinsame Faktoren aufgeklärt werden kann

gemeinsamen Faktoren geleistet wird, ist notwendig, um eine reliable Interpretation der Faktoren zu gewährleisten. Im Extraktionsverfahren einer Faktorenanalyse – wie z. B. der Hauptachsenanalyse – muss daher die aufzuklärende Varianz abgeschätzt werden.

Die *Hauptkomponentenanalyse* (Principal Component Analysis, PCA), ein anderes Verfahren zur Faktorextraktion, unterscheidet dagegen nicht zwischen Kommunalitäten und diesen Residuen, sie unterstellt, dass es keine Residuen gibt. Mit der PCA ist daher nur die im Vergleich pragmatischere dimensionenreduzierende Reproduktion der Datenstruktur und die Gruppierung der Items ohne weitergehende Interpretation intendiert (Backhaus et al., 2003).

Die oben angeführte theoretische Grundlegung der Zitationsanalyse durch Price, Marshakova und Small impliziert keine kausale Interpretation von Korrelationen bzw. Ähnlichkeiten; es gibt keine empirischen und theoretischen Anhaltspunkte zur Varianzabschätzung.

In einer PCA wird das Koordinatensystem so gedreht, dass die Punktwolken der Beobachtungswerte auf den neuen Achsen sukzessive maximale Varianz haben und die neuen Achsen nicht miteinander korrelieren.

Faktorwerte kennzeichnen die Positionen der Ursprungsvariablen auf diesen neuen Achsen, Faktorladungen entsprechen den Korrelationen zwischen den Faktorwerten und den ursprünglichen Variablenwerten. Sie geben also an, wie stark ein Faktor mit einer Ursprungsvariablen korreliert. Die Kommunalität einer Variablen gibt an, wieviel der Varianz der Variable durch alle Faktoren aufgeklärt wurde (Bortz, 2005).

Die Anzahl der Faktoren ist zunächst gleich der Anzahl der Variablen. Der Eigenwert (Summe der quadrierten Ladungen der Variablen) eines Faktors gibt an, wie viel der Gesamtvarianz aller Variablen durch diesen erfasst wird. Nach dem *Kaiser-Guttman*-Kriterium wird jeder Faktor mit einem Eigenwert kleiner oder gleich 1 weggelassen. Bei großen Variablenmengen werden nach Bortz (2005) bei diesem Kriterium jedoch zu viele unbedeutende Faktoren extrahiert; alternative Kriterien sind die Parallelanalyse oder der *Scree Test*.

Die Faktorladungsstruktur, die als Resultat der PCA vorliegt, ist dadurch geprägt, dass es auf den ersten Faktoren viele hohe Ladungen gibt und auf allen anderen viele mittlere und niedrige, was die Interpretation erschwert. Sie wird daher vorwiegend nur zur Bestimmung der Anzahl der bedeutsamen Faktoren benutzt, die dann beliebig rotiert werden können, um die Varianz umzuverteilen. Nach dem Theorem der Einfachstruktur wird mit Hilfe der Rotation versucht, eine Struktur zu erreichen, in der auf allen Faktoren einige Variablen sehr hoch und einige niedrig sowie auf verschiedenen Faktoren verschiedene Variablen hoch laden, so dass die Faktoren jeweils nur mit einer begrenzten Anzahl von Variablen korrelieren (Bortz, 2005) bzw. Variablen, die vorher auf mehreren Faktoren mittlere Ladungen hatten, eindeutig Faktoren zugeordnet werden. Eine verbreitete analytische orthogonale Rotation zur Erlangung der Einfachstruktur ist die Varimax-Methode nach Kaiser.

Zur Interpretation der Faktorenstruktur existieren Richtwerte, die darüber Auskunft geben, ab welcher Ladungshöhe Variablen einem Faktor zugeordnet werden (S. ANHANG C. RICHTLINIEN ZUR FAKTORINTERPRETATION).

Die begrenzte Kapazität von SPSS[8] – schränkt die Nutzbarkeit von Faktorenanalyse auf sehr kleine Datenmengen ein.

Problematisch ist, dass die durch eine PCA erreichte Klassifikation nicht prinzipiell disjunkt ist, wenn auch eine Orientierung an der Einfachstruktur durch Rotation und Kriterien zur Faktorinterpretation überschneidende Ladungen idealer weise verhindert werden.

Die Bestimmung der Faktorenanzahl durch den Nutzer wird durch methodische Richtlinien unterstützt, bleibt aber in der Entscheidung, welches Kriterium oder ob ein externes Kriterium angewandt wird, ein subjektiver Akt.

Wie noch im Zusammenhang mit der Eignung des Korrelationskoeffizienten diskutiert werden wird, sind lineare Beziehungen nicht optimal als Ähnlichkeitsindikator für Zitationsanalysen. Auch ist eine homogene bzw. Normalverteilung vorausgesetzt; diese ist jedoch bei bibliometrischen Daten i. d. R. nicht gegeben (vgl. 5.2.3).

8 die Faktorextraktion in den SPSS-Versionen 11 und 12 ist auf etwa 3600 Variablen beschränkt

Die Kriterien zur Faktorinterpretation, die signifikante von zufälligen Korrelationen unterscheiden sollen, verhindern als Nebeneffekt kleine Klassen, also Faktoren mit sehr wenigen Ladungen – ein Implikat, das sich durch die Zitationsdaten selbst und das Ziel, die Partitionalisierung eines Fachgebietes anhand dieser, nicht klar empirisch legitimieren lässt – a priori spricht nichts dagegen, dass auch eine Entität von zwei Artikeln eine inhaltlich abgrenzbares Thema repräsentiert.

Eine exakte Analyse der Resultate von PCA-Anwendungen in Zitationsstudien ist dadurch erschwert, dass die Richtlinien zur Faktorinterpretation (in bezug auf wenige, niedrige und überschneidende Ladungen) bei PCA-Anwendungen in Zitationsstudien oft nicht eingehalten werden. Eine Zählweise, bei der alle Faktorwerte > 0 gezählt werden, führt, wie die Journal-Studie Leydesdorffs (2006) zeigt, auch bei vorheriger Anwendung eines Schwellwertes zu stark überschneidenden Ladungsstrukturen, jedes Journal wird in dieser Studie im Schnitt 3 Faktoren zugeordnet[9]. In einer weiteren Journal-Studie zählt Leydesdorff (2004b) in einem pragmatischen Umgang mit den Richtlinien zu jedem Faktor diejenigen Journale mit der primären Ladung auf dem jeweiligen Faktor, wodurch Überschneidungen vermieden werden.

Die Größenverteilungen der Cluster werden von Leydesdorff als normalverteilt (Leydesdorff, 2006). Dies ist nicht korrekt[10]; sie weisen aber offenbar, wie es z.B. in der oben genannten Untersuchung, der Faktordistribution der Hauptdimensionen des *Social Science Citation Index* 2001, der Fall ist, eine Streuung ähnlich klein der einer Normalverteilung auf. In der genannten Faktordistribution sind 5 von 18 bzw. 72.23% der resultierenden Clustergrößen innerhalb der Standardabweichung zum Mittelwert[11].

9 bei der Gesamt-Rotationslösung wird jedes (zitierende) Journal im Durchschnitt 3.3 Faktoren zugeordnet, in der hierarchisch untergeordneten Chemie-Rotationslösung 2.5 und in der Biochemie-Rotationslösung 2.9 Faktoren

10 Median, Modalwert und Mittelwert fallen in fast allen Fällen nicht zusammen, kein glockenförmiger Verlauf

11 bei Normalverteilungen sind es 68%

4.6. Visualisierungen durch Force-directed Placement

Im *Force-directed Placement* werden Variablen einem System von Objekten und Federn (springs) mit repulsiven und anziehenden Kräften zwischen den Objekten analogisiert. Alle Vertizes (d. h. Objekte im Netzwerk) werden anfänglich zufällig auf einer Fläche verteilt und diejenigen, die eine Verbindung zueinander haben – Kanten oder edges im graphanalytischen Kontext – ziehen sich an. Während die Implementation von Eades (1984) repulsive Kräfte zwischen allen Vertizes kalkuliert und anziehende nur zwischen Nachbarn, weil es nur als wichtig erachtet wird, dass direkt verbundene Vertizes nahe beieinander liegen, wird im Kamada-Kawai-Algorithmus die ideale Distanz zwischen nicht direkt benachbarten Vertizes als proportional zum *shortest path* (s. Abschnitt 5.5.2.) zwischen ihnen bestimmt.

Kamada und Karwai formulierten das Visualisierunsproblem als iterativen Prozeß der Minimierung der Gesamtenergie: Wenn die Gesamtspannung zwischen allen Objekten minimal ist, sind die Vertizes optimal positioniert.

Nach Fruchtermann und Reingold (1991) gehorcht das *Force-directed Placement* (FDP) in bezug auf ungerichtete Graphen folgenden ästhetischen Kriterien: Vertizes sollen gleichmäßig in der Ebene verteilt werden, überkreuzende Kanten minimiert, Kanten-Längen sollen vereinheitlicht und die inhärente Symmetrie reflektiert werden. Auch benachbarte Vertizes sollen – abhängig von ihrer Anzahl und dem verfügbaren Raum – nicht zu nahe platziert werden, um die Visualisierung übersichtlich zu halten.

Aus Performancegründen kalkulieren Fruchtermann und Reingold wiederum die attraktiven Kräfte nur zwischen Nachbarn und repulsive zunächst zwischen allen, im Versuch der Beschleunigung des Algorithmus zwischen allen innerhalb einer definierten Nachbarschaft eines Vertizes[12].

Die repulsiven bzw. attraktiven Kräfte zwischen zwei Vertizes sind unterschiedliche Funktionen aus der Distanz zwischen ihnen und einem frei zu bleibenden Radius um einen Vertex, *k*, der aus der Fläche und der Vertexanzahl berechnet wird:

12 2k um einen Vertex herum

$$f_a(d) = \frac{d^2}{k}$$ (attraktiv)

$$f_r(d) = \frac{-k^2}{d}$$ (repulsiv)

$$k = C\sqrt{\left(\frac{area}{number \quad of \quad vertices}\right)}$$

wobei *C* experimentell festgelegt wurde.

Es werden für jeden Vertex der Effekt der attraktiven Kräfte und der Effekt der repulsiven Kräfte berechnet und die daraus entstehende Bewegung durch einen je nach Iterationsschritt und spezifischer Methode des *cooling schedule*[13] variierenden Maximalwert limitiert. Während die Algorithmen, die die Gesamtenergie minimieren, durch Schwellwerte für diese Gesamtenergie terminieren, ist die Terminierung im *Fruchtermann-Reingold-Algorithmus* nicht durch strenge Richtlinien bestimmt.

Das *VxInsight package* enthält die *FDP*-Prozedur *VxOrd* und die Visualisierungsengine *VxInsight*. *VxOrd* akzeptiert als Input vorkalkulierte Ähnlichkeitsmatrizen oder alternativ Listen mit direktionalen Kanten, aus denen Ähnlichkeiten berechnet werden (Boyack et al., 2002)[14]. Die Funktion

$$E_{x,x} = \left[\sum_{i=0}^{n} w_i l_i^2\right] + D_{x,y}$$

bezeichnet die Energie eines Vertizes mit *n* Kanten auf einer spezifischen Koordinate *x, y*, die Gleichung wird iterativ minimiert (Boyack et al., 2002, Davidson et al., 2003).

w_i ist die Ähnlichkeit des Vertizes mit einem durch die Kante *i* verknüpften Vertex bzw. die Gewichtung der betreffenden Kante; l_i ist die euklidische Distanz zwischen dem Vertex und dem durch die Kante *i* verbundenen Vertex und $D_{x,y}$ ist ein Dichtemaß, das die Rolle einer repulsiven Kraft einnimmt.

Das Dichtemaß wird nichtspezifisch, also bezogen auf die gesamte Vertexmenge, kalkuliert, indem jeder Vertex auf einer zweidimensionalen

13 ein hoher Maximalwert am Anfang, der rapide absinkt und dann auf einem konstant niedrigen Level bleibt

14 dazu keine genaueren Angaben, zur Berechnung von Ähnlichkeit aus gewichteten bzw. ungewichteten Graphen (basierend auf Shortest Path) vgl z.B. Egghe, 2003

Fläche eine Funktion[15] hinterlässt. In die Dichteberechnung für jeden Vertex gehen diese energy footprints in einem Radius mit quadratisch abfallender Distanz ein.

Bei der Minimierung der Energie ist es einem kleinen Teil der Vertizes[16] erlaubt, den repulsiven Dichteterm zu ignorieren und zu einem gewichteten Zentroiden aller verbundenen Kanten zu springen (barrier jumping), um das frühzeitige Verfangen der Vertizes in lokalen Minima zu verhindern (Davidson et al., 2003).

Die Ersetzung der beim FDP sonst üblichen paarweisen repulsiven Energien durch das Dichtemaß, das eher eine *allgemeine Überfüllung* misst (Davidson et al, 2003), beschleunigt die Kalkulation: Es wird bei einem Graphen mit der Vertex-Menge N $O(N)$ statt $O(N^2)$ Zeit benötigt (Börner et al., 2003, Boyack et al., 2002). Wie bei anderen FDP-Algorithmen ist das Ziel die Energieminimierung des Graphen, indem die Energiewerte der einzelnen Vertizes iterativ durch immer kleinere Bewegungen im Raum verringert werden. Der Wert der Gesamtenergie wird als Abbruchkriterium benutzt (Boyack et al., 2002; Davidson et al., 2001).

Klavans (2005) postuliert als inhärenten Vorteil einer Visualisierungssoftware die Verbesserung der coverage, da die Prozedur niedrig und hoch korrelierte Items mappen könnte und automatisch die Anzahl und Größe der Cluster determinieren würde – was im Normalfall in einer Lösung resultieren würde, in der hoch korrelierte Items ineinander gravidierten und die niedrig korrelierten als isolierte Items ohne nahe Nachbarn zu sehen wären. Diese Bewertung ist problematisch, da einerseits eine FDP-Visualisierungssoftware wie VxInsight kein Clusteralgorithmus ist, also keine diskreten, sondern kontinuierliche Strukturen schafft, andererseits Cluster erst ab einer Mitgliederzahl von zwei als solche definiert werden und deshalb singuläre Items nicht teil einer Clusterdistribution sind. Mit coverage üblicherweise und in dieser Arbeit ist jedoch die Anzahl der Daten der Clusterdistribution in Relation zur Ausgangsdatenmenge gemeint.

15 von den Autoren nicht genauer definiert

16 von den Autoren nicht genauer definiert

4.7. Netzwerkanalytische Clusterverfahren

Netzwerkanalytische Verfahren mit dem Ziel der Clusterung unterscheiden sich prinzipiell von den bisher beschriebenen Clusterverfahren dadurch, dass die Clusterung nicht anhand eines Proximitätsmaßes durchgeführt wird, sondern anhand der Struktureigenschaften des Graphen: Nach unterschiedlichen Maßgaben werden dicht verlinkte Regionen von weniger dichten separiert und als Cluster definiert. Die beiden Herangehensweisen nähern sich im Fall des *single linkage* und *complete linkage* an: Die beiden *linkage*-Verfahren entsprechen, werden sie nicht hierarchisch-agglomerativ, sondern mit initialen festen Schwellwerten angewendet, den einfach verbundenen components bzw. den komplett verbundenen cliques der Netzwerkanalyse. Die anderen Verfahren wie z. B. *average* und *median linkage* dagegen sind nicht direkt übertragbar, da graphanalytischen Methoden häufig mit nicht gewichteten Kanten arbeiten.

4.7.1. Centrality

Freeman unterteilt den Begriff *centrality* in die drei basalen Kategorien *closeness*, *degree* und *betweenness* (nach Everett und Borgatti, 2005).
Der *degree* ist ein sehr simples Zentralitätsmaß, er bezeichnet die Anzahl der Nachbarn eines Vertizes und kann benutzt werden, um schwach verlinkte Vertizes zu eliminieren.
Closeness ist die inverse Summe der geodätischen shortest path-Distanzen eines Vertizes zu allen anderen (Everett und Borgatti, 2005).

4.7.2. Edge Removal durch Betweenness (Girvan-Newman-Algorithmus)

Girvan und Newman (2002) übernahmen Freemans *Betweenness-Centrality*-Konzept, das auf Vertizes bezogen war, und wenden es auf die Kanten zwischen den Verrtizes an. Kanten mit den höchsten *betweenness*-Werten verbinden maximal viele Vertizes miteinander, sie bilden Brücken zwischen dichter verlinkten Regionen. Sie sind also am wenigsten zentral für diese

Dichteregionen, die durch die Entfernung dieser Brücken zu distinkten Clustern werden.

Der hierarchisch-divisive *Girvan-Newman (GN)-Algorithmus* besteht aus dem iterativen Entfernen der Kanten mit den höchsten *betweenness*-Werten, die nach jedem Durchgang für den gesamten Graphen neu berechnet werden, da sonst nach dem ersten Löschvorgang die anfänglich berechneten Werte nicht mehr dem veränderten Netzwerk entsprechen (Girvan & Newman, 2002; Newman & Girvan, 2004).

Beim *shortest path betweenness* werden die kürzesten (geodätischen) Wege zwischen allen Vertizes ermittelt und den paths jeweils der einheitliche Wert 1 zugewiesen. Wenn es mehrere shortest paths zwischen zwei Vertizes gibt, wird jedem ein gleicher Fraktionalwert zugewiesen, so dass ihre Totalität wieder 1 ergibt. Daraufhin werden die paths, die über eine bestimmte Kante verlaufen, für diese aufsummiert und die Kante mit dem jeweils größten Wert gelöscht.

Neben dem shortest path betweenness gibt es noch die Ansätze *random walk* und *current flow*. Beim *random walk*-Verfahren wird der Erwartungswert der Häufigkeit kalkuliert, mit der zufällige Pfade zwischen allen Vertex-Paaren eine spezifische Kante treffen.

Beim *current flow*-Konzept wird jeder Kante des Netzwerkes mit einem einheitlichen Widerstand belegt. Der Strom zwischen zwei Vertizes verläuft in diesem Modell über multiple Pfade, der größte Anteil über den Weg mit dem geringsten Widerstand. Die Anzahl der Wege je Kante wird wie bei den anderen aufsummiert.

Girvan und Newman (2002) schlagen die modularity *Q* als Maß vor für die Qualität der Partitionierung einer Clusterdistribution, formuliert in der Mischungsmatrix *e*:

$$Q = Tre - \left\| e^2 \right\|$$

Die Spur *Tr* einer Matrix ist die Summe der Diagonalenelemente, also in einer Mischungsmatrix der Zellen, die jeweils die Anzahl der Kanten innerhalb von Clustern enthalten. Durch die Normalisierung der Matrix werden alle Zelleninhalte in Wahrscheinlichkeiten verwandelt: Die Zelle e_{ij} in der

normalisierten Mischungsmatrix ist dann gleich der Wahrscheinlichkeit eines Zufallslinks zwischen *i* und *j*. $\|e^2\|$ ist das Produkt dieser Wahrscheinlichkeiten, mit denen Zufallslinks jeweils in Clustern beginnen oder enden. In einem Zufallsnetzwerk wäre dieser Wert identisch mit der Spur *Tr e*.

Ein Wert *Q = 0* indiziert, dass die Gruppenstruktur nicht stärker als bei einem Zufallsgraphen ist; während das Maximum *Q = 1* eine Clusterdistribution indizieren würde, bei der überhaupt keine Verbindungen zwischen Clustern existieren und bei der die Partitionierung nach diesem Maßstab perfekt wäre. Newman gibt an, in der Praxis habe sich herausgestellt, dass Werte, die größer als *Q = 0.3* sind, signifikante community-Strukturen indizieren. Der *Q*-Wert wird nach jedem Iterationsschritt des GN-Algorithmus berechnet und peaks werden zur Bestimmung der finalen Clusterdistribution herangezogen (Newman, 2004).

Der Rechenaufwand für *edge removal* mit *shortest path* kann insgesamt bis zu $O(m^2n)$[17], im Fall eines spärlichen Netzwerkes – wie es Netzwerke wissenschaftlicher Kollaborationen i. d. R. sind – $O(n^3)$ Zeit beanspruchen (Newman, 2004b). Radicchi et al. (2004) als auch Newman (2004b) beschreiben, die Anwendung sei schon für Netzwerke mittlerer Größe undurchführbar[18].

Die drei Methoden führen in einer Untersuchung von Girvan und Newman (2002) zu tendenziell ähnlichen Resultaten; *current flow* und *random walk* stimmen in dieser Studie im Ergebnis exakt überein. Newman und Girvan (2004b) schließen deshalb, die Wahl des konkreten *betweenness*-Maßes sei zweitrangig und ohne großen Einfluss auf das Ergebnis. Sie empfehlen *shortest path* gegenüber den anderen, da diese noch rechenintensiver sind.

