Michael Heinz, Oliver Mitesser,
Jochen Gläser & Frank Havemann

Ist die Vielfalt der Forschung in Gefahr? Methodische Ansätze für die bibliometrische Messung thematischer Diversität von Fachbibliographien

1. Einführung

Die Frage der Forschungsvielfalt scheint ins Zentrum der Debatte zur Forschungspolitik zu rücken. Neuere Instrumente der Governance von Wissenschaft binden die Grundfinanzierung und damit die Existenz von Forschungseinheiten an deren Erfolg in der Leistungsbewertung. Diese Versuche, Forschungsfinanzierung selektiver zu gestalten, reduzieren die Zahl der geförderten Einheiten, wodurch tendenziell auch die Vielfalt der Forschungsansätze verringert wird.[1] Subtiler ist die Gefahr, die der Forschungsdiversität durch angepasstes Verhalten von Wissenschaftlern erwächst. Wann immer Forschungspolitik wissenschaftliche Fehlschläge unmittelbar bestraft, zum Beispiel durch Mittelreduzierung, werden Forscher gedrängt, sich sichere Projekte vorzunehmen, d. h. solche, die von den Fachgemeinschaften gebilligt werden und hohe Erfolgsaussichten haben. Sichere Projekte folgen dem Mainstream des Fachgebietes und nutzen erprobte Methoden. Forschung abseits des Mainstreams wird so immer seltener, was die Vielfalt der Problemstellungen und Forschungsstrategien in einem Feld beeinträchtigt.[2]

1 Adams, J. / Smith, D., Funding research diversity. – In: A report from Evidence Ltd to Universities UK 1(2003)84036, S. 102.

2 Harley, S. / Lee, F. S., Research selectivity, managerialism, and the academic labor process: The future of nonmainstream economics in UK universities. – In: Human Relations 50(1997), S. 1427 – 1460.
Whitley, R., Evaluation without Evaluators: The Consequences of Establishing Research Evaluation Systems for Knowledge Production in Different Countries and Scientific Fields. – In: The Changing Governance of the Sciences: The Advent of Research Evaluation Systems. Ed. by R. Whitley und J. Gläser, Dordrecht: Springer 2007. S. 3 – 27.

Diesen plausiblen Argumenten für die so genannte Homogenisierungshypothese (homogenisation hypothesis) fehlt bisher jedoch die empirische Bestätigung. Während auf der Mikroebene die Mechanismen identifiziert werden konnten, welche Forscher zum Maistream streben lassen,[3] ist es bisher nicht überzeugend gelungen, Forschungsvielfalt auf höheren Aggregationsebenen zu messen. Wissenschaftspolitik und Wissenschaftsforschung sind deshalb bis heute auf die Meinungen von Wissenschaftlern angewiesen, die jedoch keine verlässliche Evidenz liefern. Nonkonformistische Ansätze können von der Mehrheit einer Fachgemeinschaft als schlechte Wissenschaft wahrgenommen werden. Umgekehrt können Forscher die geringe Anerkennung ihrer Ergebnisse auf deren Spezifizität zurückführen, um ihre mangelnde Qualität (vor sich selbst und anderen) zu verschleiern. Um die Homogenisierungsshypothese zu testen, müssen Verfahren verwendet werden, die unabhängig davon sind, wie Wissenschaftler das Problem wahrnehmen.[4]

Bibliometrische Verfahren bieten sich für die Konstruktion objektiver Maße von Forschungsvielfalt an, weil sie Indikatoren für Forschungsinhalte nutzen, die unabhängig von den Ansichten der Wissenschaftler über diese Forschungsinhalte sind. Das Diversitätskonzept wurde bisher in der Wissenschaftsforschung selten verwendet und noch nicht befriedigend operationalisiert. Nach unserem Wissen hat zuerst Hariolf Grupp (1990) einen bibliometrischen Ansatz zu Messung der Forschungsvielfalt vorgeschlagen.[5] Er ermittelte unter anderen die Diversität der Innovationsstrategien einiger Länder und Firmen anhand von Patentzahlen nach Patentklassen.