2005 argumentiert Newman jedoch für *random walk betweenness*, da (in Bezug auf soziale Netzwerke) *shortest path* und *current flow* in einer spezifischen konstruierten Beispielsituation jeweils kontraintuitive Ergebnisse

17 m = Anzahl der Kanten, n = Anzahl der Nodes

18 Radicchi spricht von 10 000 Vertizes

zeitigten; er kritisiert in diesem Kontext das Prinzip der idealen Routen, das den beiden Modellen inhärent ist.

Der Algorithmus wurde bisher vor allem in biologischen und sozialen Netzwerken angewandt, artikelbasierte Kozitationsstudien benötigen in der Regel. eine Datenbasis, die durch den *GN*-Algorithmus aufgrund dessen Kapazitätsbeschränkungen (s. o.) nicht bearbeitet werden kann. In artifiziellen und real world-Beispielen, häufig kleinen Organisationsnetzwerken, deren Struktur bekannt ist, führte der Algorithmus zu einer sehr gut übereinstimmenden Strukturerkennung.

4.7.3. Schnellere Alternativen zum GN-Algorithmus

Radicchi et al. (2004) schlagen einen fast algorithm, einen alternativen Ansatz zum langsamen GN-Algorithmus vor, der mit größerer Schnelligkeit[19] zu vergleichbar akkuraten Ergebnissen kommen soll: Statt repetitiv für jede Kante die *betweenness*, bezogen auf das ganze System zu berechnen, sieht ihr Vorschlag einen divisiven Algorithmus vor, der, indem der Clusterkoeffizient auf Kanten bezogen wird, nur mit lokalen Quantitäten arbeitet.

Es wird die Anzahl der zyklischen Strukturen, an denen eine Kante partizipiert – in der Grundversion die Anzahl der Dreiecke – dividiert durch die Anzahl der zyklischen Strukturen bzw. Dreiecke, zu denen er potenziell gehören könnte. Dieser hypothetische Wert wird durch den *degree* der zwei adjazenten Vertizes bestimmt:

$$c_{ij}^{(3)} = \frac{z_{ij}^{(3)} + 1}{\min\left[\left(k_i - 1\right)\left(k_j - 1\right)\right]}$$

Kanten, die unterschiedliche communities verbinden, sind nach Radicchi et al. in nur wenigen Dreiecken involviert und haben deshalb einen niedrigen Clusterkoeffizienten, während innerhalb von Clustern viele Dreiecke existieren. Ähnlich wie beim GN-Algorithmus wird repetitiv die Kante mit dem geringsten Clusterkoeffizient-Wert gelöscht. Als Kriterium zur Bestimmung

19 Newman (2004) schätzt O(m4/n2 oder O(n2)

der Clusteranzahl wird die Definition von weak bzw. strong communities herangezogen: In einer strong community muss jeder Vertex eines Clusters mehr Verbindungen innerhalb des Clusters haben als nach außerhalb; in einer weak community muss nur die Summe der Kanten nach innen größer als die nach draußen sein.

Da im Fall einer Situation, in der null Dreiecke realisiert sind, die Berechnung immer zu einem Ergebnis = 0 führte, wird der Zähler des Terms durch +1 modifiziert. Die verallgemeinerte Formel lautet dann:

$$c_{ij}^{(g)} = \frac{z_{ij}^{(g)} + 1}{s_{ij}^{(g)}}$$

mit $z_{ij}^{(g)}$ gleich der Anzahl der möglichen zirkulären Strukturen. Statt der Dreiecke arbeiten Radicchi et al. auch mit Rechtecken.

Für den Fall, dass ein Vertex nur durch eine Kante mit anderen Vertizes verbunden ist, ergäbe sich eine Division durch null, die formal nicht definiert ist, von Radicchi et al. aber plausibel als Operation mit dem Ergebnis unendlich behandelt wird, so dass – analog dem *GN*-Algorithmus – ein Vertex mit einem einzigen Kante nicht abgetrennt werden kann.

Radicchi et al. vergleichen die Resultate ihres Algorithmus` mit denen des GN-Algorithmus und analysieren, dass die Resultate sehr ähnlich seien, wenn auch die Antikorrelation zwischen *edge betweenness* und dem Clusterkoeffizienten des *Radicchi-Algorithmus* nicht vollkommen ist.

Newman (2004) merkt an, dass Untersuchungen zu unterschiedlichen Resultaten bezüglich der empirischen Häufigkeit von Dreiecken in real world-Netzwerken gekommen seien. Er behauptet, dass die Häufigkeit von triangulären Strukturen in sozialen Netzwerken sehr hoch, in anderen aber eher niedrig sei.

Es bleibt festzuhalten, dass der Algorithmus zwar die Funktionsweise des GN-Algorithmus im Prinzip simuliert, das Clusterkriterium jedoch etwas willkürlicher ist.

Newman (2004b) schlägt auch eine mit lokalen Quantitäten arbeitende Alternative zum GN-Algorithmus mit ebenfalls besserer Performance (*O((m+n)n)* oder $O(n^2)$) vor. Das modularity-Maß *Q* wird für einen

agglomerativ-hierarchischen Algorithmus genutzt, bei dem einzelne Vertizes bzw. Cluster zusammengeführt werden und auf jeder Stufe die Fusionierung von Paaren, die den größten Anstieg von *Q* bringt, realisiert wird. Dieser Prozess wird in einem Dendrogramm dargestellt und die definitive Partitionierung kann durch einen cut gemäß dem größten auf das Gesamtsystem bezogenen *Q*-Wert vorgenommen werden.

Nach den Angaben Newmans (2004b) hat der GN-Algorithmus in der Praxis häufig bessere Ergebnisse als der fast algorithm durch den Vorteil der Nutzung nicht-lokaler Informationen über das gesamte Netzwerk, er sieht den fast algorithm aber als gute Option für große Netzwerke.

Beide Algorithmen errechnen für wissenschaftliche Kollaborationsnetzwerke[20] vergleichbarer Größenordnung des gleichen Fachgebietes (etwa 12722 bzw. 9350 Vertizes) Clusterdistributionen, die einer Potenzfunktion mit dem Exponenten 2 entsprechen.

4.7.4. Bi-connected components

Bi-connected components sind Cluster mit einer minimalen Größe von drei Vertizes ohne einen cut-vertex, d. h. sie enthalten keinen Vertex, dessen Entfernung zu einer Spaltung des Clusters führen würde. Jeder Vertex muss deshalb zwingend mit mindestens zwei anderen Vertizes verbunden sein (de Nooy et al., 2005).

Nach Leydesdorff garantiere die bi-connected-Definition, dass die großen bi-connected Components durch punktuell ausfallende oder hinzukommende Vertizes nicht sehr stark beeinflusst werden (Leydesdorff, 2004b).

In der Clusteranalyse werden die bi-connected components extrahiert und alle Vertizes mit ihren Kanten, die der Anforderung nicht entsprechen, entfernt.

20 Radicchi et al. bestimmen die Kollaboration von Autoren, die in arXiv-cond-mat (Festkörperphysik) zwischen 1995 und 1999 mindestens einmal publiziert haben; Newman nimmt ohne Zeitbeschränkung die Autoren aller Branchen in arXiv, die durch den Algorithmus zunächst in 600 communities gespalten werden. 77% dieser sind in vier communities konzentriert, die mit Hochenergiephysik, Astrophysik und zweimal Festkörperphysik identifiziert werden. Die schmalere Festkörperpartition wird weiter geclustert und generiert die genannte power law-Verteilung

Vertizes, die als cut-Vertizes zwei bi-components miteinander verbinden, sind als articulation points definiert und können nach Leydesdorff hierarchisch als layer einer übergeordneten Ebene interpretiert werden (Leydesdorff, 2004a).
In seiner Studie zur Dekomposition eines Journalsets durch bi-connected components untersucht Leydesdorff (2004a) die citing-Dimension eines aus 5739 Journalen bestehenden Datensets aus den Journal Citation Reports 2001. Er unterzieht es zunächst ohne Schwellwert einer bi-connected component-Analyse. Dies führt zu einer sehr extremen Clusterdistribution mit einem Makrocluster, das fast die gesamte Journalmenge (5713) enthält sowie einigen sehr kleinen Clustern.
Nach Anwendung eines Schwellwertes (Pearson`s r $\geq$ 0.8) wird die Journalmenge auf 73% des ursprünglichen Sets bzw. 3991 Journals reduziert. Die Dekomposition anhand des bi-connected-Ansatzes führt zu einer weiteren Reduktion der *coverage* auf nun 59.9% und einem Zuwachs der Clusteranzahl auf 222. Die Distribution ist dennoch weiterhin extrem schief mit einem dominierenden Makrocluster (1417 Journals) und vielen kleinen, das nächst kleinere Cluster umfasst nur 83 Journale.
Das größte Cluster wird in einem variable level-Ansatz weiter zerlegt durch Erhöhung von Pearson's r auf $\geq$ 0.9 sowie $\geq$ 0.95.
Die 55 articulation points sind dünn über das gesamte Netzwerk verstreut und indizieren laut
Leydesdorff keine zweite Ebene, da sie zuwenig bi-connected components zusammenfassen. Unklar bleibt aber, was sie nach Leydesdorff im umgekehrten Fall genau aussagen würden.
In einem zweiten Schritt wird eine Minimalgröße von 10 Journalen für bi-connected component*s* gesetzt; dies führt zu 62 bi-connected components mit insgesamt 2726 Journalen (47.4 %, der Gesamtmenge), die dann sehr homogen sind.
Die Untersuchung Leydesdorffs zeigt auf, dass der *bi-connected*-Ansatz zur Clusterung ohne Anwendung eines Schwellwertes untauglich ist. Mit einer zusätzlichen Schwellwertanwendung sind die Ergebnisse des schwächer definierten *single linkage clustering* strukturell ähnlich.

4.7.5. Clusterdichte (Gmür)

Gmür (2004) definiert Cluster als Strukturen, die entweder wenigstens eine komplett verlinkte Gruppe aus drei Items oder eine sternförmige Formation aus fünf Items enthält. Er schlägt ein durch die Clusterdichte definiertes Verfahren vor, bei dem vom Clusterkern aus die nach unterschiedlichen Ähnlichkeitsmaßen 200 jeweils am höchsten korrelierten Items kontinuierlich in ein Cluster integriert werden, so lange die Clusterdichte nicht abnimmt. Die Dichte berechnet er durch die Differenz zwischen in-degree und out-degree, dividiert durch die Clustergröße. Er wendet den vorgeschlagenen Algorithmus in einer Artikel-Mikrostudie im Fach Organization Science zum Vergleich unterschiedlicher Ähnlichkeitsmaße sowie der Faktorenanalyse an. Während die angewandten Ähnlichkeitsmaße und ein initialer Integer-Schwellwert ≥ 42 (eine Referenz wird von mindestens 2% der Artikelmenge zitiert) die ursprüngliche Datenmenge der Referenzen im Schnitt auf etwa 65% reduzieren, liegt die *coverage* in Clustern infolge des graphanalytischen Clusterverfahrens im Durchschnitt bei 50%[21], das ist m. E. eine zu starke Reduktion.

21 43.3%; 59.3%; 48.5%;49.5%

5. Ähnlichkeitsmaße

5.1. Allgemeines

Nach Dominich (2001) haben Ähnlichkeitsmaße im Vektorraummodell die drei typischen Eigenschaften Symmetrie, Reflexivität und Normalisierung.
Symmetrie bedeutet, dass die Reihenfolge der beiden zu vergleichenden Vektoren keinen Unterschied macht; Normalisierung, dass das Maß Werte zwischen null und eins annimmt und Reflexivität, dass der maximale Wert eins erreicht wird, wenn die beiden Vektoren identisch sind.
Dominich nennt folgende Ähnlichkeits-Maße (in mengentheoretischer Schreibweise für binäre Vektoren):

Skalarprodukt der beiden Vektoren (*inner product* oder *simple matching coefficient*)

$$|Q \cap D|$$

Saltons Kosinus-Koeffizient

$$\frac{|Q \cap D|}{|Q|^{\frac{1}{2}} \times |D|^{\frac{1}{2}}}$$

Dice-Koeffizient

$$\frac{2(|Q \cap D|)}{|Q| + |D|}$$

Jaccard-Koeffizient

$$\frac{|Q \cap D|}{|Q \cup D|}$$

Overlap-Koeffizient

$$\frac{|Q \cap D|}{\min(|Q|,|D|)}$$

Alle Maße enthalten im Zähler die den beiden Vektoren gemeinsamen Eigenschaften (Kozitationen im Fall der Kozitationsanalyse) und – bis auf den Simple Matching Koeffizienten – im Nenner einen variierenden Normalisierungsfaktor, der den Wert des Zählers anhand der Längen der Vektoren relativiert.

Außer dem *simple matching*-Koeffizienten haben alle die genannten drei Eigenschaften. Der *simple matching*-Koeffizient ist wegen der nicht vorhandenen Normalisierung nicht in der Lage, den Einfluss hoch zitierter Artikel auszugleichen, der *Overlap*-Koeffizient erscheint in der alleinigen Abhängigkeit vom niedrig zitierteren Partner eines Kozitationspaares einseitig.

Der Kosinus-Koeffizient ist das am meisten verbreitete Maß unter ihnen; es gleicht laut Small und Sweeney (1985a), wie schon erwähnt, niedrig und hoch zitierte Artikel besser aus als der Jaccard-Koeffizient.

5.2. Kontroverse um Pearson`s r und Salton`s Cosine

Während Small und Sweeney (1985a), wie beschrieben, die Kozitations-Daten zunächst mit dem Jaccard-Koeffizienten und dann mit Saltons Kosinusmaß normalisierten, benutzten White und Griffith (1981) als Pioniere der Autor-Kozitationsanalysen (ACA) *Pearsons Korrelationskoeffizienten* (Pearson`scher Maßkorrelationskoeffizient, Pearson`s r), ein sehr verbreitetes Standardmaß für Intervalldaten, als Ähnlichkeitsmaß.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (xi - \overline{x})^2} \cdot \sqrt{\sum_i (xi - \overline{x})^2}}$$

Der bis heute sehr häufige Gebrauch von *Pearson`s r* als Ähnlichkeitsmaß geht auf einen allgemein als Standard akzeptierten technischen Überblicksartikel von McCain (1990) zu ACA bzw. McCains Aussagen bezüglich ihrer Präferenz von *Pearson`s r* in späteren Artikeln zurück.
Pearson`s r misst den Grad, zu dem eine lineare Funktion zwischen zwei Variablen existiert (Siegel & Castellan, 1988). Ein starker Zusammenhang, der nicht linear ist, führt deshalb zu keinem hohen Korrelationswert.
Einige Statistik-Autoren lehnten in der Vergangenheit *Pearson`s r* für Clusterzwecke pauschal ab, so z.B. Dillon und Goldstein (1984). Diese kritisieren das Faktum, dass nicht faktische Ähnlichkeit oder Identität der Werte ausschlaggebend sind, sondern, wie beschrieben, der Grad einer linearen Beziehung zwischen den Wertereihen, und empfinden das Maß wegen dieser Eigenschaft und den daraus resultierenden z.T. kontraintuitiven Resultaten (z.B. ein hoher Korrelationswert, wenn die Wertepaare zwar stark auseinander liegen, aber immer die gleiche Differenz zueinander haben) als untauglich für Clusterzwecke.
2003 lösten Ahlgren, Jarneving und Rousseau mit einem Artikel, in dem sie die Eignung von Pearson`s r für Clusteranalysen mit mathematischen Argumenten und einem Fallbeispiel in Frage stellten und als Alternativen *Saltons Kosinus-Koeffizienten*, *Chi-Quadrat* und den *Spearman-Rangkoeffizienten* vorschlugen, eine Debatte um Ähnlichkeitsmaße in Clusteranalysen aus, die als solche offenbar längere Zeit tendenziell im Hintergrund gestanden hatten gegenüber der Erschließung von Anwendungsgebieten und der Exploration neuerer Cluster- und Visualisierungsverfahren.

5.2.1. Globale und lokale Maße

Anders als die obigen Ähnlichkeitsmaße misst *Pearson`s r* nicht die direkte Ähnlichkeit zwischen zwei Autoren in Form ihrer gegebenen Falls normalisierten Kozitationsstärke (von Ahlgren et al. als lokaler Ansatz bezeichnet), sondern die Ähnlichkeit der Autoren wird in einem globalen Ansatz in Form ihrer Beziehungen zu allen anderen Autoren bestimmt, indem

jeweils alle einzelnen zentrierten Kozitationswerte zweier Autoren miteinander verglichen werden. In der ACA werden die diagonalen Werte der Zitationsmatrix häufig als fehlend behandelt, obwohl es, wie Ahlgren et al. darlegen, die bessere Lösung wäre, die tatsächliche Anzahl, wie oft ein Autor mit sich selbst kozitiert wird (d. h. wie oft zwei unterschiedliche Arbeiten eines Autors kozitiert werden, exklusive Selbstzitationen), zu verwenden, um eine mathematisch komplette Matrix zu erlangen. Im Fall der leeren Diagonale wird *Pearson`s r* zwischen zwei Autoren mit allen Kozitationshäufigkeiten außer der zwischen diesen beiden Autoren berechnet. Diese wird weggelassen, weil jeweils einer der beiden Variablenwerte im Term die leere Diagonalen-Zelle wäre. Das kann als logische Schwäche der ACA mit *Pearson`s r* angesehen werden.

5.2.2. Normalisierung

Ahlgren et al. (2003) verweisen in ihrer Erörterung der Eigenschaften von *Pearson`s r* auf Dominich (2001), dessen Postulat der Normalisierung dadurch verletzt wird, dass *Pearson`s r* Werte von –1 bis 1 annehmen kann. Da es sich jedoch nur um eine Definitionsfrage handelt – Dominich spricht von typischen Eigenschaften, die von allen außer dem von ihm ebenfalls genannten einfachen Skalarprodukt, das in dieser Form nicht normalisiert und reflexiv ist, eingehalten werden – wird hierdurch nicht die Qualität der Resultate, sondern nur die Frage, ob es sich bei *Pearson`s r* um ein echtes Ähnlichkeitsmaß handelt, berührt. Ahlgren et al. schlagen vor, jeden Wert *r* durch (*r*+1)/2 zu transformieren. Der Median wäre dann nicht mehr 0, sondern 0.5.

5.2.3. Anforderungen linearer statistischer Modelle

Lineare statistische Modelle wie der Korrelationskoeffizient setzen voraus, dass sich die Haupteffekte additiv verhalten, dass die Varianz konstant[22] ist und eine Normalverteilung vorliegt (Leydesdorff, 2005).

22 d.h. unabhängig von der Größe des Mittelwertes

Bei einer Normalverteilung müssen Median, Modalwert und Mittelwert zusammenfallen; der Graph hat eine symmetrische Glockenform und nähert sich asymptotisch der x-Achse, die er nicht berührt (Bortz, 2005). 68% der Werte liegen im Bereich der Wendepunkte – zwischen ± 1 Standardabweichungen zum Mittelwert – und 96% im Bereich zwischen ± 2 Standardabweichungen.

Bibliometrische Daten dagegen sind in der Regel hochgradig schief verteilt (vgl. z.B. Van Raan (2005), der jeweils ein annäherndes *power law*-Verhalten für die Verteilungsfunktionen von Referenzen zu Artikeln bzw. von Zitationen zu Artikeln zeigt[23]).

Weil bibliometrische Daten die Anforderungen linearer Modelle nicht erfüllen, wird in der Literatur (z.B. Ahlgren et al., 2003) häufiger eine Logarithmierung von bibliometrischen Daten vorgeschlagen. Die Originaldistribution[24] wird als Lognormalverteilung[25] behandelt, die durch die Logarithmierung einer Normalverteilung angeglichen werden und auch den Forderungen der Additivität und konstanten Varianz entsprechen soll[26].

Eine logarithmische Transformation der Daten unterminiert, da sie die ursprüngliche Varianz der Daten vermindert, jedoch die Basis der nachfolgenden Klassifikation, die ja die Daten in ihrer Heterogenität repräsentieren soll. Es liegt also nahe, dass die Qualität einer Klassifikation durch die Logarithmierung verschlechtert wird, wie Leydesdorff & Bensman (2006) auch nachweisen: In einer Journal-Studie werden 21 Journale (die für jeweils mehr als ein Prozent der Zitationen des Journal of the American Chemical Society verantwortlich sind und aus den Journal Citation Reports 2003 ermittelt wurden) einer logischen Induktions- und Analogie-Analyse nach den subject headings und class numbers und einer parallelen Faktorenanalyse (die, wie beschrieben, auf einer Pearson-Korrelationsmatrix

23 jeweils abgeschwächt im Bereich sehr niedriger Referenzen- bzw. Zitationswerte; es findet aber im Unterschied zu unserem Untersuchungssetting keine Exklusion von Datentypen wie reviews etc. statt

24 diese soll eingipflig und rechtsschief sein

25 eine Lognormalverteilung hat Ähnlichkeit mit einer power law-Verteilung und nimmt in Log-logdarstellung partiell die Form einer Geraden an (Mitzenmacher, 2005)

26 Leydesdorff & Bensman (2005) formulieren allerdings inkorrekt: „Logarithmically transformed data may exhibit log-normality, and thus allow for using the Pearson correlation coefficient„

basiert) unterzogen. Die asymmetrische Matrix der cited-Dimension[27] wird dabei einmal logarithmiert und einmal unlogarithmiert verarbeitet.
In der Faktorenanalyse zeigen Leydesdorff & Bensman, dass sich die Varianzreduktion tatsächlich in der Klassifikation auswirkt; so sinkt die Zahl der Faktoren mit einem Eigenwert größer eins von sechs auf vier ab und zwei Journale, die vorher eine separate Gruppe bildeten, gehören nach der Logarithmierung einem anderen Cluster an. Bei einer erzwungenen 6-Faktorenlösung der logarithmierten Datenstruktur verändert sich die Struktur nicht grundlegend, aber die Faktoren überlappen einander stärker als zuvor. Die Struktur des Journalsets erweist sich somit zwar als relativ robust, doch büßt sie durch die Transformation leicht an Differenzierung ein.

5.2.4. Skalierung

Ein wichtiger Unterschied zwischen dem Kosinus-Maß und dem Korrelationskoeffizienten liegt darin, dass ersteres auf den Originalwerten als Abweichung vom Ursprung, letzterer auf zentrierten Werten (Abweichungen vom Mittelwert) basiert. Das Kosinus-Maß erscheint besonders adäquat, wenn der Nullpunkt bedeutungsvoll und die Originalwerte eine absolute Bedeutung haben (d. h. wenn eine Rationalskalierung vorliegt). Wenn der Nullwert hingegen zufällig gewählt wurde und die Originalwerte nur in Relation zueinander und zu ihren Mittelwerten bedeutungsvoll (also intervallskaliert) sind, ist der Korrelationskoeffizient ein adäquateres Maß ist (Anderberg, 1973). Beide Maße sollten dementsprechend nicht bei ordinalskalierten Daten angewandt werden.
Ahlgren et al. (2003) argumentieren, dass Kozitationsdaten als ordinalskaliert betrachtet werden könnten, wenn man davon ausgehe, dass es keine absolute Null und damit keine absoluten Abstände gebe, insofern nicht mit absoluter Sicherheit, sondern immer nur aufgrund einer konkreten Datenbasis wie der ISI-Datenbank (die natürlich selbst wie auch die jeweiligen Testmengen Veränderungen unterworfen ist) eine Aussage darüber gemacht werden könne, dass zwei Autoren nicht kozitiert würden.

27 die Matrix umfasst die 21 Journale, die von allen Journals aus der Datenbank zitiert werden

Diese Annahme ordinalskalierter Daten richtet sich natürlich gegen beide Maße (vgl. Anderberg, 1973).

Ausgehend von der basalen Voraussetzung, dass die Ergebnisse einer Kozitationsanalyse nur Aussagen mit einer durch die Datenbasis bestimmten Reichweite zulassen (diese Reichweite wird einerseits durch den Zeitpunkt der Untersuchung, andererseits durch die empirischen Grenzen des SCI[28] bestimmt), so ist es m. E. nicht sinnvoll, von einer anderen als einer Rationalskala auszugehen, da die Messwerte (die Anzahl von Kozitationen) Mengen sind und als solche alle Voraussetzungen (bedeutungsvoller Nullpunkt, Multiplikation ist eine mögliche Operation) erfüllen.

Von dieser Einschätzung unabhängig ist die Tatsache, dass die Auswahl der konkreten Untersuchungsmenge, sollte diese nicht den gesamten SCI umfassen, natürlich fundiert geschehen sollte, um entweder eine statistische Repräsentativität oder die weitreichende Abdeckung einer natürlichen Teilentität wie eines Fachgebietes[29] zu erreichen. Diese Aspekte verändern m. E. nicht die Skalierung (ein Nullpunkt, der auf ein Fachgebiet bezogen ist, ist in diesem Rahmen genau so sinntragend wie einer, der auf den gesamten SCI bezogen ist), sondern die Reichweite und gegebenen Falls die Qualität der Ergebnisse.

5.2.5. Das Adding Zeros-Problem

Ahlgren et al. (2003) fokussieren ihre Kritik von *Pearson`s r* auf die Tatsache, dass das Maß sensitiv für das sogenannte adding zeros-Problem ist. Dieses impliziert, dass einer Matrix mit Autoren, deren Korrelationen schon berechnet wurden, eine weitere hinzugefügt wird. Ahlgren et al. stellen die beiden Axiome auf, dass zum einen der Korrelationskoeffizient zweier Autoren der ersten Teilmatrix, die beide niemals mit der Gruppe der neu hinzufügten Autoren kozitiert werden, nach der Expansion nicht sinken darf und zum anderen, dass, wenn ein Korrelationskoeffizient zwischen zwei Autoren vor der Expansion kleiner als der zwischen zwei anderen ist und alle

28 keine Aufnahme von Monographien, Unvollständigkeit im Proceedingsbereich und insb. bei nicht-englischsprachiger Literatur

29 Schwierigkeit der bibliometrischen Eingrenzung eines Fachgebietes durch Überlappung der Literatur (vgl. Bradford`s Law of Scattering und Garfield`s Law of Concentration)

vier Autoren nicht mit der Gruppe neu hinzugefügter Autoren kozitiert werden, diese Relation in etwa so bleiben muss. Diese beiden Forderungen werden als notwendig für ein Kozitations-Ähnlichkeitsmaß betrachtet.

In der kleinen Testmenge zeigt sich in Bezug auf die erste Forderung, dass ein negativer oder niedriger Korrelationskoeffizient zwar ansteigt, aber ein hoher Korrelationswert durch adding zeros absinkt und ebenso die Differenz zwischen zwei *r*-Werten im Vergleich zur ersten Berechnung kleiner wird; beiden Forderungen wird also nicht entsprochen, wenn dessen ungeachtet *Pearson`s r* bei der sehr kleinen und stark dichotomen Testmenge aus viel zitierten Autoren der *Bibliometrie* einerseits und *Retrieval Research* andererseits kohärente Ergebnisse zeitigte.

Sie schlagen als alternative Maße *Chi-Quadrat* und *Salton`s Cosine* vor, die den beiden aufgestellten Forderungen genügen.

Ahlgren et al. weisen daraufhin, dass ihre in einem ACA-Ansatz nachgewiesene Kritikpunkte auch für andere Formen sozialer Interaktionsforschung (Ko-Wort, Dokument-Kozitation etc.) gelten, da es keinen mathematischen Unterschied gebe.

White (2003) verteidigt daraufhin die Verwendung von *Pearson's r* in den Arbeiten der Drexel University mit den pragmatischen Hinweisen, dass die ACA der Drexel University auf größeren Datenmengen basierten, Korrelationskoeffizienten fixiert und nicht neu berechnet würden und dass diese Daten in der Regel stark miteinander korrelierten, während Ahlgren et al. auf relativ artifizielle Weise dichotome Blöcke miteinander in Beziehung setzten, und vor allem mit dem Hinweis, dass die Unterschiede empirisch vernachlässigbar seien. Er nutzt die rohe Kozitations- und Pearson`s Korrelationsmatrix, die Ahlgren et al. veröffentlichten, als Input für MDS[30] und SPSS Cluster (agglomerativ-hierarchisch mit *complete linkage*) – mit unterschiedlichen Variationen der Diagonalenbehandlung – und zeigt mit den Dendrogrammenen und MDS-maps, dass beide Maße zu substanziell ähnlichen Clustern führen. Auch eine Clusterdistribution, die auf *Chi-Quadrat* als Distanzmaß basiert, ist ähnlich denen der beiden Ähnlichkeitsmaße. Zum

30 SPSS ALSCAL für ordinale Skalierung

Befund fast identischer Resultate kamen auch schon Leydesdorff und Zaal (1988), als sie *Pearson`s r*, *Saltons Kosinuskoeffizienten*, den *Jaccard-Koeffizienten* und die *Euklidische Distanz* in einer Ko-Wort-Analyse verglichen.

White konzediert, dass die Logarithmisierung der Daten eigentlich notwendig sei, weist aber auch in diesem Fall daraufhin, dass es in einem Test bis auf kleine Änderungen in der Feinstruktur zu keinen Unterschieden in der Interpretation der Cluster gekommen sei.

Er unternimmt es jedoch genauso wenig wie Bensman (2004), der ebenfalls Ahlgrens Vorgehensweise kritisiert, die Gültigkeit der beiden Axiome logisch zu widerlegen, obwohl beide spekulieren, möglicherweise handele es sich beim adding zeros-Phänomen sogar um einen Vorteil – Bensman argumentiert nur auf einer allgemeinen Ebene, die Invarianz der *r*-Werte könne Forscher dazu bringen, die Natur der untersuchten Beziehungen der ersten Testmenge genauer in den Blickwinkel zu nehmen. M. E. ist daher die logische Korrektheit der Kritik Ahlgrens nicht widerlegt.

6. Entropie

Die *Shannon-Wiener*-Entropieformel ist ein verbreitetes Diversitätsmaß: Sie kann benutzt werden, um den Grad der Unterteilung – die Diversität – einer Distribution zu messen. Ausgehend von einer Anzahl von Ereignissen mit feststehenden Wahrscheinlichkeiten lautet Shannons Grundüberlegung, dass der Informationsgehalt einer Nachricht, die das Eintreffen eines Ereignisses beinhaltet, desto kleiner sei, je größer der Wahrscheinlichkeitswert desselben ist. Wenn die Wahrscheinlichkeit eines Ereignisses jedoch sehr klein ist und dieses dann dennoch eintrifft, ist der Informationsgehalt umgekehrt sehr groß.

Shannon schlug daher vor, den Informationsgehalt bzw. die Entropie einer Nachricht mittels einer (von ∞ zu 0) absteigenden Funktion nach ihrer Wahrscheinlichkeit zu messen:

$$h(p) = \log \frac{1}{p} = -\log p$$

Diese Funktion hat den Vorteil der Additivität (Theil, 1972), die Entropie einer gesamten Verteilung lässt sich berechnen, indem die Einzelentropien mit ihren jeweiligen Wahrscheinlichkeitswerten gewichtet und dann addiert werden:

$$H(I) = \sum_i p_i h(p_i) = \sum_i p_i \log \frac{1}{p_i} = -\sum_i \frac{n_i}{n} \log \frac{n_i}{n}$$

Von der ursprünglichen Funktion der Formel als Erwartungswert einer Nachricht, die das Eintreffen eines Ereignisses auf der Basis einer gegebenen Wahrscheinlichkeits- bzw. Häufigkeitsstruktur berechnet, kann die Entropie übertragen werden auf die Diversität einer Cluster-Distribution: Danach bestimmt sich die Wahrscheinlichkeit eines Items aus einem bestimmten Cluster über die Größe des Clusters in Proportion zur Größe der Gesamtmenge

$$p_i = \frac{g_i}{n} .$$

Die Entropie ist immer positiv, da die Einzelentropien und die Wahrscheinlichkeiten, mit denen diese gewichtet werden, positiv sind. Sie nimmt den Wert 0 an, wenn eines der Ereignisse die Wahrscheinlichkeit 1 hat. Die Entropie ist maximal, wenn alle Ereignisse eine gleich große Wahrscheinlichkeit haben, wenn also n Ereignisse den Wahrscheinlichkeitswert *1/n* haben und daher eine maximale Unsicherheit besteht.

Theil (1972) betont die funktionelle Ähnlichkeit der Entropie mit der Varianz einer Zufallsvariablen (die die quadrierte Distanz zum Mittel benutzt), jedoch ist ein wichtiger Unterschied, dass die Varianz im Unterschied zur Entropie nicht auf Nominalwerte angewandt werden kann. Mit der Entropie dagegen kann die Diversität jeder beliebigen Struktur gemessen werden, da sie rein probabilistisch definiert ist.

Da der Maximalwert von der Anzahl der Cluster abhängt, ist es, wenn unterschiedliche Clusterdistributionen verglichen werden sollen, sinnvoll, die jeweiligen Entropiewerte anhand der Maximalwerte der Entropien zu normalisieren:

$$\frac{H(I)}{H_{\max}(I)}$$

Für den Maximalwert werden alle Items als ungeclustert betrachtet und gehen mit der Wahrscheinlichkeit *1/n* in die Berechnung ein.

In der folgenden Umformung der Entropieformel

$$H = -\sum_i \frac{n_i}{n} \ln \frac{n_i}{n} = \ln n - \frac{1}{n} \sum_i n_i \ln n_i$$

kann man leicht sehen, dass die maximale Entropie einfach durch ln*n* berechnet werden kann, da der gesamte Term auf der rechten Seite wegfällt, weil lnn_i null ergibt.

7. Konzeption und Durchführung der **feasibility study**

Das Untersuchungssetting umfasst den SCI-Jahrgang 1998 als Datenbasis, aus dem anhand der Zugehörigkeit zu Zeitschriften der Elektrochemie ein Artikelmassiv selektiert wurde. Zur Bildung von Kozitationsclustern wird eine *single linkage*-Methode mit variierten Schwellwerten auf der Basis des Kosinus-Koeffizienten angewandt, daraufhin werden durch Projektion der Kozitations-Clusterdistribution auf den Ausgangsjahrgang *research fronts* konstituiert und die Diversität der resultierenden *research fronts*-Distribution wird anschließend mit der Entropieformel gemessen. Analog wird die Entropieformel benutzt, um, bezogen auf die Anteile fünf ausgewählter Länder an der Artikelmenge und den *research fronts*, die Diversität der Forschungslandschaften dieser Länder zu berechnen.

14 Eletrochemie-Journale wurden aus verschiedenen Quellen ermittelt: der ISI journal list, einer Suche im Web of Science mit den Strings „electroch*“ und „elektroch*“ sowie einer Clusteranalyse von SCI-Journalen von Leydesdorff (2004). Mit Hilfe dieser Liste wurden im SCI-Jahrgang 1998 alle Artikel selektiert, die in einer der 14 Zeitschriften publiziert wurden. Daraus resultierte ein Datenset von 4 522 Records. Es wurde entschieden, nur research papers in der Untersuchung zu verwenden: als solche werden üblicherweise die Artikel der Kategorien articles, notes und letters definiert, da in diesen drei Publikationsformen im Unterschied z.B. zu reviews Forschungsergebnisse kommuniziert werden. 4 257 articles und 62 letters wurden selektiert; notes waren im Datenset nicht vorhanden. Die Untersuchungsmenge besteht also aus 4 319 Records, von denen 110 jedoch keine Referenzen haben. Die Zahl der unterschiedlichen Referenzen dieser Artikelmenge beträgt 69 764[31].

Single linkage clustering ist ein sehr simpler Clusteralgorithmus, der trotz struktureller Schwächen aufgrund seiner universalen Anwendbarkeit – er ist mit jedem beliebigen Proxitimitätsmaß durchführbar – und vor allem seiner Schnelligkeit und Eignung für große Datenmengen in der bibliometrischen

31 gezählt anhand des SCI, kein weiteres Synomymen-clearing

Forschung relativ verbreitet ist. Wir wollen seine Tauglichkeit für Zitationsstudien in dieser Studie evaluieren.

Saltons Kosinus-Koeffizient ist ein in der Zitationsanalyse bewährtes Maß zur Normalisierung unterschiedlich langer Referenzenlisten. Small hält ihn aufgrund des stärkeren Ausgleichs hoch und niedrig zitierter Referenzen für überlegen gegenüber dem *Jaccard-Koeffizienten* (Small & Sweeney, 1985a). Gegenüber dem *Korrelations-Koeffizienten* (*Pearson`s r*) erscheint er uns als die bessere Wahl angesichts der theoretischen und praktischen Probleme, die *Pearson`s r* den Anwender in Bezug auf bibliometrische Untersuchungen stellt. Wir folgen insbesondere der Position von Dillon & Goldstein (1984), nach der die Ähnlichkeit, die gemessen werden soll, nicht essentiell linear ist und dass die Anforderungen, die *Pearson`s r* an die Eigenschaften der Daten stellt (Additivität, konstante Varianz und Normalverteilung) bei bibliometrischen Daten nicht zutreffen und eine logarithmische Transformation nachteilige Folgen hat.

Anfängliche Experimente werden in Anlehnung an die Studienreihe Smalls mit einer absoluten cut off-Schwelle und einer fraktionalen Zitationsschwelle zur Nivellierung langer Referenzenlisten in Kombination mit einer Kosinus-Zitationsschwelle durchgeführt. Durch die Kombination der Schwellen wird die Datenbasis jedoch zu massiv verringert.

Der Verzicht auf die fraktionale Schwelle wird im Hinblick auf ihre Hauptfunktion, massive Differenzen zwischen Publikations- und Zitationsmodi auszugleichen, von uns als vertretbar angesehen, da die Untersuchung in diesem Stadium nicht disziplinenübergreifend ist. Die einfache cut off-Schwelle, die ihrerseits die Bevorzugung kurzer Referenzenlisten, wie sie der fraktionalen Schwelle inhärent ist, nivellieren soll, wird im Gegenzug auf größer eins gesetzt. Dadurch begrenzen wir die Referenzenmenge auf diejenigen, anhand derer Artikel bibliographisch gekoppelt werden können.

Der Kozitationsschwellwert wird variiert, um einerseits auf dieser Ebene eine möglichst große Coverage zu erreichen bei einer andererseits möglichst feinen Clusterung. Wir berechnen die Entropie der Clusterdistribution für Salton-normalisierten Kozitationsschwellwerte $s > 0.1$ bis $s > 0.9$, bezogen

auf die Gesamtmenge der Referenzen, die mehr als einmal zitiert wurden. Diese Werte können als Indikatoren für die Entropiewerte auf der Ebene der *research fronts* gelten.

Die ausgewählte Clusterdistribution wird auf die Ausgangsebene, den Artikeljahrgang 1998, rückprojiziert. Diese Projektion entspricht faktisch dem Vorgang einer Bibliographischen Kopplung, die sich auf die Kozitationscluster statt einzelner Referenzen, *concept symbols*, bezieht: Alle Artikel, die mindestens eine Referenz aus dem gleichen Kozitationscluster zitieren, bilden zusammen eine *research front*. Sie sind mit allen anderen Artikeln der *research front* bibliographisch gekoppelt in Beziehung auf das jeweilige Kozitationscluster. Zudem sind sie mit mindestens einem anderen Artikel in Beziehung auf eine gemeinsame Referenz gekoppelt.

Zusätzlich bilden wir mittels einfacher Bibliographischer Kopplung *research fronts* aus allen Referenzen, die 1998 mehr als einmal zitiert, aber nicht zu dem jeweils spezifizierten Schwellwert kozitiert wurden. Durch die Kombination der beiden unterschiedlichen methodischen Ansätze sollen Forschungslinien, die, ablesbar an der Existenz von Kozitationsstrukturen, in die Vergangenheit zurückreichen, durch jene ergänzt werden, die sich aktuell konstituieren, ablesbar nur an gemeinsamen Referenzen.

Für eine Kombination der beiden methodischen Ansätze spricht auch die massive Erhöhung der Coverage, die mit ihr erreicht werden kann. Wie aus der Untersuchung Jarnevings (2005) deutlich wird, unterscheiden sich *research fronts*-Distributionen, die methodisch auf Bibliographischer Kopplung basieren, und solche, die auf Kozitationscluster zurückgehen, unter analogen Bedingungen der Schwellwertgestaltung nicht wesentlich in Bezug auf die *coverage*. Werden jedoch beide kombiniert, kann die *coverage* signifikant erhöht werden. Die theoretische Begründung für die Methodenfusion kann somit quantitativ gerechtfertigt werden.

Research fronts können sich in unserem Modell überschneiden, da Artikel Referenzen aus verschiedenen Clustern zitieren können. Während wir diese Überschneidungen anfänglich auf die Gesamtmenge der Artikel addierten, da die betreffenden Artikel formal mehrmals gezählt werden und dadurch das Verhältnis der Artikel in *research fronts* zu ungebundenen verfälschen,

revidierten wir dieses Vorgehen als unbefriedigend und gingen dazu über, überschneidende Artikel den jeweiligen *research fronts* fraktional zuzurechnen. D. h., während Artikel, die Referenzen aus einem Kozitationscluster zitieren, mit dem Wert 1 der entsprechenden *research front* zugerechnet werden, wird ein Artikel, der Referenzen aus zwei Kozitationsclustern zitiert, mit jeweils dem Wert 1/2 den beiden *research fronts* zugerechnet.

Bei der Messung der Diversität der Verteilung werden von uns auch diejenigen Artikel, die nicht an *research fronts* partizipieren, also keine Kozitationscluster zitieren und nicht bibliographisch gekoppelt sind, mit eingeschlossen. Diese einzelnen Artikel gehen als singuläre Entitäten mit der Größe 1 in die zu messende Verteilung ein und werden als fachlich abgegrenzte Einzelforschung interpretiert. Ein Weglassen dieser singulären Artikel, wie es oftmals bei der Auswertung von Clusterdistributionen geschieht, wäre durch die *Shannon-Wiener-Entropie* nicht zu rechtfertigen und würde die Reichweite der Messung auf unsinnige Weise beschränken. Wird ihre Anzahl aber zu groß, sollte der Clusteralgorithmus prinzipiell in Frage gestellt bzw. durch einen besser geeigneten Algorithmus ersetzt werden, da eine große Menge von singulären Entitäten gegen die Plausibilität der Repräsentation spricht.