Mit dem Ziel, Interdisziplinarität bibliometrisch messbar zu machen, wurde sie von Bordons et al. (2004) sowie von Rafols & Meyer (2007) auf thematische Diversität zurückgeführt.[6] Eine generelle Diskussion der Anwendung von Diversitätsmaßen in der Wissenschaftsforschung hat Stirling 2007 vorgelegt.[7] Kürzlich

3 Gläser, J. / Laudel, G., Evaluation without Evaluators: The impact of funding formulae on Australian University Research. – In: The Changing Governance of the Sciences: The Advent of Research Evaluation Systems. Ed. by R. Whitley u. J. Gläser. Dordrecht: Springer 2007. S. 127 – 151.
Gläser, J. / Lange, S. / Laudel, G. / Schimank, U., Evaluationsbasierte Forschungsfinanzierung und ihre Folgen. – In: Wissen für Entscheidungsprozesse. Hrsg. v. F. Neidhardt, R. Mayntz, P. Weingart u. U. Wengenroth. Bielefeld: transcript 2008. S. 145 – 170.

4 vgl. den Vortrag von Uwe Schimank zur „*Typologie institutioneller Handlungsbedingungen von Wissenschaftlern – Institutionelle Governance-Strukturen und ihre Auswirkungen auf die Forschung an Hochschulen*" auf der Tagung der Gesellschaft für Wissenschaftsforschung am 24. und 25. März 2006 in Berlin zum Thema „*Wissenschaft und Technik in theoretischer Reflexion*".

5 Grupp, H., The concept of entropy in scientometrics and innovation research. – In: *Scientometrics* 18(1990), 3 – 4, S. 219 – 239.

haben Rafols & Meyer uns einen noch unveröffentlichten Aufsatz zum Thema zugänglich gemacht.[8]

2. Diversitätsmaße

Sind in einem Waldstück *W* mehr Baumarten vertreten als in einem Vergleichsgebiet *W**, so neigt man dazu, den Baumbstand in *W* als vielfältiger zu bezeichnen. Wenn aber der Bestand im Gebiet *W* von Kiefern dominiert wird und alle anderen Baumarten nur mit vereinzelten Exemplaren vertreten sind, kann *W** trotz geringerer Artenzahl einen vielfältigeren Eindruck machen als *W*. Es kommt also nicht nur auf die Artenzahl an, sondern auch darauf, wie gleichmäßig die Individuen auf die Arten verteilt sind, was als evenness oder auch als die Balance der Verteilung bezeichnet werden kann.[9] Ein Diversitätsmaß, das Artenzahl und Balance berücksichtigt, ist der mittlere Informationsgehalt der Aussage, ein Individuum im Biotop gehöre einer bestimmten Art an. Sie ergibt sich aus den relativen Häufigkeiten der Arten im Biotop nach der bekannten Boltzmannschen Formel für die Entropie und wird auch als Shannon-Index bezeichnet.[10] Ein weiteres Maß für die Diversität, welches Balance und Artenzahl berücksichtigt, ist die Wahrscheinlichkeit, dass zwei zufällig ausgewählte Individuen verschiedenen Arten angehören. Dieses Maß geht auf Simpson (1949) zurück und wird auch als Simpson-Index bezeichnet.[11]

6 Bordons, M. / Morillo, F. / Gomez, I., Analysis of cross-disciplinary research through bibliometric tools. – In: Handbook of quantitative science and technology research. Ed. by H. F. Moed, W. Glänzel, u. U. Schmoch, Chapter 21, S. 437–456. Dordrecht: Kluwer 2004; Rafols, I. / Meyer, M., Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. – In: Proceedings of ISSI 2007. Volume 2. Ed by D. Torres-Salinas and H. F. Moed. Madrid 2007. S. 631 – 637.

7 Stirling, A., A general framework for analysing diversity in science, technology and society. – In: Journal of The Royal Society Interface. 4 (2007), 15, S. 707 – 719.

8 Rafols, I. / Meyer, M., Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. Preprint 2008.;http://www.sussex.ac.uk/spru/documents/rafols-meyer-diversity2008.pdf.