Wir wählten sechs Länder (USA, Frankreich, Großbritannien, Deutschland, Japan und Russland), deren Anteile an der Distribution der *research fronts* und deren Entropie berechnet wurden. Hierfür werden Artikel jeweils den Ländern zugeordnet, denen mindestens einer der Autoren angehört. Internationale Kollaborationen zwischen zwei oder mehreren der sechs Länder führen dazu, dass Artikel mehreren Ländern zugeschlagen werden können, was in diesem Fall kein methodisches Problem bedeutet, da die Entropien für jedes Land getrennt berechnet werden.

Die Verarbeitung der Daten wurde von mir mit perl scripts und dem freien Netzwerk-Tool Pajek vorgenommen. Aufbauend auf einem script[32], das die Zitationsdaten der CD-Rom nach Autoren geordnet in einen Textstring

32 von Michael Heinz

ausliest, schrieb ich zunächst ein perl-Programm (ANHANG D.1. DATENIMPORT), das die Textdatei in die relationale Struktur einer MySQL-Datenbank[33] einliest. Im nächsten script (ANHANG D.2. SCHWELLWERTBERECHNUNG(1)) werden für die Referenzen die absolute Zahl der Zitationen bestimmt und die fraktionalen Zitationswerte berechnet sowie diese für jede Referenz addiert. Im dritten perl script (ANHANG D.3. SCHWELLWERTBERECHNUNG(2)) werden die Kozitationspaare konstituiert und die Kosinus-normalisierte Kozitationsschwelle berechnet. Für die modifizierte Version des Untersuchungssettings werden die Kozitationspaare, bei denen beide Partner eine Zitationsanzahl größer eins erreicht haben, abgelegt und daraufhin werden durch if-then-Schleifen Kozitationspaare gemäß des inkrementierten Kozitationsschwellwertes selektiert. Diese werden in das Importformat für Pajek umgeschrieben und als Textdateien ausgelesen (ANHANG D.4. CLUSTERVORBEREITUNG). Die *single linkage*-Clusterprozedur wird in das Netzwerktool Pajek ausgelagert, dessen implementierte *Force-directed Placement*-Prozeduren auch zur Visualisierung der Clusterstrukturen genutzt werden können.

Pajek kann verschiedene Netzwerkdatenformate lesen, darunter Matrixformate und Ucinet-DL-files. Aufgrund der Ähnlichkeit zur relationalen Struktur verwendete ich eines der Pajek-nativen network-Formate[34] zum Import der Daten.

Pajek kreiert aus den Daten ein ungerichtetes, ungewichtetes Netzwerk. Durch den Befehl Net>Components werden Cluster konstruiert. Pajek definiert ein Netzwerk oder Subnetzwerk als connected bzw. weakly connected, wenn alle Vertizes durch einen semipath mit allen anderen verbunden sind. Ein semipath ist ein zusammenhängender Weg zwischen Start- und Endvertex, bei dem Start- und Endvertex identisch sein können, aber zwischen ihnen kein Vertex mehrmals vorkommen darf – die Restriktion hat die Funktion, die Zahl der redundanten möglichen Pfade einzuschränken. Die Unterscheidung zwischen weakly und strongly connected ist relevant nur

33 Programmierung der Datenbank durch Alexander Struck

34 die edge list, in der zunächst Vertizes definiert und dann Kanten zwischen ihnen aufgelistet werden. Kanten sind im Unterschied zu arcs per Definition ungerichtete Verbindungen

in Bezug auf gerichtete Graphen, für ungerichtete Graphen gibt es nur eine Form der connectedness, die dem weakly connectedness der gerichteten Graphen und dem single linkage der Clustertheorie entspricht[35]. In einer Visualisierung werden die einzelnen components verschiedenfarbig dargestellt und der Graph im *Fruchtermann-Reingold-Algorithmus* gemäß den inhärenten Dichtestrukturen dargestellt (vgl. 4.6), wodurch die Binnenstruktur der Cluster gut visuell inspiziert werden kann.

Mit dem nächsten perl-script werden die Pajek-Output-Dateien der Cluster in die Datenbank eingelesen und die Entropie der Clusterdistribution, bezogen auf die Referenzenmenge mit dem Zitationswert größer eins, berechnet (ANHANG D.5. ENTROPIEN DER REFERENZEN).

Wie benutzen an dieser Stelle die schon erwähnte Umformung der Entropieformel

$$H = -\sum_i \frac{n_i}{n} \ln \frac{n_i}{n} = \ln n - \frac{1}{n} \sum_i n_i \ln n_i$$

,

die, weil alle Entitäten mit der Größe eins in der Summe wegfallen, rechengünstiger ist.

Die *research fronts* werden konstituiert, Überschneidungen fraktional umgerechnet und gleichzeitig durch Rückgriff auf die Zitationsdaten die Referenzen, die mehr als einmal zitiert, aber nicht kozitiert wurden, bibliographisch gekoppelt (ANHANG D.6. BERECHNUNG DER FORSCHUNGSFRONTEN). Daraufhin berechne ich die Entropie der *research fronts*; die Anteile der Länder an ihnen und die Entropie der Länder (ANHANG D.7. ENTROPIEN DER ARTIKEL/LÄNDER).

Die Trennung der Scripts nach Aufgaben begründete sich im Arbeitsprozess dadurch, dass insbesondere die beiden scripts zur Konstruktion der Kozitationspaare und der Berechnung der Schwellwertgrundlagen sehr rechenintensiv sind und bei den diversen Variationen der Schwellwerte nicht wiederholt werden müssen. Das Netzwerktool Pajek ist nicht extern

35 bei strong communities wird für die Definition des paths die Richtung der Kanten berücksichtigt - beide Konzepte sind strikt nur auf die Binnenstruktur der Cluster gerichtet und nicht auf das Verhältnis zu den nach außen führenden Kanten, wie beim gleichnamigen Konzept, das im Algorithmus von Radicchi et al. (2004) benutzt wird

ansteuerbar und kann somit nicht in perl-Routinen integriert werden, so dass auch an dieser Stelle eine Trennung der scripts zwangsläufig ist.
Ergebnisdaten werden nicht in perl-eigenen Datenstrukturen, sondern generell in MySQL-Tabellen gespeichert, um sie unabhängig von Programmlaufzeiten für Zugriffe in späteren Prozessphasen und zur Überprüfung vorzuhalten.

8. Ergebnisse

Bei einer absoluten Zitationsschwelle $c \geq 5$, einer fraktionalen Zitationsschwelle $f \geq 1.5$ und einer Kozitationsschwelle $s \geq 0.17$[36] ergeben sich auf der Kozitationsebene nur 2 Cluster mit je 6 und 2 Mitgliedern. Bei einer Verringerung der fraktionalen Schwelle auf $f \geq 1$ fusionieren die beiden Cluster, bei einer fraktionalen Schwelle $f \geq 1.1$ liegt die Anzahl noch bei zwei Clustern, die gegenüber dem Anfangszustand etwas angewachsen sind (18 Mitglieder).

Bei einer Verringerung der fraktionalen Schwelle auf $f \geq 0.5$ und der absoluten Zitationsschwelle auf $c \geq 3$ bei gleichzeitiger Steigerung der Kozitationsschwelle auf $s \geq 0.26$ entstehen 12 Cluster mit insgesamt 116 Mitgliedern. Das größte Cluster enthält 31 Referenzen.

Alle Resultate, die aus der Kombination der drei Schwellwerte entstehen, disqualifizieren sich durch die extrem geringe *coverage*.

Bei einer konstanten absoluten Zitationsschwelle $c > 1$ und einer Erhöhung des Kosinus-normalisierten Kozitationsschwellwertes von $s > 0.1$ auf $s > 0.9$ nimmt die Anzahl der in Clustern gebundenen Referenzen von 12 723 beim Kozitationsschwellwert $s > 0.1$ auf 3 869 bei $s > 0.9$ ab.

Die Anzahl der Referenzen-Cluster steigt von 18 beim Schwellwert $s > 0.1$ auf 1435 beim Schwellwert 0.8 und sinkt dann beim Schwellwert $s > 0.9$ auf 1226 ab.

Die Entropie der Distribution der Referenzen-Cluster steigt von $H = 0.078$ bei $s > 0.1$ auf $H = 9.049$ bei $s > 0.9$ an.

Das Verhältnis $H(I)/H_{max}(I)$ liegt beim Schwellwert $s > 0.1$ bei 0.08 und ist 0.9573 beim Kozitationsschwellwert $s > 0.9$, also nahe dem Maximalwert (Tabelle 1).

Die sehr niedrige Entropie am Anfang ist durch die niedrige Clusteranzahl bei hoher *coverage* der Clusterdistribution bedingt: Ein überwältigender Anteil der Referenzen ist in sehr wenigen Clustern und vor allem in einem

36 in Anlehnung an die Schwellwertempfehlungen Smalls mit dem Unterschied, dass die fraktionale Berechnung in Bezug auf die ungekürzten Referenzenlisten berechnet wurden

Makrocluster organisiert. Am anderen Pol ist die Entropie annähernd maximal, weil die Anzahl der Referenzen in Clustern stark geschrumpft ist und alle nicht der Clusterdistribution angehörenden Referenzen als singuläre Entitäten in die Berechnung eingehen.

Die Erhöhung des Kozitationsschwellwertes schlägt sich in einer stark wachsenden Anzahl von Mikroclustern nieder, die von größeren *communities* abgespalten werden. Eine zu massive Anzahl kleiner Cluster spricht gegen die Adäquatheit der Distribution, wie andererseits auch die Existenz von Makroclustern, von denen man annehmen kann, dass sie mit großer Wahrscheinlichkeit inhaltlich heterogene Substrukturen zusammenketten. Makrocluster sind ein in der Literatur diskutiertes Phänomen basierend auf den *chaining*-Effekten des *single linkage*-Clusterns (vgl. 4.1).

Zwischen den Schwellwerten $s > 0.4$ und $s > 0.5$ ereignet sich gleichzeitig ein massiver Anstieg der Clusteranzahl von 31 auf 990 und ein massiver Rückgang der Coverage von 12 544 auf 8 960 Referenzen (Tabelle 2).

Der immense Sprung in der Clusteranzahl zwischen den Schwellwerten $s > 0.4$ und $s > 0.5$ findet am stärksten zwischen den Schwellwerten $s > 0.49$ und $s > 0.5$ statt. Dies ist durch die sehr große Anzahl der Kozitationspaare, die einmal kozitiert werden, während beide Partner dieser Paare jeweils zweimal zitiert werden (dies ergibt in der Kosinus-Normalisierung 0.5), begründet. Ein kleinerer Sprung zwischen $s > 0.4$ und $s > 0.41$ findet exakt bei 0.409 statt; dabei handelt es sich um die Paare, die ebenfalls einmal kozitiert werden, während die beiden Partner je zweimal und dreimal zitiert werden.

Beim Kozitationsschwellwert $s > 0.4$ enthält das größte Cluster 12 421 Referenzen; der Schwellwert $s > 0.49$ ergibt 127 Cluster und das größte Cluster enthält immer noch 11 413 Referenzen. Beim Schwellwert $s > 0.5$ wird das größte Cluster signifikant kleiner, es enthält nun nur noch 2630 Referenzen. Die massive Verkleinerung des Makroclusters auf diesem Level ist jedoch nur bei dem gleichzeitigen immensen Anstieg der Clusteranzahl auf 990, darunter allein 460 Cluster mit der Minimalgröße 2, möglich. Diese Clusteranzahl erscheint vollkommen inadäquat; die immense Vielzahl der Mikrocluster mit zwei oder drei Mitgliedern kann keine sinnvolle Repräsentanz von inhaltlich begründeten Substrukturen sein.

Tabelle 1 : Clusterdistribution der Referenzen s > 0.1 bis s > 0.9

Kozitations-schwelle	Cluster-anzahl	Referenzen in Clustern	größtes Cluster	H(c)	$H(c)/H_{max}(c)$
> 0.1	18	12723	12 644	0.077	0.008
> 0.2	18	12723	12 644	0.077	0.008
> 0.3	20	12705	12 618	0.097	0.010
> 0.4	31	12544	12 421	0.255	0.027
≥ 0.5	126	11867	11413	1.037	0.11
> 0.5	990	8960	2630	6.505	0.688
> 0.6	1262	8209	144	7.8	0.825
> 0.7	1379	7385	73	8.185	0.866
> 0.8	1435	6198	48	8.580	0.908
> 0.9	1226	3869	25	9.050	0.957

Tabelle 2: Clusterdistribution der Referenzen s > 0.41 bis s > 0.49

Kozitations-schwelle	Cluster-anzahl	Referenzen in Clustern	größtes Cluster	H(c)	$H(c)/H_{max}(c)$
> 0.41	94	12075	11704	0.807	0.085
> 0.42	95	12073	11699	0.811	0.085
> 0.43	98	12603	11683	0.824	0.087
> 0.44	109	12027	11620	0.873	0.092
> 0.45	115	11976	11555	0.924	0.098
> 0.46	115	11969	11549	0.930	0.098
> 0.47	114	11955	11538	0.939	0.099
> 0.48	125	11870	11418	1.033	0.109
> 0.49	127	11869	11413	1.037	0.11

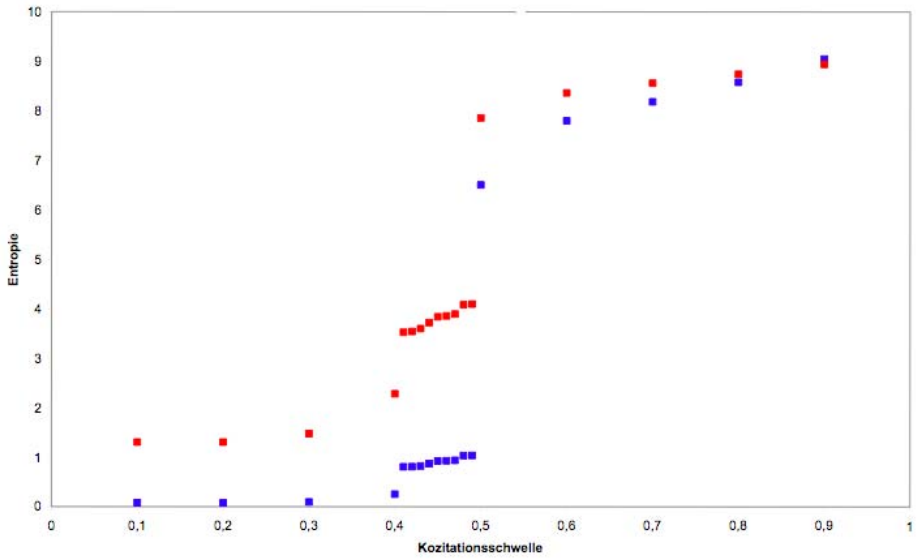

Abbildung 1: Entropien der Kozitationscluster (blau) und research fronts (rot)

Diagramm 1 zeigt die Entwicklung der Entropiewerte der Distributionen der Kozitationscluster (blaue Punkte) und der research fronts (rote Punkte) vom Kozitationsschwellwert s > 0.1 bis s > 0.9. Die Verteilungen auf der Ebene der research fronts sind im Vergleich zu denen der Kozitationsebene diverser: Neben der durch die bibliographische Kopplung (um das fast 8-fache) vergrößerten Menge der Partitionen ist das Makrocluster relativ kleiner geworden. So umfasst die größte research front bei s > 0.4 81% der geclusterten Artikel (beim Schwellwert s ≥ 0.5 nur 58%), während auf der Kozitationsebene das größte Referenzen-Cluster 99% der geclusterten Referenzen beim Kozitationsschwellwert s > 0.4 enthält.

Diese Unterschiede werden durch die *Shannon-Wiener-Entropie* repräsentiert, in der die Anzahl der Partitionen und deren Größen synthetisiert werden.

Aufgrund des reziproken Sprungs in Clusteranzahl und *coverage*, der bei der Kozitationsschwelle *s > 0.5* stattfindet, bietet sich ein Versuch mit *s > 0.49* oder *s ≥ 0.5* (beide Schwellen verhalten sich auf der Kozitationsebene sehr

ähnlich) an. Die Projektion der Referenzen-Clusterdistribution auf die Artikel-Ebene der *research fronts* ergibt 1010 *research fronts* mit einer Entropie H_r = *4.096* and $H_r/H_{r\,max}$ = *0.49*.

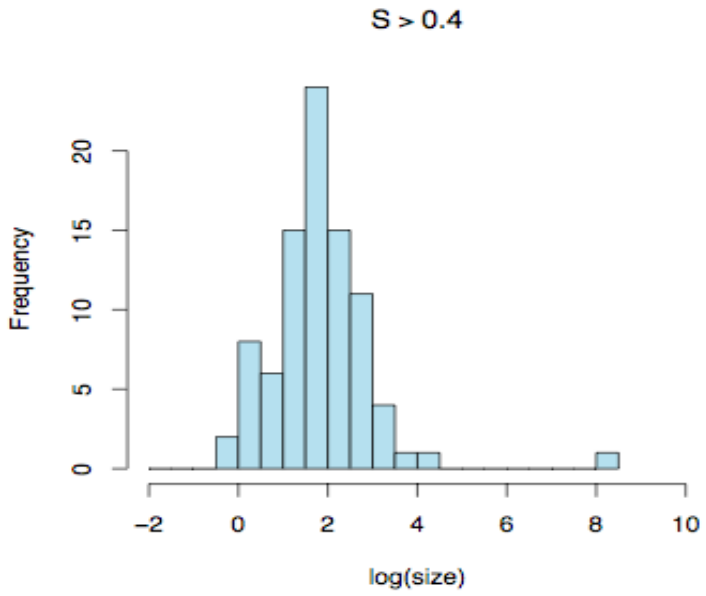

Abbildung 2: Größen-Häufigkeitsverteilung der research fronts für s > 0.4

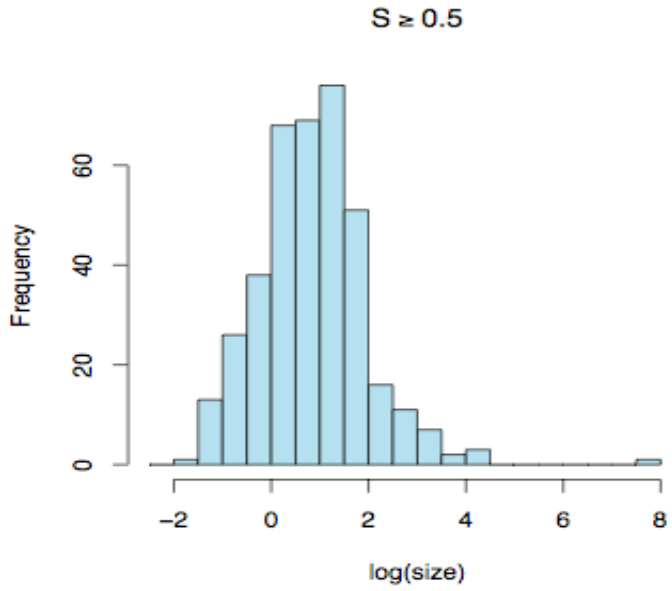

Abbildung 3: Größen-Häufigkeitsverteilung der research fronts für s ≥ 0.5[37]

[37] beide Diagramme sind Schmidt (2006) entnommen

Die Diagramme 2 und 3 zeigen die Größen-Häufigkeitsverteilungen der *research fronts* für *s > 0.4* und *s ≥ 0.5*. Die große Anzahl der Partitionen bei *s ≥ 0.5* zieht, wie deutlich zu erkennen ist, sehr viele kleine Entitäten nach sich. Dies spricht gegen die Wahl dieses Schwellwertes, zur Repräsentation eines einzigen Fachgebietes ist diese Partitionierung zu fraktioniert. Beide Verteilungen sind als Ganze nicht *power law*-verteilt (vgl. 9.1); der Abstand zum Makrocluster und die Anzahl der Mikrocluster sind jeweils zu groß. Abgesehen vom Makrocluster sind die Verteilungen relativ symmetrisch. Wir benutzen den Kozitationsschwellwert *s > 0.4* für die weitere Kalkulation der Länderanteile an den *research fronts* und der Länder-Entropien. Auf diesem Level sind 3865 von 4319 Artikel (89%) in 239 *research fronts* gebunden (genau so viele wie bei ≥ 0.5[38]), nur 454 Artikel bleiben singulär. Die Entropie für die Distribution der *research fronts* ist 2.284 und $H(r)/H_{max}(r)$ = 0.3167

Tabelle 3: Entropien der Länderbeteiligungen an den research fronts-Distributionen (Kozitationsschwellwert > 0.4)

	Anzahl research fronts	Alle Artikel mit Beteiligung des Landes	Singuläre Artikel	Anzahl der Artikel in research fronts	H(r)
USA	114	763	93	670	2.05
Deutschland	80	393	48	345	1.89
Russland	48	320	49	271	1.88
Japan	105	658	61	597	1.86
Frankreich	81	356	23	333	1.72
GB	56	245	19	226	1.55

38 die Anzahl der in research fronts gebundenen Artikel, 3865, ist für alle Distributionen von s > 0.1 bis s > 0.9 identisch. Das ist notwendigerweise so, weil Artikel, die ein Kozitationscluster zitieren, immer mit mindestens einem anderen Artikel bibliographisch gekoppelt sind. Durch Schwellwerterhöhungen werden Cluster aufgespalten und Referenzen, die mit nicht ausreichender Kozitationsstärke mit anderen zitiert werden, zu singulären Items, die davon unbeeinflusst weiterhin von mindestens zwei Artikeln zitiert werden. Die Anzahl bibliographisch gekoppelter Artikel ist unabhängig von Kozitationsschwellwerten, deshalb ist die Anzahl der Artikel in research fronts konstant

Die *single linkage*-Clusterdistribution stellt ein problematisches Fundament der Untersuchung dar. Die Struktur fasst offenbar einen großen Bereich der mainstream-Forschung mit unbekannter Subgliederung in einer Partition zusammen und erlaubt nur an den Rändern Diversifikation – inwieweit bzw. in welcher Zusammensetzung diese aus nach dem *Bradford-Gesetz* in Elektrochemie-Journalen publizierten Artikeln benachbarter Fachgebiete oder aus sehr abgegrenzten Ansätzen innerhalb der Elektrochemie resultieren, lässt sich im Rahmen dieser Untersuchung nicht aufklären.