9 Vgl. Stirling, A., 2007, a.a.O.

10 Mathematische Einzelheiten können unserem online frei verfügbaren Konferenzbeitrag vom Juli 2008 entnommen werden: Mitesser, O. / Heinz, M. / Havemann, F. / Gläser, J., Measuring Diversity of Research by Extracting Latent Themes from Bipartite Networks of Papers and References. – In: Proceedings of WIS 2008: Fourth International Conference on Webometrics, Informetrics and Scientometrics Ninth COLLNET Meeting, Berlin. Ed. by H. Kretschmer and F. Havemann. Berlin: Gesellschaft für Wissenschaftsforschung 2008. http://www.collnet.de/Berlin-2008/MitesserWIS2008mdr.pdf.

11 Simpson, E., Measurement of diversity. – In: Nature. 163(1949)4148, S. 688.

Ein weiterer Aspekt von Vielfalt betrifft die Unterschiedlichkeit der Arten (disparity). Wenn im Waldstück *W* nur Nadelbäume wachsen, *W** aber ein Mischwald ist, dann wird *W** auch bei gleicher Artenzahl und Balance als vielfältiger empfunden werden. Wir gehen damit von der binären Relation mit den Werten gleich und ungleich über zu einer graduellen Unterschiedlichkeit.[12] Der Rao-Index der Diversität misst entsprechend die mittlere Disparität eines zufälligen Paares.[13]

Die Disparität *D* mit Werten zwischen 0 und 1 kann durch eine Distanz *d* ersetzt werden, für die auch Werte $d > 1$ auftreten. Dann misst man Diversität durch eine mittlere (taxonomisch oder genetisch definierte) Entfernung zwischen den Individuen des Biotops. Wenn nur eine Art betrachtet wird, kann ihre genetische Diversität durch die mittlere genetische Entfernung zwischen ihren Individuen gemessen werden, welche i. a. alle genetisch verschieden sind.

3. Kozitationsanalyse

Eine oft erprobte Methode, die thematische Struktur wissenschaftlicher Zeitschriftenliteratur sichtbar zu machen, ist die auf Irina Marshakova (1973) und auf Henry Small (1973) zurückgehende Kozitationsanalyse.[14] In einem Zeitschriften-Jahrgang oft zitierte Quellen dienen dabei als Symbole für Standardkonzepte. Werden sie auch oft kozitiert, zeigt das ihre thematische Nähe an. Mit dem Salton-Index der Kozitierung als Ähnlickeitsmaß wurden dann von Small und Sweeney (1985) mittels *single-linkage clustering* Kozitationscluster hochzitierter Referenzen gebildet.[15] Dabei können die minimalen Werte von Zitierung und Kozitierung (*thresholds*) variiert werden. Diese Cluster werden dann auf den Jahrgang der zitierenden Aufsätze reprojiziert, indem die jeweils ein Cluster zitierenden Arbeiten als eine Forschungsfront angesehen werden..

12 Shimatani, K., On the measurement of species diversity incorporating species differences. – In: Oikos. 93(2001)1, S. 135 – 147.

13 Ricotta, C. / Szeidl, L., Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. – In: Theoretical Population Biology. 70(2006)3, S. 237 – 243.

14 Marshakova, I., System of document connections based on references. – In: Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy i Sistemy. 6(1973), S. 3 – 8. (in Russisch); Small, H., Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. – In: Journal of the American Society for Information Science. 24(1973), S. 265 – 2 69.

15 Small, H. / Sweeney, E., Clustering the Science Citation index using cocitations. – In: Scientometrics. 7(1985)3, S. 391 – 409.

Ein Versuch, die Größen von Kozitationsclustern und von Forschungsfronten in einem Fachgebiet als Ausgangsgrößen für das Entropiemaß der Forschungsvielfalt zu verwenden, zeigte aber sofort, dass diese dafür wenig geeignet sind.[16] Da es bei der Messung der Forschungsvielfalt nicht nur um deutlich sichtbare Frontgebiete der Forschung gehen kann, sondern auch um die vielen kleinen, wenig sichtbaren Themen – die gerade die Vielfalt ausmachen –, war es notwendig den Zitationsschwellwert auf ein Minimum herabzusetzen. Dadurch kam aber eine negative Eigenschaft des *single-linkage clustering* zur Wirkung: das *chaining*. Darunter versteht man, dass langgezogene Kozitationscluster entstehen, deren Enden thematisch wenig miteinander zu tun haben. Ab einem bestimmten Schwellwert für den Salton-Index der Kozitation werden dadurch fast alle zitierten Quellen in einem großen Cluster versammelt.

Dieses negative Ergebnis machte deutlich, dass offenbar die Klassifizierung der Literatur eines Fachgebiets nach disjunkten Themen – als Analogon zu den disjunkten Arten in einem Biotop – auf Schwierigkeiten stößt. Auch wenn durch eine andere Clustermethode der *chaining*-Effekt möglicherweise vermeidbar ist – es bleibt die Schwierigkeit, dass eine Arbeit genau einem thematischen Cluster zugordnet werden muss. Das ist eine nicht zu rechtfertigende Einschränkung.[17]

4. Bibliographische Kopplung

Weil es schwierig ist, jede Zeitschriftenpublikation genau einem thematischen Cluster zuzuordnen, liegt es nahe, vom Analogon der biologischen Artenvielfalt zu dem der genetischen Vielfalt innerhalb einer Art überzugehen.[18] Bei Populationen einer Art können keine scharf definierten Teilmengen gebildet werden. Die Diversität wird hier genetisch gemessen. Durch die Bestimmung der genetischen Ähnlichkeit kann man ein Abstandsmaß gewinnen und benutzt dann den mittleren Abstand als Maß für die genetische Diversität der Population. Die genetische Information eines Individuums verweist auf dessen Vorfahren. In der wissenschaftlichen Literatur sind einige der unmittelbaren geistigen Vorfahren eines Werkes in der Liste der zitierten Quellen aufgeführt. Diese bibliographische In-

16 Schmidt, M. / Gläser, J. / Havemann, F. / Heinz, M., A Methodological Study for Measuring the Diversity of Science. – In: International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, 10–12 May, Nancy. S. 129 – 137. SRDI – INIST-CNRS 2006.

17 Gläser, J., Wissenschaftliche Produktionsgemeinschaften: Die Soziale Ordnung der Forschung. Frankfurt am Main: Campus 2006. Vgl. insbesondere S. 137–139 und 160–162.

18 Havemann, F. / Heinz, M. / Schmidt, M. / Gläser, J., Measuring Diversity of Research in Bibliographic-Coupling Networks. – In: Proceedings of ISSI 2007. Volume 2. Ed. by D. Torres-Salinas and H. F. Moed. Madrid 2007. S. 860 – 861. Poster abstract.

formation ist allerdings weitaus unvollständiger als die genetische: keinesfalls ist aus ihr der gesamte geistige Stammbaum ablesbar.[19] Dazu müsste man rekursiv im Zitationsgraphen alle Vorfahren ermitteln, d. h. auch die Quellen der Quellen in die Analyse einbeziehen bis zu den geistigen Stammmüttern oder -vätern.

Vor einem solchen Unterfangen schreckten wir wegen des Aufwandes zurück und versuchten, allein mit den unmittelbaren Vorfahren ein Abstandsmaß zwischen Artikeln zu konstruieren, welches sich zur Messung thematischer Vielfalt eignet. Dies bedeutet, Artikel als fachlich nah anzusehen, wenn sie viele zitierte Quellen gemeinsam haben oder, mit anderen Worten, wenn sie stark bibliographisch gekoppelt sind. Das zur Kozitationsmethode komplementäre Konzept der bibliographischen Kopplung wurde von Kessler 1963 eingeführt.[20] Die Zahl gleicher zitierter Quellen kann für die Berechnung eines relativen Maßes der Kopplungsstärke, wie dem Salton- oder dem Jaccard-Index, benutzt werden.

Das Netzwerk bibliographisch gekoppelter Artikel eines Jahrgangs in einem Fachgebiet weist jedoch wegen der Unvollständigkeit der bibliographischen Information eine sehr geringe Netzwerkdichte auf: nur wenige der $n(n-1)/2$ möglichen Kopplungen zwischen n Artikeln sind realisiert (< 1 Prozent). Eine auf dieser Basis berechnete mittlere Entfernung kann deshalb kein sinnvoller Indikator für die Forschungsdiversität sein.