Vor diesem Hintergrund ist eine wirklich substanzielle Interpretation der Länderentropien nicht möglich, da eine solche auf einer validen Clusterstruktur aufbauen müsste.

Es ist leicht festzustellen, dass die Entropie nicht direkt mit der Größe des Outputs und auch nicht mit der Summe der *research front*-Beteiligungen und der singulären Artikel eines Landes korreliert (Tabelle 3): Während die USA den größten Output an Artikeln insgesamt, die höchste Beteiligung an *research fronts*, die höchste Anzahl an singulären Artikeln und die größte Diversität haben, folgt mit der zweitgrößten Diversität Deutschland, das an nur etwa halb so vielen Artikeln wie die USA beteiligt und auch weniger produktiv als Japan ist. Japan, Deutschland und Russland haben fast die gleiche Entropie, obwohl Japan an etwa doppelt so vielen Artikeln und auch sehr viel mehr *research fronts* zuzüglich singulären Artikeln wie Russland beteiligt ist.

Es ist folgerichtig, dass die Entropie, die sowohl die Anzahl der Partitionen als auch die Gleichmäßigkeit von deren Größenverteilung misst, zu einer eigenen Rangfolge kommt, die von der Rangfolge der reinen oder normalisierten Anzahl der *research fronts* zuzüglich singulärer Artikel abweicht, die als solche durchaus auch eine Aussagekraft hat. In der Forschung zur Biodiversität werden diese beiden Komponenten eines Diversitäts-Maßes als *species richness* und *evenness* umschrieben (Rousseau & Van Hecke, 1999). Während beim Vergleich der Schwellwerte zueinander oder der Referenzen-Cluster zu den *research fronts* Vergrößerungen der *coverage* meist mit der Verringerung des Makroclusters einherging, sind die Differenzen hier diffiziler: So haben Frankreich und

Deutschland eine fast gleiche Anzahl von *research fronts* (80 und 81), eine ähnliche Anzahl von geclusterten Artikeln, aber Deutschland eine höhere Anzahl von singulären Artikeln, die trotz des leicht größeren Makroclusters für eine höhere Diversität sorgen. Allerdings sind die Unterschiede in unserer Studie nicht groß; für die Interpretation der Entropien in einer Folgestudie mit verbesserter Clusterstruktur wäre es wichtig, Entscheidungskriterien dafür zu erarbeiten, ab wann diese als signifikant zu bewerten sind.

Auf dieser Ebene ist die Messung und Einbeziehung der maximalen Entropie, wie sie bisher berechnet wurde, nicht sinnvoll, weil die Summe der *research fronts* und der singulären Artikel (693), wie sie durch die totale *research fronts*-Distribution vorgegeben ist, kleiner ist als die Gesamtzahl aller Artikelbeteiligungen der größten Landes USA, das somit keine in diesem Sinn größte maximale Entropie erreichen könnte.

9. Diskussion

9.1. Power law-*Verteilungen*

In der bibliometrischen Forschung werden oftmals *power law*-Verteilungen beobachtet, so im Fall der *Zipf-Verteilung* und *Lotka-Verteilung*. Auch skaleninvariante Netze sind nach Barabási und Albert (1999) dadurch definiert, dass, unabhängig vom Zeitpunkt und der Größe eines Netzes, die Wahrscheinlichkeit *P(k)*, dass ein Vertex mit *k* anderen verbunden ist, nach einem Potenzgesetz absteigt:

$$P(k) \approx k^{-\gamma}$$

In einer Computersimulation modellieren Barabási und Albert ein skaleninvariantes Netz und inkorporieren dabei die beiden Strukturprinzipien *growth* (die Größe des Gesamtnetzes wächst) und *preferential attachment* (die Wahrscheinlichkeit eines Vertizes, mit anderen verbunden bzw. von anderen zitiert zu werden, ist proportional zum bisherigen Vernetzungsgrad des Vertizes[39]), die sie als hauptverantwortlich für die *Skalenfreiheit* ansehen. In dieser basalen Modellierung beträgt der Exponent 3, was von Redner (zitiert nach Barabási & Albert, 1999) in einer Untersuchung eines Zitationsgraphen bestätigt wurde.

Katz (1999) untersucht Wachstumsprozesse von wissenschaftlichen communities und postuliert ein Potenzgesetz zwischen den Größen von communities (gemessen an der Anzahl der Artikel, die zwischen 1981 und 1991 im SCI indexiert wurden) und ihrer Wahrnehmung (gemessen an der Anzahl der Zitationen der Artikel dieser communities). Er generiert 152 (überlappende) communities[40] über die SCI-eigene Klassifikation, die jeweils zwischen 500 und 300000 Mitglieder haben. Von diesen communities bestimmt er die Zitationen innerhalb eines dreijährigen Zeitfensters. In seinem Modell, in dem die Anzahl der Zitationen von communities als Funktion der Größen der communities aufgetragen wird, ist, anders als im

39 das Barabási/Albert-Modell bezieht sich auf lineares preferential attachment (Barabási & Albert, 1999)

40 die referenzierten Artikel wurden auf Artikel, notes und reviews aus dem SCI reduziert

Barabási/Albert-Modell der Exponent positiv und die Gerade in der Log-log-Darstellung deshalb nicht monoton fallend, sondern ansteigend.

Van Raan (1991) untersucht eine ISI-Kozitations-Clusterdistribution des SCI/SSCI-Jahrganges von 1984 (vgl. 3.2 und 3.3) und weist eine *power law*-Verteilung der *research fronts* im Fall der C2-Stufe und der C3-Stufe nach.

In einer neueren Bibliographischen Kopplungsstudie untersucht Van Raan (2005) Clusterverteilungen, die auf Bibliographischer Kopplung basieren (ein Cluster konstituiert sich durch die Anzahl der bibliographisch gekoppelten Arbeiten über jeweils eine Referenz) und weist eine annähernde *power law*-Verteilung bis zu einem cut-off bei Clustergrößen von über 1000 Artikeln nach, wenn alle Referenzen eines Jahrgangs in die Untersuchung einfließen (die Referenzenlisten des Ausgangsjahrganges also komplett zur Bibliographischen Kopplung verwendet werden). Im Fall einer Segmentierung der Referenzen in Dreijahresschritten nähern sich die Verteilungen der Segmente älterer Referenzen eher Exponentialverteilungen an.

Katz führt seine Ergebnisse auf *preferential attachment* zurück: Objekte wachsen in ihrer Größe bzw. Vernetzung proportional zu ihrer bisherigen Größe bzw. dem bisherigen Vernetzungsgrad, weil die Attraktivität eines jeden von seiner bisherigen Größe bzw. Wahrnehmbarkeit abhänge. Die Wahrscheinlichkeit, dass sich viele Wissenschaftler mit einem Fach bzw. auf der Mikroebene mit einem bestimmten Thema oder einer Fragestellung beschäftigen, hängt in diesem Modell von der Präsenz und den damit verbundenen Erfolgsaussichten (in Bezug auf Publikationsaussichten in wichtigen Zeitschriften, Finanzierung etc.) ab.

Zu beachten wäre zudem, dass, während Artikel über die Zeit Zitationen ansammeln und nicht verlieren, auf der Ebene der communities die Annahme eines konstanten Wachstums nicht adäquat wäre. Nur wenn man davon ausginge, dass sich diese generell aufspalten, sobald sie eine bestimmte Größe erreicht haben, müsste kein negatives Wachstum oder Absterben modelliert werden.

Katz ignoriert jedoch, dass in größeren Fachgebieten – wie z.B. Biowissenschaften im Vergleich zu Mathematik – mehr relevante Literatur

vorhanden ist, die von den Wissenschaftlern wahrgenommen und zitiert werden kann. Wenn also die proportional häufigere Zitierung einiger communities mit proportional längeren Referenzenlisten der zitierenden einhergeht (wie im Fall des biomedizinischen Bias in den ISI-Studien postuliert wurde (vgl. 3.1 und 3.2), wäre die Form der resultierenden Verteilung nicht bzw. nicht vollständig dem *preferential attachment*-Effekt geschuldet. Genaue Daten zu diesem Verhältnis wären für die Interpretation der Verteilungsstruktur interessant.
Gleichzeitig nehmen wir an, dass Konzentrationsprozesse durch die aktuellen Paradigmen der Wissenschaftspolitik zunehmend initiiert bzw. verstärkt werden. Wenn es eine solche Tendenz gibt, müsste sie in Zeitreihen-Untersuchungen heraustreten.

9.2. Clusteralgorithmen

Die Evaluation von Clusteralgorithmen ist methodisch schwierig; aufgrund der Größe der Datendistributionen und des Fehlens objektiver Zuordnungskriterien wären empirische Evaluationen anhand von Interviews mit Fachexperten nicht weniger subjektivitätsbehaftet als die Clusteranalyse selbst.
Analytische Kriterien zur Bewertung einer Clusterdistribution, wie sie für die multivariaten statistischen Verfahren angewendet werden, sind nicht direkt auf graphanalytische Verfahren übertragbar, da sie ein spezifisches Abstandsmaß, häufig den euklidischen Abstand voraussetzen, wie dies z.B. bei der Methode der Doppelkreuzvalidierung der Fall ist (Bortz, 2005).
Primär für die Auswahl bzw. Bewertung der Eignung eines oder mehrerer Clusteralgorithmen sollten daher, wie in Abschnitt 2.1. erläutert, Überlegungen zur Übereinstimmung der Eigenschaften eines Algorithmus` mit den deduzierbaren oder bekannten empirischen Merkmalen des Datenmaterials sowie *scalability* und *coverage* sein.
Single linkage clustering impliziert ein sehr schwaches Clusterkriterium, das minimale Anforderungen an die Homogenität der Cluster stellt. Die Verteilungen, die wir generieren, sind – bei niedrigeren

Kozitationsschwellwerten – dominiert von einem fast die gesamte Verteilung umfassenden Makrocluster. Bei sukzessiver Erhöhung der Schwellwerte gibt es einen fast übergangslosen Wechsel zu einer Vielzahl kleiner Cluster; die Clusterverteilungen der Schwellwerte 0.4 und 0.5 sind nicht *power law*-verteilt.

Das *single linkage*-Kriterium erscheint auch theoretisch zu schwach, insofern einzelne Artikel, welche die strukturelle Basis sich unterschiedlich entwickelnder Gebiete bilden, in dieser Funktion Substrukturen durch Zitationsverbindungen zusammenketten. Eine tatsächliche Diversifikation wird also nicht immer in der Clusterdistribution repräsentiert. Durch eine Normalisierung anhand des Ähnlichkeitsmaßes wird dieses Phänomen etwas abgemildert, da die Wirkung hoch zitierter Artikeln abgeschwächt wird. Doch legen die Resultate nahe, dass die Verteilung trotzdem noch zu sehr von Verkettungseffekten geprägt ist. Die Begrenzung der Studie auf ein einzelnes Fachgebiet macht die zusätzliche Anwendung des fraktionalen Schwellwertes, der, wenn er auch insbesondere interdisziplinär wirksam ist, einen zusätzlichen Normalisierungseffekt anhand der Länge der Referenzenlisten ausüben könnte, faktisch unmöglich. Die Studie zeigt, dass eine Kombination mehrerer Schwellwerte die Datenbasis zu massiv verringert.

Wie beim *single linkage* ist auch beim *complete linkage* das Clusterkriterium statisch, d. h. unabhängig von Dichte bzw. Homogenität des Graphen. Je spärlicher ein Netzwerk ist, desto unwahrscheinlicher ist es, dass in Relation zum Gesamtgraphen größere Subgraphen komplett verlinkt sind; Zitationsgraphen sind i. d. R. keine dichten Netzwerke (Leydesdorff & Bensman, 2005).

Es erscheint inadäquat, eine inhaltliche community darüber zu definieren, dass alle Mitglieder miteinander verbunden seien; dies würde ein ideales System lückenlosen Zitierens voraussetzen. Garfield (zitiert nach Umstätter, 2005) nennt als Ursachen für *uncitedness*, dass Arbeiten als irrelevant eingeschätzt werden, vergessen oder übersehen werden oder zu bekannt sind (d. h. Autoren verzichten auf die Zitation, weil Konzepte als allgemein bekannt vorausgesetzt werden). Zudem gibt es nach Umstätter (2005) das

Phänomen des bewussten Ignorierens: Artikel werden nicht zitiert, weil z.B. die Konzepte den eigenen eines Autors ganz oder teilweise widersprechen, fehlerhafte Theorien aussterben sollen oder die Falsifikation nicht geleistet werden kann.

Ein *complete linkage*-Clusteralgorithmus wird mit großer Wahrscheinlichkeit auch eigentlich zusammengehörende Gebiete aus den genannten Gründen trennen, er ist deshalb nicht adäquat für Zitationsdaten. Tatsächlich bestätigt die in der Literatur genannte Eigenschaft der dilatierenden Clusterbildung, die zu gleichmäßig kleinen Clustern führt, die inhaltlichen Überlegungen.

Das Clusterkriterium des *bi-connected components*-Algorithmus ist ebenfalls statisch. Die theoretische Begründung – die Dichtestrukturen der großen *bi-connected components,* die durch das Kriterium konstituiert werden, stabilisierten die components gegen Ausfälle von einzelnen Vertizes[41] – erscheint zu unspezifisch. Tatsächlich verhält sich der Algorithmus, wie beschrieben, ähnlich wie der *single linkage*-Algorithmus; Differenzen im Resultat sind wohl überwiegend auf größere Mindestclustergröße von drei Mitgliedern zurückzuführen. Es fällt auf, dass diese Definition genau der Haupteigenschaft skaleninvarianter Netze, der Stabilität gegenüber Zufallsausfällen, entspricht. Wenn Zitationsgraphen tendenziell skaleninvariante Netze sind, kann der bi-connected components-Algorithmus kaum strukturentdeckend sein, wie auch die empirischen Ergebnisse zu bestätigen scheinen. Für ihn gilt daher die gleiche Bewertung wie für *single linkage*, nämlich dass das Clusterkriterium für Zitationsgraphen zu schwach ist.

Einige Algorithmen operieren mit statischen Anteilen im Clusterkriterium, so die Clusterkern-Definition Gmürs (2004) und des Algorithmus von Radicchi et al. (2004). Dieser, der wie der „fast algorithm“ von Newman (2004b) den als erfolgreich geltenden GN-Algorithmus mit einfacheren Mitteln simuliert, normalisiert die vorhandenen Dreiecke mit einer hypothetischen Anzahl von Dreiecken, die anhand des degrees der beiden adjazenten Vertizes berechnet werden, d. h. der Dichte der direkten Umgebung. Damit richtet

41 ´This bi-connectedness stabilizes the cluster against changes in the initial selection when producing the database. Thus, the inclusion or exclusion of journals by ISI would not directly affect the large bi_components in the network data' Leydesdorff, 2004a

sich die Konstitution der Cluster einerseits nach der Dichte der Umgebung, ist aber andererseits von der Existenz zirkulärer Strukturen abhängig: In einem Netzwerk ohne solche kann der Algorithmus keine optimalen Ergebnisse bringen. So funktioniert er z.B. in einem Netzwerk von Sexualkontakten erwiesenermaßen nicht. Auch für Kozitationsgraphen lässt sich eine immanente Notwendigkeit zirkulärer Strukturen nicht deduzieren, wenn auch die empirischen Resultate zumindest für Kollaborationsnetzwerke Adäquatheit indizieren (z.B. die recht gute Übereinstimmung beider fast algorithms zueinander und zum GN-Algorithmus und die Tatsache, dass beide *power law*-Verteilungen generieren, was sich gut in die vorhandenen Modelle einfügen lässt (vgl. 9.1).

Der statische Anteil der Clusterdefinition Gmürs sieht auch ein komplett verlinktes Dreieck oder alternativ eine sternförmige Fünfer-Struktur vor, beide können ebenfalls nicht theoretisch deduziert werden.

Alle erwähnten statischen Clusterkriterien erscheinen entweder zu schwach oder zu stark oder zumindest mit einer gewissen Willkür behaftet und nicht begründbar – eine empirische Analyse über die Häufigkeit von triangulären, rechteckigen oder sternförmigen Strukturen in Kozitationsgraphen könnte jedoch genaueren Aufschluss über die Legitimität der genannten Kriterien geben.

Das Clusterkriterium relativ zu den gegebenen Ähnlichkeits- bzw. Dichtestrukuren eines Graphen zu bestimmen, scheint angesichts der Tatsache, dass Cluster in Zitationsgraphen formal-strukturell nicht exakt definiert werden können, die bessere Option zu sein. Bei den hierarchisch-agglomerativen Verfahren *average*, *centroid*, *median linkage* sowie beim *Ward-Algorithmus* ist dies der Fall ebenso wie bei beim *GN*- und *Newman-Algorithmus*. Da *centroid*, *median* und der *Ward-Algorithmus* auf euklidischen Distanzen basieren, wäre von den statistischen Verfahren *average* zu bevorzugen, welches mit dem Kosinus-Koeffizienten angewendet werden kann.

Die Verkettungstendenz hochgradig verlinkter Vertizes wird beim *GN*-Algorithmus und den Algorithmen von Radicchi et al (2004) und Newman (2004b), die sich an ihm orientieren, durch die Löschung von Kanten mit

Brückenfunktion nivelliert. So bleiben Binnenstrukturen zurück, in denen Vertizes in einem variablen Dichtegrad miteinander verlinkt sind, während Gruppen, die nur über einen oder wenige Links verbunden sind, auseinander fallen. Das ist prinzipiell ein sinnvolles Clusterkriterium, das den Eigenschaften der Zitationsdaten gut entgegenkommt.

Bei allen hierarchisch-agglomerativen Algorithmen wird eine Clusterdistribution nicht über alle möglichen Kombinationen optimiert, da dies sehr rechenintensiv ist; die Clusterbildungen basieren immer auf den vorherigen Clusterungen. Nur bei den Algorithmen *single linkage* und *complete linkage* gibt es faktisch keinen Einfluss der Reihenfolge, weil die beiden Kriterien statisch sind: Unabhängig davon, welche Objekte auf welcher Stufe fusioniert werden, müssen immer die gleichen Partitionen resultieren.

Ein Clusterverfahren, das hierarchisch oder nur mit lokalen Informationen arbeitet, wird mit großer Wahrscheinlichkeit einem Verfahren, das alle Clusterbildungen über den gesamten Graphen optimiert, unterlegen sein, wie es Radicchi et al. und Newman auch in Bezug auf ihre fast algorithms gegenüber dem *GN-Algorithmus* konstatieren.

Der divisive *GN-Algorithmus* führt eine Optimierung des gesamten Systems durch, indem *betweenness*-Werte immer wieder auf das ganze System bezogen berechnet werden. Er ist deshalb auch in seiner Kapazität sehr stark begrenzt und kann auf Artikel-Kozitationsstudien nicht angewandt werden.