Andererseits sind fast alle Artikel eines Jahrgangs in der Hauptkomponente des Netzwerks versammelt (> 90 Prozent), d. h. sie hängen wenigstens indirekt zusammen. Diesen Befund kann man sich zunutze machen und die Länge des kürzesten Pfades zwischen zwei Artikeln in der Hauptkomponente als Entfernung zwischen ihnen definieren. Dieses Vorgehen ist ganz ähnlich dem von Botafogo *et al.* (1992) und von Egghe & Rousseau (2003), die ein Maß für die Kompaktheit von Netzwerken aus mittleren Längen kürzester Pfade ableiten.[21]

Wir haben für die elf Jahrgänge 1995–2005 von 13 elektrochemischen Zeitschriften[22] jeweils die Länge aller kürzesten Pfade in der Hauptkomponente berechnet. Die Entfernung zwischen zwei direkt bibliographisch gekoppelten Artikeln i und j haben wir als $d_{ij} = -\log(J_{ij})$ angesetzt, wo $J_{ij} < 1$ den Jaccard-Index der bibliographischen Kopplung bezeichnet, welcher als Verhältnis der Länge

19 Referenzen auf Quellen, die auch nicht unbedingt geistige Vorfahren sein müssen, werden oft mit einer gewissen Beliebigkeit hinzugefügt oder weggelassen. Dennoch ist die Methode der bibliographischen Kopplung, wie auch die Kozitationsmethode, geeignet, Beziehungen zwischen Artikeln zu ermitteln, wenn es um statistische Aussagen und um große Zahlen geht, um große Zahlen von Artikeln oder um große Zahlen von Zitationen eines Artikels.

20 Kessler, M. M., Bibliographic coupling between scientific papers. – In: American Documentation. 14(1963), S. 10 – 25.

von Durchschnitt und von Vereinigung der Referenzenlisten der beiden Artikel definiert ist.

Die mittlere Entfernung schwankt für die Doppeljahrgänge 1995/1996 bis 2000/2001 um den Wert 12,6, um danach mit einem deutlichen Trend bis 2004/2005 auf 11,9 abzusinken.[23] Wir fragten dann, ob diese Tendenz nicht auf andere Ursachen, als auf eine sinkende Forschungsvielfalt, zurückgeführt werden könnte. Tatsächlich bewegt sich entgegengesetzt zur mittleren Entfernung in der Hauptkomponente das (geometrische) Mittel der Länge der Referenzenlisten der Artikel ab 2000/2001 von 18,6 nach oben auf 21,7.[24]

Mehr Referenzen pro Artikel führen zu mehr Links im Netzwerk. Damit verkürzen sich viele kürzeste Pfade zwischen Artikeln, weil sie jetzt über Abkürzungen (short cuts) laufen können.

Um zu prüfen, ob die Abnahme der mittleren Distanz völlig durch die Zunahme der Kantenzahl erklärt werden kann, konstruierten wir aus unseren empirischen Netzwerken Modellgraphen, indem wir auf zufällige Weise zitierte Quellen in den Referenzenlisten löschten, bis die mittlere Referenzenzahl in allen Doppeljahrgängen gleich war. Um eine Zeitreihe vergleichbarer mittlerer Entfernungen zu erhalten, haben wir weiterhin aus allen Jahrgängen jeweils gleich große Zufallsstichproben von Artikeln gezogen, denn die Artikelzahl nimmt zum Ende der untersuchten Zeitspanne ebenfalls rapide zu, was zu größeren mittleren Entfernungen führen kann, aber auch zu kleineren (wenn dadurch mehr Verbindungen im Netzwerk entstehen). Um zu sehen, wie stark die Ergebnisse jeweils von der zufälligen Stichprobe abhängen, haben wir jedem Doppeljahrgang fünf Stichproben entnommen.

Tatsächlich verschwand durch diese Prozedur die fallende Tendenz für die mittlere Entfernung vollkommen. Beim Messen von Diversität sind zufällige Stichproben von Individuen zulässig, aber zufälliges Streichen von Referenzen

21 Botafogo, R. / Rivlin, E. / Shneiderman, B., Structural analysis of hypertexts: identifying hierarchies and useful metrics. – In: ACM Transactions on Information Systems (TOIS) 10(1992) 2, S. 142–180; Egghe, L. / Rousseau, R., BRS compactness in networks: Theoretical considerations related to cohesion in citation graphs, collaboration networks and the internet. – In: Mathematical and Computer Modelling. 37(2003)7–8, S. 879 – 899; Rafols, I. / Meyer, M., Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. – In: Proceedings of ISSI 2007.Volume 2. Ed. by D. Torres-Salinas and H. F. Moed. Madrid 2007. S. 631–637.