Ein Nachteil vieler Algorithmen ist, dass sie viele unverbundene Vertizes, die im Lauf der Erhöhung der Kozitationsschwelle von Clustern abgespalten werden, generieren. Während beim *GN-Algorithmus* und beim *Radicchi-Algorithmus* Vertizes mit nur einer Kante durch diese immer einer community verbunden bleiben, geht bei den hierarchisch-agglomerativen Verfahren die relationale Umgebung dieser schwach verlinkten Vertizes in der finalen Clusterdistribution und damit für die weitere Analyse verloren; wie es auch in der von uns benutzten *single linkage*-Variante mit festen Schwellwerten der Fall ist. Die beiden graphanalytischen Verfahren *bi-connected components* und *Clusterdichte*, in Kombination mit Schwellwerten bzw. dem

graphanalytischen Dichtekriterium, erzeugen ebenfalls eine massive Anzahl unverbundener Items.
Bei der Faktorenanalyse/PCA kommt es dagegen zu keiner Isolierung von Variablen, i. d. R. gehen alle in die Faktorendistribution ein. Allerdings erweist sich die *PCA/Faktorenanalyse* aufgrund ihrer Anforderungen an die Homogenität der Datenstruktur (Normalverteilung), der Fokussierung auf lineare Beziehungen sowie methodischer Probleme – der Überlappung von Clusterstrukturen – als problematisch. Überdies hat die PCA eine sehr begrenzte Kapazität der Verarbeitung der Variablen.

9.3. Fazit

Für eine Weiterentwicklung des methodischen Gerüstes der Studie wären m. E. der *GN-Algorithmus*, falls ein bibliographisches Kopplungsset quantitativ durch ihn berechenbar wäre, oder einer oder beide der graphanalytischen fast algorithms interessant, weil ihre Clusterkriterien theoretisch plausibel erscheinen und die Vermeidung singulärer Items positiv zu bewerten ist. Die resultierenden *power law*-Verteilungen der Kollaborationsnetzwerke bieten interessante Voraussetzungen: Es ist wahrscheinlich, dass die Clusteralgorithmen bei Kozitations- oder Bibliographischen Kopplungs-Netzwerken vergleichbare Verteilungen erzeugen. Diese Übereinstimmung mit theoretischen Konzepten wäre ein interessanter Befund. Eine mögliche Alternative, mit dem Nachteil einer nicht so guten *coverage* der Clusterdistribution, wäre *average linkage*.
Es spricht nichts gegen die Weiterverwendung des Kosinus-Koeffizienten als Ähnlichkeitsmaß.
Da sich unser Interesse nur auf die *research fronts* bezieht, wäre ein kompletter Wechsel zu bibliographischer Kopplung eine nahe liegende Perspektive. Zwar weist Jarneving keine strukturellen Unterschiede nach, doch ist das Verfahren der Kombination der beiden Methoden umständlich und auch methodisch angreifbar: Die einfache Kopplung über das Kozitationscluster – d. h., dass es ausreicht, nur einen Artikel aus einem Cluster zu zitieren - ist sehr schwach und lässt die Frage offen, wie kohärent

die daraus resultierenden Strukturen wirklich sind. Des weiteren sind die Überschneidungen zwischen den Artikeln in Hinsicht auf die Validität der *research fronts*-Distribution problematisch, andererseits ist ein befriedigendes Behandeln dieser Überschneidungen schwierig: Das aktuelle Vorgehen beruht auf der Überlegung, dass, je mehr Referenzen aus unterschiedlichen Clustern ein Artikel zitiert, er inhaltlich desto weniger aussagekräftig für oder zugehörig zu jeder einzelnen *research front* ist. Allerdings bewerten wir das Ausmaß der Zugehörigkeit für diejenigen Artikel, die nur ein Cluster zitieren, nicht, obwohl es angesichts der begonnenen Differenzierung konsequent erschiene, diese von der Anzahl der zitierten Referenzen abhängig zu machen. Das entspräche bibliographischer Kopplung.

Eine weitere Optimierungsmöglichkeit bestünde darin, die Artikel des Untersuchungsdatensets nicht, wie bisher über die Zugehörigkeit zu vorher ermittelten Fachjournalen zu selektieren, sondern die Artikel über eine Titelwortrecherche direkt aus dem SCI zu recherchieren. Es wäre nach dem *Bradford-Gesetz* korrekter, in die Artikelsuche auch solche mit einzubeziehen, die möglicherweise in anderen Journalen publiziert wurden. In unserem setting gibt es dagegen den reziproken Effekt, dass sich im Datenset in geringem Ausmaß auch eher fachgebietsfremde Artikel befinden könnten. Möglicherweise gleichen sich Auswirkungen also auf die Diversität aus. Durch eine direkte Artikelsuche würde das Problem der unsicheren Feldabgrenzung auf der Journalebene umgangen.

10. Literaturverzeichnis

Ahlgren, P., B, Jarneving and R. Rousseau (2003): Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. Journal of the American Society of Information Science and Technology 54 (2003), *6,* 550-560

Amano, K., H. Nakamura and H. Ichikawa (2003): Self-Organizing Clustering: A Novel Non-Hierarchical Method for Clustering Large Amounts of DNA Sequences. Genome Informatics 14(2003), 575-576

Anderberg, M. R. (1973): Probability and Mathematical Statistics. New York: Academic Press, 1973 (A Series of Monographs and Textbooks)

Bacão, F., V. Lobo and M. Painho (2005): Self-organizing Maps as Substitutes for K-Means Clustering. Lecture Notes on Computer Science 3516(2005), 476-483

Backhaus, K., B. Erichson and W. Plinke and R. Weiber (2003): Multivariate Analysemethoden. Berlin et al.: Springer, 2003

Barabási, A. - L. and R. Albert (1999): Emergence of Scaling in Random Networks. Science 286(1999), 509-512

Barabási, A.-L. and E. Bonabeau (2004): Skalenfreie Netze. Spektrum der Wissenschaft 7(2004), 62-69

Batagelj, V. and Mrvar, A. (1998): Pajek – Program for large network analysis

Bensman, S. J. (2004): Pearson`s r and the Author Cocitation Analysis: A Commentary on the Controversy. Journal of the American Society of Information Science and Technology 55(2004)*10*, 935-936

Börner, K., C. Chaomei and K. W. Boyack (2003): Visualizing Knowledge Domains. Annual Review of Information Science and Technology 37(2003), 179-255

Bortz, J. (2005): Statistik für Human- und Sozialwissenschaftler. Heidelberg et al.: Springer, 2005

Boyack, K. W., B. N. Wylie and G. S. Davidson (2002): Domain Visualization with VxInsight. Journal of the American Society of Information Science and Technology 53(2002)*9*, 764-774

Boyack, K. W. (2004): Mapping knowledge domains: Characterizing PNAS. Proceedings of the National Academy of Sciences 101(2004)*Suppl. 1*, 5192-5199

Boyack, K. W., R. Klavans and K. Börner (2005): Mapping the Backbone of Science. Scientometrics 64 (2005)*3*, 351-374

Braam, R. R., H. F. Moed and F. J. van Raan (1991): Mapping of Science by Combined Cocitation and Word Analysis. I. Structural Aspects. Journal of the American Society of Information Science 42(1991)*4*, 233-251

Chakrabarti, S. (2003): Mining the Web. Discovering Knowledge from Hypertext Data. Amsterdam (et al.): Morgan Kaufman, 2003

Davidson, G. S., B. N. Wylie and K. W. Boyack (2001): Cluster Stability and the Use of Noise in Interpretation of Clustering. Proceedings of IEEE Information Visualization 2001, 23-30

Davidson, G. S., B. Hendrickson and D.K. Johnson and C.E. Meyers and B.N. Wylie (1998): Knowledge Mining with VxInsight: Discovery through Interaction. Journal of Intelligent Information Systems 11(1998), 259-285

De Nooy, W., A. Mrvar and V. Bagatelj (2005): Exploratory Social Network Analysis with Pajek. Cambridge: Cambridge University Press, 2005 (Structural Analysis in the Social Sciences, 27)

Dillon, W. R. and M. Goldstein (1984): Multivariate Analysis. Methods and Applications. New York (et al.): Wiley, 1984 (Wiley Series in Probability and Mathematical Statistics)

Dominich, S. (2001): Mathematical Foundations of Information Retrieval. Dordrecht et al.: Kluwer, 2001 (Mathematical Modelling and Applications)

Eades, P. (1984): A Heuristic for Graph Drawing. Congressus Nutnerantiunt 42(1984), 149-160

Eckey, H.-F., R. Kosfeld and M. Rengers (2002): Multivariate Statistik. Grundlagen – Methoden – Beispiele. Wiesbaden: Gabler, 2002

Egghe, L. and R. Rousseau (2003): A Measure for the Cohesion of Weighted Networks. Journal of the American Society of Information Science and Technology 54(2003)*3*, 193-202

Everett, M. and S.P. Borgatti (2005): Extending Centrality. P.J. Carrington, J. Scott and S. Wassermann: Models and Methods in Social Network Analysis. Cambridge: University Press, 2005

Fruchterman, T.M.J. and E.M. Reingold (1991): Graph Drawing by Force-directed Placement. Software – Practice and Experience 21(1991)*11*, 1129-1164

Girvan, M. and M. E. J. Newmann (2002): Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences 99(2002)*12*, 7821-7826

Gmür, M. (2004): Cocitation Analysis and the Search for Invisible Colleges: A Methodological Evaluation. Scientometrics 57(2004)*1*, 27-57

Griffith, B.C. and H.D. White (1983): Authors as Makers of Intellectual Space: Co-Citation in Studies of Science, Technology and Society. Journal of Documentation 38(1983)*4*, 255

Jarneving, B. (2005): A Comparison of two Bibliometric Methods for Mapping of the Research front. Scientometrics 65(2005)*2*, 245-263

Katz, J. S. (1999): The self-similar science system. Research Policy 28(1999), 501-517

Klavans, R. and K.W. Boyack (2005): Mapping World-wide Science at the Paper Level. Proceedings International Society for Scientometrics and Infometrics, 2005, 426-436

Klavans, R. and K. W. Boyack (2006): Identifying a better Measure of Relatedness for Mapping Science. Journal of the American Society of Information Science and Technology 57(2006)*2*, 251-263

Kostoff, R.N., R. Tshiteya, K.M. Pfeil and J. A. Humenik (2002): Electrochemical Power Text Mining using Bibliometrics and Database Tomography. Journal of Power Sources 110(2002)*1*, 163-176

Leydesdorff, L. (2004a): Clusters and Maps of Science Journals Based on Bi-Connected Graphs in the Journal Citation Reports. Journal of Documentation 60(2004)*4*, 317-427

Leydesdorff, L. (2004b): Top-Down Decomposition of the Journal Citation Reports of the Social Science Citation Index: Graph- and Factor-Analytical Approaches. Scientometrics 60(2004)*2*, 159-180

Leydesdorff, L. (2005): Similarity Measures, Author Cocitation Analysis, and Information Theory. Journal of the American Society of Information Science and Technology 56(2005)*7*, 769-772

Leydesdorff, L. (2006): Can Scientific Journals be Classified in terms of Aggregated Journal-Journal Citation Relations using the Journal Citation Reports? Journal of the American Society of Information Science and Technology 57(2006)*5*, 601-613

Leydesdorff, L and S. E. Cozzens (1993): The Delineation of Specialties in Terms of Journals using the Dynamic Journal Set of the SCI. Scientometrics 26(1993)*1*, 135-156

Leydesdorff, L. and S. Bensman (2006): Classification, Powerlaws: The Logarithmic Transformation. Journal of the American Society of Information Science and Technology 57(2006)*11*, 1470-1486

Leydesdorff, L. and L. Q. Vaughan (2006): Co-occurrence matrices and their Applications in Information Science: Extending ACA to the Web Environment. Journal of the American Society of Information Science and Technology 57(2006)*12*, 1616-1628

Liu, Zao (2005): Visualizing the Intellectual Structure in Urban Cities: A Journal Cocitation (1992-2002). Scientometrics. 62(2005)*3*, 385-402

Marshakova, I. V. (1973): System of Document Connections Based on References. Nauchno-Tekhnicheskaya Informatsiya Seriya 2– Informatsionnye Protsessy I Sistemy, 2(1973)*6*, 3-8

Merelo-Guervos, J. J., B. Prieto and A. Prieto and G. Romero and C. Valdiviesco and F. Tricas (2004): Clustering Web-Based Communities using Self-Organizing Maps. Proceedings IADIS International Conference Web Based Communities 2004, 158-165

M. Mitzenmacher (2005): A Brief History of Generative Models for Power Law and Lognormal Distributions. http://citeseer.ist.psu.edu/461916.html

Newman, M. E. J. (2004a): Detecting Community Structure in Networks. The European Physical Journal B 38 (2004) 321-330

Newman, M. E. J. (2004b): Fast Algorithm for Detecting Community Structures in Networks. Physical Review E 69(2004) *066133*

Newman, M. E. J and M. Girvan (2004): Finding and Evaluating Community Structure. Physical Review E 69(2004) *026113*

Newman, M. E. J. (2005) (: A Measurement of Betweenness Centrality based on Random Walks. Social Networks 27(2005), 39-54

Persson, O. (2001): All Author Citations versus First Author Citations. Scientometrics 50(2001), 339-344

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto and D. Parisi (2004): Defining and Identifying Communities in Networks. Proceedings of the National Academy of Sciences 101(2004)*9*, 2658-2663

Schmidt, M, J. Gläser, F. Havemann and M. Heinz (2006): A Methodological Study for Measuring the Diversity of Science. Proceedings of the International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, 129-137

Small, H. (1973): Cocitation in the Scientific Literature. Journal of the American Society of Information Science 24(1973), 265-269

Small, H. (1978): Cited Documents as Concept Symbols. Social Studies of Science 8(1978)*3*, 327-340

Small, H. (1993): Macro-Level Changes in the Structure of Cocitation Clusters: 1983-1989. Scientometrics 26(1993)*1*, 5-20

Small, H. (2003): Paradigms, Citations and Maps of Science: A Personal History. Journal of the American Society of Information Science and Technology 54(2003)*5*, 394-399

Small, H. and B. C. Griffith (1974a): The Structure of Scientific Literatures I: Identifying and Graphing Specialties. Science Studies 4(1974), 17-40

Small, H. B. C. Griffith, J. A. Stonehill and Sandra Dey (1974b): The Structure of Scientific Literatures II: Toward a Macro- and Microstructure. Science Studies 4(1974)*4*, 339-365

Small, H. and E. Sweeney (1985a): Clustering the Science Citation Index using Cocitations. Scientometrics 7(1985)*3-6*, 391-409

Small, H. E. Sweeney and E. Greenlee (1985b): Clustering the Science Citation Index using Cocitations II: Mapping Science. Scientometrics 8(1985)*5-6*, 321-340

Theil, Henri (1972): Statistical Decomposition Analysis. With Applications in the Social and Administration Sciences. Amsterdam, London: North-Holland, 1972 (Studies in Mathematical and Managerial Economics, 14)

Rousseau, R. and p. Van Hecke (1999): Measuring Biodiversity. Acta Biotheoretica 47(1999), 1-5

Van Raan, A. F. J. (1991): Fractal Geometry of Information Space as represented by Cocitation Clustering. Scientometrics 20(1991)*3*, 439-449

Van Raan, A. F. J. (2000): On Growth, Aging, and Fractal Differentiation of Science. Scientometrics 47(2000)*2*, 347-362

Van Raan, A. F. J. (2005): Reference-based Publication Networks with Episodic Memories. Scientometrics, 63(2005)*3*, 549-566

Wang, Y. C. Yang, K. Mathee and G. Narasimhan (2005): Clustering using Adaptive Self-Organizing Maps (ASOM) and Applications. Lecture Notes in Computer Science 35(2005)*15*, 944-951

Weingart, P. R. Sehringer and M. Winterhager (1988): Bibliometric Indicators for Assessing West German Science. In: A. F. J. Raan: Handbook of Quantitative Studies of Science and Technology. Amsterdam et al.: North-Holland, 1988

White, H. D. (2003): Author Cocitation Analysis and Pearson`s r. Journal of the American Society of Information Science and Technology 54(2003)*13*, 1250-1259

White, H. D. and B. C. Griffith (1981): Author Cocitation: A Literature Measure of Intellectual Structure. Journal of the American Society of Information Science 32(1981)*5*, 163-171

Umstätter, W. (2005): Die Bedeutung des Bradford Law of Scattering für die Bibliothekswissenschaft. http://www.ib.hu-berlin.de/%7Ewumsta/infopub/pub2001f/Bradford05fold.pdf

Umstätter, W. (2004): Szientometrische Verfahren. In: R. Kuhlen: Grundlagen der praktischen Information und Dokumentation. 5., vollst. neu gefasste Aufl. München: Saur, 2004

11. Anhang

A. Heuristische Bestimmung der Schwellwerte von Small & Sweeney und Resultate der Vergleichsstudie

Der aus den Voruntersuchungen ermittelte Integerschwellwert ist ≥ 17; der Fraktionalschwellwert wird auf der Basis der Konvergenz in der absoluten Anzahl der selektierten Zitationen (*citations at threshold*) auf ≥ 0.77 gesetzt; d. h. die beiden Schwellwerte selektieren eine fast äquivalente Anzahl an Zitationen (etwa 700 000). Von diesen schwankt die Anzahl der einzelnen zitierten Artikel (*No. of cited documents selected at threshold*) wegen des Normalisierungseffektes der fraktionalen Zählweise zwischen 23440 (integer) und 43931 (fraktional). Sie ist bei der fraktionalen Zählweise höher, weil mehr niedrig zitierte Artikel selektiert werden. Die Zahl der selektierten distinkten Kozitationspaare schwankt zwischen 1 834 390 (integer) und 1 103 607 (fraktional). Als Cluster-Maximalgröße wird 49 festgelegt. Die Kosinus-normalisierten Kozitationsschwellwerte bzw. *starting levels* werden analog über eine definierte Gesamtanzahl der in Verbindung mit dem jeweiligen Schwellwert selektierten distinkten Kozitationspaare bestimmt (47 000) – der Kosinus-Koeffizient im *fractional/variable*-Modell ist ≥ 0.18 und konvergiert mit ≥ 0.28 beim *fractional/constant*-Modell, mit ≥ 0.224 beim *integer/variable*-Modell und ≥ 0.33 beim *integer/constant*-Modell. Die 47 000 Paare der Kosinus-Eingangsschwelle entsprechen 2.56% bzw. 4.26% der über die Integer- bzw. fraktionale Schwelle selektierten Paare.

B. Schwellwertempfehlungen von Small & Sweeney

Small & Sweeney empfehlen zu einem wegen der effektiven Kürzung der Referenzenlisten erhöhten Fraktionalschwellwert $f \geq 1.5$ eine Integer-Zitationsschwelle $c \geq 5$. Für die referierten Experimente wurde bei der *fraktional*/*variable*-Kombination eine maximale Clustergröße von 49, als Eingangs-Kozitations-Level ≥ 0.18 und ein Increment-Wert von 0.01

verwandt. Als generelle Empfehlung geben Small & Sweeney an, dass, je höher die Maximal-Größe, desto größer die *coverage*, also die Anzahl der in Clustern gebundenen Items, sei; die maximale Größe daher so groß wie möglich sein sollte, ohne Makrocluster zu kreieren. Hohe Maximalgrößen vermindern die Anzahl der Cluster; niedrige Starting Levels der Kozitationsschwelle vergrößern Recall und Clusteranzahl, aber voraussichtlich auch die Diversität der Cluster.

C. Richtlinien zur Faktorinterpretation

Nach Bortz (2005) sollten mindestens 4 Ladungen auf einem Faktor einen Wert von über 0.60 oder mindestens 10 Ladungen einen Wert von über 0.40 aufweisen; wenn weniger als 10 Variablen Ladungen über 0.40 haben, sollte nur interpretiert werden, wenn die Stichprobe aus mindestens 300 Objekten besteht, weil sonst eine zufällige Ladungsstruktur vorliegen kann. Ladungen unter 0.40 sollten bei der Faktorinterpretation nicht berücksichtigt werden. Wenn eine Variable nach der Rotation mittlere Ladungen auf mehreren Faktoren hat, sollte sie nach dem Fürntratt-Kriterium (Bortz 2005) nur dann einem Faktor zugeordnet werden, wenn mindesten 50% ihrer Varianz auf diesen entfallen.

D. Programmierarbeiten

D.1. Datenimport

```
#!/usr/bin/perl
#liest Datensätze aus Textfile in die Datenbank ein
use DBI;
use strict;
#Regulierung, ob Testdurchlauf (vorherige Tabellen werden geloescht)
my ($debug) = 1;
my ($fn) = "./daten/docs_1998.txt";
my ($infile) = $fn;
my ($outfile) = $fn."-control.txt";
my ($dsn)= "DBI:mysql:SCI98:localhost"; # data source name
my ($user_name) = "rot";
my ($pass_word) = "fh123!";
```

```
my ($dbh, $sth);
my ($sql);
my (@ary);
my ($rv);
my ($key);
my (%content);
my (@sources);
my (@authors);
my (@addresses);
my (@references);
my ($item);
my ($garbage) = "";
my (@marker) = qw(
              id ti qu dt sp ye vo is sa se be au ad ci
                 );
my ($maxf) = 13;
my ($field_no) = 0;
my ($rec_no) = 0;
my ($i) = 0;
my ($auth_no) = 0;
my ($addr_no) = 0;
my ($ref_no) = 0;
my ($source);
my ($abb);
my ($jou_id);
my ($aut_id);
my ($add_id);
my ($ref_id);
my ($year);
#Datenbankkomunikation
$dbh=DBI->connect($dsn, $user_name, $pass_word, {RaiseError=>0}) or die "Couldn`t
connect to database".DBI->errstr;
#Loeschung temporaer genutzer Tabellen
if($debug == 2) {
    #!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
        $sql = "DELETE FROM `records`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `journals`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `jou_rec`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `authors`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `aut_rec`;";
        $sth=$dbh->prepare($sql);
```

```
        $sth->execute();
        $sql = "DELETE FROM `refs`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `ref_rec`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `addresses`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $sql = "DELETE FROM `add_rec`;";
        $sth=$dbh->prepare($sql);
        $sth->execute();
        #!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
}
# infiles und outfiles:
open(IN,$infile) || die "Cannot open $infile.";
open(OUT,">$outfile")|| die "Cannot make $outfile.";
while(<IN>){  #input next line
##-- Parsen des Texts
        chomp;
        s/\r//g;
        s/\"/\'/g;
        $key = substr($_,0,2); # Beginn der Zeile (die ersten beiden Buchstaben #sind
Feldnamen -
        # falls Zeile nicht mit 3 Leerzeichen beginnt (dann ist es eine Leerzeile)
        if($key ne ""){ #(not $_ =~ /^   /i) ||
                #print "DBG: key = '$key'\n";
#vordefinierte Liste der 13 Feldnamen wird iteriert
while($field_no <= $maxf){
if($key eq $marker[$field_no]){
                                                                                last;              }
else{
$field_no++;
last;                                                                           }
                                                                                   }
         if($key ne "be" && $key ne $marker[$field_no]){
                                $field_no--;
           if($marker[$field_no] eq "au"){
                $auth_no = @authors-1;
                                $authors[$auth_no] .= $_;
                        }
                        elsif($marker[$field_no] eq "ad"){
                               $addr_no = @addresses-1;
                                $addresses[$addr_no] .= $_;
                        }
                        elsif($marker[$field_no] eq "ci"){
                                $ref_no = @references-1;
                                $references[$ref_no] .= $_;
                        }
```

```
elsif($marker[$field_no] eq "id") {
                        $content{"id"} .= $_;
                                $year = substr($content{"id"}, 3, 4);
                           }
                           else{
                                $content{$marker[$field_no]} = $_;
                           }

                }
                else{
                        if($key eq "au"){
                                push(@authors, substr($_,3));
                        }
                        elsif($key eq "ad"){
                                push(@addresses, substr($_,3));
                        }
                        elsif($key eq "ci"){
                                push(@references, substr($_,3))
                        }
elsif($key eq "id") {
                                $content{"id"} = substr($_,8);
                                $year = substr($_, 3, 4);
                        }
                        else{
                                $content{$key} = substr($_,3);

                        }
                }

        }
        else{
 @sources = split /;/, $content{qu};
                $source = $sources[0];
                $abb = $sources[1];
                if(
                $content{qu}                                                                    =~
/(.*?)electroch(.*?)|(.*?)elektroch(.*?)|(.*?)electroanal(.*?)|(.*?)elektroanal(.*?)|
(.*?)j(.*?)new(.*?)mat(.*?)electr(.*?)sys|(.*?)j(.*?)power(.*?)sources|(.*?)solid(.*?
)state(.*?)electr(.*?)|(.*?)plat(.*?)surf(.*?)finish|(.*?)prot(.*?)met+(.*?)|(.*?)sol
id(.*?)state(.*?)ionics(.*?)|t(.*?)I(.*?)met(.*?)finish|(.*?)chem(.*?)vapor(.*?)depos
ition|(.*?)sensor(.*?)actuator(.*?)b-chem(.*?)|(.*?)corros(.*?)sci/i
                ){
                &db_eintraege();
                }
                &reset_vars();
        }
} # END: while(<IN>)
# nochmal den letzten in die DB, weil nach dem letzten keine Leerzeile kommt
@sources = split /;/, $content{qu};
                $source = $sources[0];
```

```
                $abb = $sources[1];
                $content{qu}     =~     /(.*?)electroch(.*?)|(.*?)elektroch(.*?)|(.*?)
electroanal(.*?)|(.*?)elektroanal(.*?)|j new mat electr sys|j power sources|j solid
state electr|plat surf finish|prot met+|solid state ionics|t i met finish|chem vapor
deposition|sensor actuator b-chem |corros sci/i
                ){
                &db_eintraege();
                }
&reset_vars();
##########################################Garbage################
print                    "*****************************\n==                    GARBAGE
==\n$garbage\n*****************************\n\n\n";
$dbh->disconnect();
exit(0);
        #Subroutinen
#Leerung der arrays
sub reset_vars{
        $rec_no++;
        $field_no = 0;
        $jou_id = 0;
        foreach $key (keys(%content)){
                $content{$key} = "";
        }
        foreach $item (@authors){
                $item = "";
        }
        @authors = ();
        foreach $item (@addresses){
                $item = "";
        }
        @addresses = ();
        foreach $item (@references){
                $item = "";
        }
        @references = ();
        print "\n\n\n======== rec($rec_no) zuende ==========\n\n";
}
sub exec_sql_select{
        $sth=$dbh->prepare($sql);
        $sth->execute();
        $rv=$sth->fetchrow_array();
        $sth->finish();
}
sub db_eintraege{
##-- Datenbank Eintraege
        # wenn record schon in records, exit;
        $sql = qq{select `rec_id` from `records` where `rec_id` = '$content{id}'};
        &exec_sql_select();
        if($rv){
```

```
                print "\n!!!!! Record mit der ID '$content{id}' existiert schon in
Tabelle records!\n\n\n";
        }
        else{
        print "\tstarte DB-Eintrag mit der ID '$content{id}'!\n";
                $ref_no = @references;
                $auth_no = @authors;
                $addr_no = @addresses;
                # schreiben von Eintrag in records
                $sql = qq{insert into `records`  ( `title` , `language` , `doc_type`
, `nr_refs` , `nr_authors` , `nr_addresses` , `rec_id` )
                values (?, ?, ?, ?, ?, ?, ?)};
                $sth = $dbh->prepare($sql);
                $sth->bind_param(1, $content{ti});
                $sth->bind_param(2, $content{sp});
                $sth->bind_param(3, $content{dt});
                $sth->bind_param(4, $ref_no);
                $sth->bind_param(5, $auth_no);
                $sth->bind_param(6, $addr_no);
                $sth->bind_param(7, $content{id});
                $sth->execute();
                if($debug == 2) {print "SQL = $sql\n\n";}
                # schauen, ob Journal schon existiert
                $sql = qq{select `jou_id` from `journals` where `journal` = '$source'
and `short_name` = '$abb'};
                &exec_sql_select();
                #wenn ja
                if($rv){
                        #jou_id merken
                        $jou_id = $rv;
                }
                #wenn nein
                else{
                        #hoechste jou_id holen+1, neu schreiben, neue jou_id #merken
                        $sql = qq{select max( `jou_id` )FROM `journals`};
                        if($debug == 2) {print "SQL = $sql\n\n";}
                        &exec_sql_select();
                        if($debug == 2) {print "max_jou_id = $rv\n";}
                        $jou_id = ++$rv;
                       $sql  =  qq{insert  into  `journals`  (`jou_id`,  `journal`,
`short_name`)
                        values (?,?,?)};
                        $sth = $dbh->prepare($sql);
                        $sth->bind_param(1, $jou_id);
                        $sth->bind_param(2, $source);
                        $sth->bind_param(3, $abb);
                        $sth->execute();
                        if($debug == 2) {print "SQL = $sql\n\n";}

                }
```

```
		# schreiben von jou_id und rec_id in jou_rec
		$sql = qq{insert into `jou_rec` ( `rec_id` , `jou_id` , `vol` , `iss`
, `b_pag` , `e_pag` , `year` ) values('$content{id}', '$jou_id', '$content{vo}',
'$content{is}', '$content{sa}', '$content{se}', '$content{ye}')};
		if($debug == 2) {print "SQL = $sql\n\n";}
		$sth = $dbh->prepare($sql);
		$sth->execute();
#Autoren eintragen
		for($i = 0; $i < $auth_no; $i++){
			#gibt es den Autor schon in authors?
			$sql = qq{select `aut_id` from `authors` where `author` =
'$authors[$i]'};
			&exec_sql_select();
			#wenn ja, dann aut_id merken
			if($rv){
				$aut_id =$rv;
			}
			#wenn nein, abspeichern
			else{
				#hoechste aut_id holen+1, neu schreiben, neue aut_id
#merken
				$sql = qq{select max( `aut_id` )FROM `authors`};
				if($debug == 2) {print "SQL = $sql\n\n";}
				&exec_sql_select();
				$aut_id = ++$rv;
			      $sql = qq{insert into `authors` (`aut_id`, `author`)
				values (?, ?)};
				$sth = $dbh->prepare($sql);
				$sth->bind_param(1, $aut_id);
				$sth->bind_param(2, $authors[$i]);
				$sth->execute();
				if($debug == 2) {print "SQL = $sql\n\n";}
			}
			#Eintrag in aut_rec mit rec_id und aut_id
			$sql = "insert into `aut_rec` (`rec_id`, `aut_id`, `aut_pos`)
values('$content{id}', '$aut_id', '".(1+$i)."')";
			if($debug == 2) {print "SQL = $sql\n\n";}
			$sth = $dbh->prepare($sql);
			$sth->execute();
		}
		#Addressen eintragen
		for($i = 0; $i < $addr_no; $i++){
			#gibt es die Adresse schon in addresses?
			$sql = qq{select `add_id` from `addresses` where `address` =
'$addresses[$i]'};
			&exec_sql_select();
			#wenn ja, dann add_id merken
			if($rv){
				$add_id =$rv;
			}
```

```
                        #wenn nein, abspeichern
                        else{
                                #hoechste add_id holen+1, neu schreiben, neue aut_id
#merken
                                $sql = qq{select max( `add_id` )FROM `addresses`};
                                if($debug == 2) {print "SQL = $sql\n\n";}
                                &exec_sql_select();
                                $add_id = ++$rv;
                             $sql = qq{insert into `addresses` (`add_id`, `address`)
                                values (?, ?)};
                                $sth=$dbh->prepare($sql);
                                $sth->bind_param (1, $add_id);
                                $sth->bind_param (2, $addresses[$i]);
                                $sth->execute();
                                if($debug == 2) {print "SQL = $sql\n\n";}
                        }

                        #Eintrag in add_rec mit rec_id und add_id
                        $sql = "insert into `add_rec` (`rec_id`, `add_id`, `add_pos`)
values('$content{id}', '$add_id', '".(1+$i)."')";
                        if($debug == 2) {print "SQL = $sql\n\n";}
                        $sth = $dbh->prepare($sql);
                        $sth->execute();
                }
                #Zitate eintragen
                for($i = 0; $i < $ref_no; $i++){
                        #gibt es das Zitat schon in refs?
                        $sql = qq{select `ref_id` from `refs` where `ref` =
'$references[$i]'};
                        &exec_sql_select();
                        #wenn ja, dann ref_id merken
                        if($rv){
                                $ref_id =$rv;
                                print "\t\tref ($ref_id) schon vorhanden!\n";
                        }
                        #wenn nein, abspeichern und ref_id merken
                        else{
                                #hoechste aut_id holen+1, neu schreiben, neue aut_id
merken
                                $sql = qq{select max( `ref_id` )FROM `refs`};
                                if($debug == 2) {print "SQL = $sql\n\n";}
                                &exec_sql_select();
                                $ref_id = ++$rv;
                                $sql = qq{insert into `refs` (`ref_id`, `ref`) values
(?, ?)};
                                $sth=$dbh->prepare($sql);
                                $sth->bind_param(1, $ref_id);
                                $sth->bind_param(2, $references[$i]);
                                $sth->execute();
                                if($debug == 2){print "SQL = $sql\n\n";}
```

```
                        }

                        #Eintrag in ref_rec mit rec_id und ref_id
                        $sql   =   "insert   into   `ref_rec`   (`rec_id`,   `ref_id`)
values('$content{id}', '$ref_id')";
                        if($debug == 2) {print "SQL = $sql\n\n";}
                        $sth = $dbh->prepare($sql);
                        $sth->execute();
                }
        }
}
```

D.2. Schwellwertberechnung(1)

```
#! /usr/bin/perl
#Schwellwerte fuer  Referenzen
#berechnet Integer-  und Fraktionalzitationsanzahl
use DBI;
use strict;
my ($debug) =1;
my ($dsn)="DBI:mysql:SCI98test:localhost";
my ($user_name)="rot";
my ($pass_word)="fh123!";
my ($dbh, $sth);
my ($sql);
my (@ary);
my ($rv);
my ($quotsum);
my ($ref_id);
#Verbindung zur Datenbank durch DBI-Modul
$dbh=DBI->connect($dsn, $user_name, $pass_word) or die "Couldn't connect
to database".DBI->errstr;
#Loeschung der temporaer genutzten SQL-Tabellen
if ($debug == 1) {
$sth=$dbh->prepare(qq{delete from threshold1});
$sth->execute();
$sth=$dbh->prepare(qq{delete from threshold2});
$sth->execute();
$sth=$dbh->prepare(qq{delete from threshold3});
$sth->execute();
$sth=$dbh->prepare(qq{delete from threshold4});
$sth->execute();
}
#Berechnung der Anzahl der Zitationen je Referenz
$sth=$dbh->prepare(qq{insert into threshold1 (nr_cit, ref_id) select
count(*), ref_id from ref_rec group by ref_id});
$sth->execute();
```

```
#1. Teilberechnung der fraktionalen Werte
#Berechnung fraktionaler Werte durch Division
#durch die Anzahl  aller Referenzen (1/n) je Artikel
#( im Moment nicht gebraucht)
$sth=$dbh->prepare(qq{insert into threshold2(rec_id, ref_id, quotient)
select records.rec_id, threshold1.ref_id,
(1/nr_refs) from threshold1, records, ref_rec where
threshold1.ref_id = ref_rec.ref_id and ref_rec.rec_id = records.rec_id
order by ref_id});
$sth->execute();
#2. Teilberechnung der fraktionalen Werte:
#Addition der fraktionalen Werte aller Artikel, in denen
#eine Referenz vorkommt
#( im Moment nicht gebraucht)
$sth=$dbh->prepare(qq{insert into threshold3(ref_id,
frac_cit_count, nr_cit) select threshold2.ref_id,
sum(quotient), threshold1.nr_cit from threshold2, threshold1 where
threshold1.ref_id = threshold2.ref_id group by
ref_id});
$sth->execute();
#Zusammenfassung der Informationen
$sth=$dbh->prepare(qq{insert into threshold4(rec_id, ref_id,
frac_cit_count, nr_cit) select ref_rec.rec_id, threshold3.ref_id,
threshold3.frac_cit_count, threshold1.nr_cit from ref_rec, threshold3,
threshold1 where
ref_rec.ref_id=threshold3.ref_id and threshold1.ref_id =
threshold3.ref_id});
$sth->execute();
$dbh->disconnect();
exit(0);
```

D.3. Schwellwertberechnung(2)

```
#! /usr/bin/perl
#Konstruktion der Kozitationspaare, Berechnung des Kozitationsschwellwertes
#die letzte Tabelle haelt alle Daten des Untersuchungssettungs vor, so dass
#die Programme muessen 1-3 bei Schwellwert-Variationen
#nicht wiederholt werden muessen
use DBI;
use strict;
my ($debug) = 1;
my ($dsn)="DBI:mysql:SCI98test:localhost"; # DB-Name
my ($user_name)="rot";
my ($pass_word)="fh123!";
my ($dbh, $sth);
my ($sql);
my (@ary);
my ($rv);
```

```
my ($ref_id);
#Verbindung zur Datenbank durch DBI-Modul
$dbh=DBI->connect($dsn, $user_name, $pass_word) or die "Couldn't connect
to database".DBI->errstr;
#Loeschung temporaer genutzter Tabellen
if ($debug == 1) {
$sth=$dbh->prepare(qq{delete from self_join});
$sth->execute();
$sth=$dbh->prepare(qq{delete from self_join2});
$sth->execute();
$sth=$dbh->prepare(qq{delete from self_join3});
$sth->execute();
$sth=$dbh->prepare(qq{delete from self_join4});
$sth->execute();
$sth=$dbh->prepare(qq{delete from auswertung});
$sth->execute();
$sth=$dbh->prepare(qq{delete from auswertung4});
$sth->execute();
}
#durch einen self-join auf rec_id werden Paare von Referenzen gebildet,
#die vom selben Artikel zitiert werden
$sql=qq{insert into self_join (ref1_id, ref2_id) select
m1.ref_id, m2.ref_id from threshold4 as m1, threshold4 as m2 where
m1.rec_id=m2.rec_id and m1.ref_id < m2.ref_id};
$sth=$dbh->prepare($sql);
$sth->execute();
print "$sql\n";
#Selektion der distinkten Paare
$sth=$dbh->prepare(qq{delete from self_join where ref1_id=ref2_id});
$sth->execute();
#Bestimmung der Kozitationsanzahl
$sql=qq{insert into self_join2 select ref1_id, ref2_id,
count(*) from self_join group by ref1_id, ref2_id};
$sth=$dbh->prepare($sql);
$sth->execute();
print "$sql\n";
#die Zitationsanzahl von ref1_id (des ersten Partners eines Paares)
#wird dazugeholt;
#dadurch werden fuer spaetere Zugriffe rechenintensive Tabellenjoins
#eingespart
$sql=qq{insert into self_join3 select self_join2.ref1_id,
self_join2.ref2_id, nr_cit, common_cit_count from self_join2,
threshold3_2 where self_join2.ref1_id=threshold3_2.ref_id};
$sth=$dbh->prepare($sql);
$sth->execute();
print "$sql\n";
#die Zitationsanzahl von ref2_id (des zweiten Partners eines Paares)
#wird dazugeholt;
$sql=qq{insert into self_join4 select self_join3.ref1_id,
self_join3.ref2_id, cit_count1, nr_cit, common_cit_count from
```

```
self_join3, threshold3_2 where self_join3.ref2_id=threshold3_2.ref_id};
$sth=$dbh->prepare($sql);
$sth->execute();
print "$sql\n";
#Berechnung der Kosinus-Normalisierung
$sql=qq{insert into auswertung select ref1_id, ref2_id,
cit_count1, cit_count2, common_cit_count,
(common_cit_count/SQRT(cit_count1*cit_count2)) from self_join4 order by
ref1_id};
$sth=$dbh->prepare($sql);
$sth->execute();
#aktuell gueltiger Schwellenwert: beide Partner eines Kozitationspaares
#muessen jeweils mehr als einmal zitiert werden
$sql=qq{insert into auswertung4 select * from auswertung where cit_count1 > '1' and
cit_count2 > '1'};
$sth=$dbh->prepare($sql);
$sth->execute();
$dbh->disconnect();
exit(0);
```

D.4. Clustervorbereitung

```
#! /usr/bin/perl
#bereitet import-file fuer Pajek vor
#in der Schleife am Anfang muss der Bereich der Kozitationsschwellwerte, #die
berechnet werden sollen, spezifiziert werden
use DBI;
use strict;
my ($debug) = 1;
my ($dsn)="DBI:mysql:SCI98test:localhost"; # DB-Name
my ($user_name)="rot";
my ($pass_word)="fh123!";
my ($dbh, $sth);
my ($rv);
my ($ref);
my (@rows);
my ($infile);
my ($i);
my ($j);
#zur Berechnung mehrerer Schwellwerte
for ($i = 0.1; $i < 0.9; $i=$i + 0.1) {
$j=$i;
$j=~ s/\.//;
#output-files
open(OUT,">./Daten/Vertices_$j.txt");
open(OUT2,">./Daten/Edges_$j.txt");
#Datenbank-Kommunikation
$dbh=DBI->connect($dsn, $user_name, $pass_word) or die "Couldn't connect
```

```
to database".DBI->errstr;
if ($debug == 1) {
#Loeschung temporaer genutzer Tabellen
$sth=$dbh->prepare(qq{drop table vertices$j});
$sth->execute();
$sth=$dbh->prepare(qq{drop table vertices2_$j});
$sth->execute();
$sth=$dbh->prepare(qq{drop table network1_$j});
$sth->execute();
$sth=$dbh->prepare(qq{drop table network2_$j});
$sth->execute();
}
#alle Referenzen, die mehr als einmal zitiert werden, sammeln:
#ref1_ids aus der Schluss-Tabelle des letzten Programms holen
$sth=$dbh->prepare(qq{create table vertices$j select distinct
ref1_id as ref_id from auswertung4 where square_root >'$i'});
$sth->execute();
#ref2_ids dazu holen
$sth=$dbh->prepare(qq{insert into vertices$j select distinct ref2_id as
ref_id from
auswertung4 where square_root > '$i'});
$sth->execute();
#doppelte Referenzen werden eliminiert
$sth=$dbh->prepare(qq{create table vertices2_$j select distinct ref_id
from
vertices$j order
by ref_id});
$sth->execute();
#Nummerierung der Referenzen: fuer Pajek notwendige Maskierung
$sth=$dbh->prepare(qq{alter table vertices2_$j add column nr int(5) not
null auto_increment, add primary key(nr)});
$sth->execute();
$sth=$dbh->prepare(qq{select max(nr) from vertices2_$j});
$sth->execute();
$rv=$sth->fetchrow_array;
#die Referenzen (Vertizes) werden in Textdatei ausgelesen
$sth=$dbh->prepare(qq{select nr, ref_id from vertices2_$j});
$sth->execute();
print OUT "*Vertices $rv\n";
while
(@rows=$sth->fetchrow_array()){
print OUT "$rows[0]  $rows[1]\n";
}
close(OUT);
#die ref1_id/ref2_id-Kozitationspaare werden dazugeholt
#sowie Nummerierung von ref1_id dazu
$sth=$dbh->prepare(qq{create table network1_$j select
auswertung4.ref1_id, ref2_id, nr as nr_1 from auswertung4, vertices2_$j
where auswertung4.ref1_id=vertices2_$j.ref_id and square_root > '$i'});
$sth->execute();
```

```
#Nummerierung von ref2_id dazu
$sth=$dbh->prepare(qq{create table network2_$j select nr_1,
network1_$j.ref1_id, network1_$j.ref2_id, nr as nr_2 from
network1_$j,
vertices2_$j where network1_$j.ref2_id=vertices2_$j.ref_id
order by nr_1, nr_2});
$sth->execute();
#die Kozitationspaare werden in der Maskierung
# in Textdatei ausgelesen
$sth=$dbh->prepare(qq{select nr_1, nr_2 from
network2_$j order by
nr_1, nr_2});
$sth->execute();
print OUT2 "*Edges\n";
while
(@rows=$sth->fetchrow_array()){
print OUT2 "$rows[0]  $rows[1]\n";
}
close(OUT2);
print "Durchlauf $j beendet!\n";
}
$dbh->disconnect();
exit(0);
```

D.5. Entropien der Referenzen

```
#!/usr/bin/perl
#liest Pajek-Export-Datei in DB ein, generiert Cluster-Tabellen und berechnet
#Entropien der Kozitationscluster.
#in der Schleife am Anfang muss der Bereich der Kozitationsschwellwerte, #die
berechnet werden sollen, spezifiziert werden
use DBI;
use strict;
my ($debug) = 1;
my ($dsn)= "DBI:mysql:SCI98test:localhost"; # DB-Name
my ($user_name) = "rot";
my ($pass_word) = "fh123!";
my ($dbh);
my ($sth);
my ($ref);
my ($anzahl0);
my ($anzahl);
my ($rv);
my ($i);
my ($j);
my ($k);
my ($nr);
my ($clusteranzahl);
```

```
my ($clustergroesse);
my ($sql);
my ($log);
my ($ergebnis);
my ($ergebnis1);
my ($ergebnis2);
my ($ergebnis3);
my ($gesamtmenge);
#Datenbank-Kommunikation
$dbh=DBI->connect($dsn, $user_name, $pass_word, {RaiseError=>0}) or die "Couldn`t
connect to database".DBI->errstr;
for ($i=0.1; $i < 0.9; $i = $i + 0.1) {
$j=$i;
$j=~ s/\.//;
#Loeschung temporaer genutzer Tabellen
if ($debug == 1) {
$sth=$dbh->prepare(qq{drop table allcluster_$j});
$sth->execute();
$sth=$dbh->prepare(qq{drop table allcluster2_$j});
$sth->execute();
$sql=qq{delete from cluster_shannon_sum_$j};
$sth=$dbh->prepare($sql);
$sth->execute();
 }
$sth=$dbh->prepare(qq{create table allcluster_$j (cluster_nr int(5)) type
= myisam});
$sth->execute();
# input-file:
my ($fn) = "./Daten/Pajek/network$j.clu";
my ($infile) = $fn;
my ($outfile) = $fn."-control.txt";
open(IN,$infile) || die "Cannot open $infile.";
open(OUT,">$outfile")|| die "Cannot make $outfile.";
#Einlesen der Pajek-Exportdateien
#die Pajek-Exportdatei besteht nur aus Clusterzugehoerigkeiten #(Nummern),
#die in Reihenfolge der eingelesenen Vertizes ausgegeben werden
while(<IN>){  #input
if($_ =~/\*/) {
next}
$nr = $_;
       $sth=$dbh->prepare(qq{insert into allcluster_$j (`cluster_nr`)
values ('$nr')});
$sth->execute();
}
#die Zugehoerigkeit der Cluster zu Referenzen wird
#durch Hinzufuegung eines Inkrement-Wertes
#wiederhergestellt -
$sth=$dbh->prepare(qq{alter table allcluster_$j add column nr int(5) not
null auto_increment, add primary key(nr)});
$sth->execute();
```

```
# - da durch die Nummerierung die Referenzen identifizierbar sind
$sth=$dbh->prepare(qq{create table allcluster2_$j select cluster_nr,
allcluster_$j.nr, ref_id from allcluster_$j, vertices2_$j where
allcluster_$j.nr=vertices2_$j.nr order by cluster_nr});
$sth->execute();
#Berechnung der Entropien
#Gesamtmenge Referenzen mit Zitationsanzahl groesser 1
     $sql=qq{select count(distinct ref_id) from threshold4 where nr_cit > '1'};
       $sth=$dbh->prepare($sql);
       $sth->execute();
          $rv=$sth->fetchrow_array();
       $gesamtmenge=$rv;
       print "Gesamtmenge refs: $gesamtmenge\n";
#Bestimmung der Clusteranzahl
$sth=$dbh->prepare(qq{select max(cluster_nr) from
`allcluster2_$j`});
$sth->execute();
$rv=$sth->fetchrow_array();
print "Anzahl Cluster: $rv\n";
$clusteranzahl=$rv;
#Bestimmung der Anzahl in Clustern gebundener
#Referenzen
$sth=$dbh->prepare(qq{select count(distinct ref_id) from
`allcluster2_$j`});
$sth->execute();
$rv=$sth->fetchrow_array();
print "Anzahl der in Clustern gebundenen refs: $rv\n";
#Schleife ueber alle Cluster
for ($k = 1; $k < $clusteranzahl+1; $k++) {
#Bestimmung der Clustergroessen
       $sql=qq{select count(*) from `allcluster2_$j` where cluster_nr =
'$k'};
       $sth=$dbh->prepare($sql);
       $sth->execute();
       $rv=$sth->fetchrow_array;
$clustergroesse=$rv;
#Berechnung der logarithmisierten Terme
       $log = log($clustergroesse);
       $ergebnis1 = $clustergroesse*$log;
#Einspeisung in temporaere Ergebnistabelle
$sql=qq{insert into cluster_shannon_sum_$j (`ergebnis`) values
('$ergebnis1')};
$sth=$dbh->prepare($sql);
$sth->execute();
}
#Addition der einzelnen Terme
$sql=qq{select sum(ergebnis) from cluster_shannon_sum_$j};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
```

```
$log = log($gesamtmenge);
print "log(n) = $log\n";
$ergebnis =$rv/$gesamtmenge;
$ergebnis3 =$log-$ergebnis;
print "Ergebnis Shannon-Index: $ergebnis3\n";
print "Durchgang $i beendet!\n";
close(IN);
close(OUT);
}
$dbh->disconnect();
exit(0);
```

D.6. Berechnung der Forschungsfronten

```
#! /usr/bin/perl
#berechnet Forschungsfronten der vorhandenen Cluster sowie der restlichen
#Referenzen, die bei der Kozitationsschwelle herausfallen. #Kozitationschwelle (ohne
Komma)  derjenigen Clusterverteilung, fuer die #research fronts berechnet werden
sollen, muss am
#Anfang spezifiziert werden
use DBI;
use strict;
my ($debug) = 1;
my ($dsn)="DBI:mysql:SCI98test:localhost";
my ($user_name)="rot";
my ($pass_word)="fh123!";
my ($dbh, $sth);
my ($rv);
my ($cluster_anzahl);
my ($sql);
my ($i) = 04;
my ($j);
my ($k);
my ($ref);
my ($rec);
my (@rows);
my (@recs);
#Datenbankkommuniaktion
$dbh=DBI->connect($dsn, $user_name, $pass_word) or die "Couldn't connect
to database".DBI->errstr;
#Loeschung temporaer genutzter Tabellen
if($debug==1){
        #!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
$sql =qq{drop table allresearchfronts};
$sth=$dbh->prepare($sql);
$sth->execute();
$sql=qq{drop table allresearchfronts2};
$sth=$dbh->prepare($sql);
```

```
$sth->execute();
$sql=qq{drop table allresearchfronts3};
$sth=$dbh->prepare($sql);
$sth->execute();
$sql=qq{drop table singles};
$sth=$dbh->prepare($sql);
$sth->execute();
}
        $sql=qq{create table allresearchfronts (rec_id int(10), rs_nr int
(10) default NULL) type=myisam};
$sth=$dbh->prepare($sql);
$sth->execute();
#Ermitteln aller research papers (die Tabelle ref_rec2 wurde aus der #ursprünglichen
Relationentabelle ref_rec über die Selektion der #Dokumententypen ‚article', note'
und letter' gebildet), die Referenzen
#aus Kozitationsclustern zitieren
$sql=qq{insert into allresearchfronts (rec_id, rs_nr) select
distinct ref_rec2.rec_id as
rec_id, allcluster2_$i.cluster_nr as rs_nr from ref_rec2,
allcluster2_$i
where ref_rec2.ref_id=allcluster2_$i.ref_id
order by rec_id};
$sth=$dbh->prepare($sql);
$sth->execute();
$sql=qq{select max(rs_nr) from allresearchfronts};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$k=$rv+1;
#Ermitteln der Referenzen, die beim jew. Schwellwert unverbunden #geblieben sind
(degenerated clusters)
$sql=qq{create table singles (ref_id int(10), rs_nr int(10) auto_increment, primary
key(rs_nr)) auto_increment = $k};
$sth=$dbh->prepare($sql);
$sth->execute();
$sql=qq{insert into singles (ref_id) select distinct threshold4.ref_id from
threshold4
left join vertices2_$i on vertices2_$i.ref_id=threshold4.ref_id where
vertices2_$i.ref_id is NULL and threshold4.nr_cit > '1'};
$sth=$dbh->prepare($sql);
$sth->execute();
#die zitierenden research papers dieser Referenzen werden mit fortlaufender #research
fronts-Nummerierung in die research fronts-Tabelle eingefuegt
$sql=qq{insert into allresearchfronts (rec_id, rs_nr) select ref_rec2.rec_id as
rec_id, rs_nr from ref_rec2, singles where ref_rec2.ref_id=singles.ref_id order by
rec_id};
$sth=$dbh->prepare($sql);
$sth->execute();
#Zaehlen der Ueberschneidungen: wie viele (degenerated) Cluster werden von ident.
Artikeln zitiert
```

```
$sql=qq{create table allresearchfronts2 select rec_id, count(rec_id) as count from
allresearchfronts group by rec_id};
$sth=$dbh->prepare($sql);
$sth->execute();
$sql=qq{create table allresearchfronts3 (rec_id int(10), rs_nr int(10),
proporz_wert numeric(35,30)) type=myisam};
$sth=$dbh->prepare($sql);
$sth->execute();
#diese Werte werden fraktional umgerechnet
$sql=qq{insert allresearchfronts3 select allresearchfronts.rec_id, rs_nr, 1/count as
proporz_wert from allresearchfronts, allresearchfronts2 where
allresearchfronts.rec_id=allresearchfronts2.rec_id};
$sth=$dbh->prepare($sql);
$sth->execute();
$dbh->disconnect();
exit(0);
```

D.7. Entropien der Artikel/Länder

```
#! /usr/bin/perl
#Entropie der Forschungsfronten; Berechnung der Laender und Laenderentropien
use DBI;
use strict;
my ($debug) = 1;
my ($dsn)="DBI:mysql:SCI98test:localhost";
my ($user_name)="rot";
my ($pass_word)="fh123!";
my ($dbh, $sth);
my ($rv);
my ($cluster_anzahl);
my ($sql);
my ($i);
my ($j);
my ($k);
my ($ref);
my ($rec);
my (@rows);
my (@recs);
my ($rv);
my ($rs_anzahl);
my ($rs_groesse);
my ($p);
my ($log);
my ($logn);
my ($maxentropie);
my ($ergebnis);
my ($ergebnis2);
```

```
my ($ergebnis3);
my ($gesamtmenge);
my ($gesamtmenge1);
my ($gesamtmenge2);
my ($gesamtmenge3);
my ($gesamtmenge4);
my ($pi);
my ($anzahl_rest_recs);
my (@array);
my (@array2);
my ($rec_nr);
my ($land);
my ($ref);
my ($log);
my ($logn);
my ($ergebnis);
my ($ergebnis2);
my (@ergebnis);
my (@decimal);
my ($anzahl_rest_recs);
#######Verbindung zur Datenbank###############
$dbh=DBI->connect($dsn, $user_name, $pass_word) or die "Couldn't connect
to database".DBI->errstr;
##########Berechnung der research fronts#################
##temporaere Tabellen loeschen
if($debug==1){
$sql=qq{delete from shannon_sum};
$sth=$dbh->prepare($sql);
$sth->execute();
}
##Berechnung der Entropie
#Gesamtmenge          Records          in          den          research          fronts

$sql=qq{select count(distinct rec_id) from allresearchfronts};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$gesamtmenge=$rv;
print "records in researchfronts: $gesamtmenge\n";
# Gesamtmenge alle research articles
$sql=qq{select  count(*)  from  records  where  (doc_type  =  'Article'  or  doc_type  =
'Letter' or doc_type = 'Note')};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$gesamtmenge2=$rv;
print "Gesamtmenge records: $gesamtmenge2\n";
$logn = log($gesamtmenge2);
print "log(n) = $logn\n ";
#Bestimmung der Anzahl der research fronts
```

```
$sql=qq{select count(distinct rs_nr) from allresearchfronts};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$rs_anzahl2=$rv;
print "rs_anzahl: $rs_anzahl2\n";
#Bestimmung der höchsten research front-Nummer zum Iterieren
#kann evt. von der absoluten Anzahl abweichen,
#da Kozitationscluster auf der Basis von allen Artikeln gebildet wurden,
#research fronts aber nur aus articles, notes und letters bestehen
$sql=qq{select max(rs_nr) from allresearchfronts};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$rs_anzahl=$rv;
#Berechnung der Entropie der research fronts
for ($j=1; $j < $rs_anzahl+1; $j++) {
       $sql=qq{select sum(proporz_wert) from allresearchfronts3 where rs_nr=$j};
       $sth=$dbh->prepare($sql);
       $sth->execute();
       $rv=$sth->fetchrow_array;
#falls research front wegfaellt, da zitierende Artikel reviews
#o. a. sind
       if ($rv){
       $rs_groesse=$rv;
       # p(i) berechnen
$pi=$rs_groesse/$gesamtmenge2;
# p(i)ln p(i) berechnen
$log = log($pi);
$ergebnis = $pi * $log;
        $sql=qq{insert into shannon_sum (`ergebnis`) values
('$ergebnis')};
        $sth=$dbh->prepare($sql);
        $sth->execute();
        }
}
#Berechnung der Entropie fuer die restlichen Records, die nicht teil der #research
fronts sind
$anzahl_rest_recs=$gesamtmenge2 - $gesamtmenge;
print "Anzahl restliche rec_ids: $anzahl_rest_recs\n";
for ($k=1; $k < $anzahl_rest_recs+1; $k++) {
$rs_groesse=1;
        $pi = 1/$gesamtmenge2;
$log=log($pi);
$ergebnis=$pi*$log;
      $sql=qq{insert into shannon_sum (`ergebnis`) values ($ergebnis)};
       $sth=$dbh->prepare($sql);
        $sth->execute();
 }
```

```
#Addition der einzelnen Logarithmus-Produkte
$sql=qq{select sum(ergebnis) from shannon_sum};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$ergebnis2 = -$rv;
print "Ergebnis Shannon-Index: $ergebnis2\n";
print OUT "Ergebnis Entropie research fronts: $ergebnis2 Anzahl research fronts:
$rs_anzahl  logn: logn";
close (OUT);
##########Berechnung der Laender-Anteile##############
##Bestimmung der Records und research front-Zugehoerigkeiten der Laender
if($debug==1) {
#Loeschung der temporaeren Tabellen
$sql=qq{drop table rs_laender};
$sth=$dbh->prepare($sql);
$sth->execute();
}
$sql=qq{create table rs_laender (rec_id int(10), item varchar(80),
rs_nr int(10), proporz_wert numeric(30,25)) type=myisam};
$sth=$dbh->prepare($sql);
$sth->execute();
#print "SQL=$sql";
@array = qw(USA JAPAN RUSSIA GERMANY FRANCE ENGLAND SCOTLAND NORTH-IRELAND WALES);
foreach $k(@array) {
$sql=qq{insert into rs_laender select distinct allresearchfronts3.rec_id,
item, rs_nr, proporz_wert
from countries, par_add, add_rec, allresearchfronts3
where par_add.cou_id=countries.cou_id and
par_add.add_id=add_rec.add_id and add_rec.rec_id=allresearchfronts3.rec_id
and item='$k' order by item};
$sth=$dbh->prepare($sql);
$sth->execute();
}
########da  urspranglich  England,  Nord-Irland,  Wales  und  Schottland  #getrennt
aufgefuehrt wurden
########werden diese durch GB korrigiert
$dbh->do(qq{update rs_laender set item='ENGLAND' where item
='SCOTLAND'});
$dbh->do(qq{update rs_laender set item='ENGLAND' where item ='WALES'});
$dbh->do(qq{update rs_laender set item='ENGLAND' where item
='NORTH-IRELAND'});
###########Berechnung der Entropie fuer Laender##########

@array=qw(USA JAPAN GERMANY FRANCE RUSSIA ENGLAND);
if($debug==1) {
foreach $k(@array){
$sth=$dbh->prepare(qq{drop table shannon_sum_$k});
$sth->execute();
```

```
}
$sth=$dbh->prepare(qq{delete from laender_entropie});
$sth->execute();
}
foreach $k(@array) {
$sql=qq{create table shannon_sum_$k (ergebnis numeric(65,60)) type
= myisam};
$sth=$dbh->prepare($sql);
$sth->execute();
#Gesamtmenge Records in den researchfronts je Land
$sql=qq{select count(distinct rec_id) from rs_laender where
item='$k'};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$gesamtmenge=$rv;
print "recs in researchfronts fuer $k: $gesamtmenge\n";
#Gesamtmenge Records je Land
$sql=qq{select rec_nr from laender_rec where item='$k'};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$gesamtmenge2=$rv;
print "Gesamtmenge alle recs fuer $k: $gesamtmenge2\n";
$logn = log($gesamtmenge);
print "log(n) = $logn";
#Gesamtzahl research fronts je land
$sql=qq{select count (distinct rs_nr) from rs_laender where
item='$k'};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$gesamtmenge3=$rv;
print "Gesamtmenge researchfronts fuer $k: $gesamtmenge3\n";
#Berechnung der einzelnen research fronts
#Iterierung über alle hypothetischen research fronts
for ($j=1; $j < $rs_anzahl+1; $j++) {
       $sql=qq{select sum(proporz_wert) from rs_laender where item =
'$k' and rs_nr='$j'};
       $sth=$dbh->prepare($sql);
       $sth->execute();
       $rv=$sth->fetchrow_array;
if ($rv){
       $rs_groesse=$rv;
#p(i) berechnen

$pi=$rs_groesse/$gesamtmenge2;
$log=log($pi);
        $ergebnis=$pi*$log;
       $sql=qq{insert into shannon_sum_$k (`ergebnis`) values
```

```
('$ergebnis')};
        $sth=$dbh->prepare($sql);
        $sth->execute();
}
}
        #Berechnung fuer einzelne ungebundene Referenzen
$anzahl_rest_recs=$gesamtmenge2 - $gesamtmenge;
   print "Anzahl rest-rec_ids: $anzahl_rest_recs\n";
        for ($j=1; $j < $anzahl_rest_recs+1; $j++) {
       $rs_groesse=1;
$pi = 1/$gesamtmenge2;
        $log=log($pi);
$ergebnis=$pi*$log;
$sql=qq{insert into shannon_sum_$k (`ergebnis`) values ($ergebnis)};
        $sth=$dbh->prepare($sql);
        $sth->execute();
#Addition der logarithmierten Terme
$sql=qq{select sum(ergebnis) from shannon_sum_$k};
$sth=$dbh->prepare($sql);
$sth->execute();
$rv=$sth->fetchrow_array();
$ergebnis2 = -$rv;
        print "Ergebnis Shannon-Index fuer $k: $ergebnis2\n\n";
###########Ergebnistabelle
$sql=qq{insert into laender_entropie (land, entropie) values ('$k',
'$ergebnis2')};
$sth=$dbh->prepare($sql);
$sth->execute();
}
$dbh->disconnect();
exit(0);
```

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen verwendet habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.