22 Wir benutzten den gleiche Satz von Zeitschriften wie Schmidt *et al.* (2006, a.a.O.) und verwendeten alle *records* vom Dokumenttyp *Article* und *Letter* aus dem *Web of Science* (WoS).

23 vgl. Fußnote 10 und http://www.collnet.de/Berlin-2008/MitesserWIS2008mdr.pdf

24 Weil die Verteilung der Länge der Referenzenlisten schief ist, benutzen wir das geometrische und nicht das arithmetische Mittel als Maßzahl der zentralen Tendenz (vgl. vorige Fußnote).

macht die Stichproben zu konstruierten Modellen, von denen nicht sicher auf die empirischen Gegebenheiten rückgeschlossen werden kann.

Der Ansatz, Forschungsvielfalt als mittlere kürzeste Distanz in einem Netzwerk bibliographisch gekoppelter Zeitschriftenaufsätze zu bestimmen, scheitert also daran, dass dieser Indikator zu sensibel auf Änderungen im Zitationsverhalten reagiert, welche nichts mit Änderungen der Diversität zu tun haben.

5. Extraktion latenter Themen

Artikel eines Jahrgangs können zusammen mit den in ihren Referenzenlisten auftretenden Quellen als ein bipartites Netzwerk aufgefasst werden, in dem nur Links von Artikeln zu den von ihnen zitierten Quellen vorhanden sind. Wir ignorieren dabei den Umstand, dass Artikel auch schon in ihrem Publikationsjahr zitiert werden können und damit auch zu den Quellen gehören.

Bei der Kozitationsanalyse werden Quellen nach ihrem Auftreten in den Referenzenlisten der Artikel zusammengefasst, bibliographische Kopplung von Artikeln wird über ihre gemeinsam zitierten Quellen vermittelt. In diesem Sinne sind beide Methoden komplementär zueinander. Eine Methode zur Bestimmung zusammenhängender Knoten, die beide Ebenen eines bipartiten Netzwerks symmetrisch behandelt und damit in unserem Fall Kozitation und bibliographische Kopplung gleichermaßen einschließt, beruht auf der Singulärwertzerlegung (SVD = *singular value decomposition*) der das Netzwerk beschreibenden Rechteckmatrix. Diese Methode heißt *latent semantic analysis* (LSA), wenn nicht zitierte Quellen, sondern Wörter aus den Artikeltexten als die zweite Sorte von Knoten gewählt werden.[25] Mittels SVD extrahiert man latente Themen, wobei Artikel wie Quellen (bzw. Wörter) nicht nur einem Thema zugeordnet werden. Das entspricht den Verhältnissen in der Literatur zu einem Fachgebiet – wie oben schon festgestellt – weitaus besser als ein hartes Clustern.[26]

Wir beschreiben das bipartite Netzwerk von n Artikeln eines Jahrgangs in einem Fachgebiet und den m in ihnen zitierten Quellen durch eine Rechteckmatrix X mit m Zeilen und n Spalten. Empirisch ist fast immer $m > n$. Element x_{ij} von X ist gleich 1, wenn im Artikel j die Quelle i zitiert wird und sonst 0. Es sei $r \leq n < m$ gleich dem Rang der Matrix X. Die Singulärwertzerlegung von X ist gegeben

25 Deerwester, S. / Dumais, S. / Furnas, G. / Landauer, T. / Harshman, R., Indexing by latent semantic analysis. – In: Journal of the American Society for Information Science. 41(1990)6, S. 391 – 407.

26 Janssens, F. / Glänzel, W. / De Moor, B., A Hybrid Mapping of Information Science. – In: Proceedings of ISSI 2007. Volume 1. Ed. by D. Torres-Salinas and H. F. Moed. Madrid 2007, S. 408 – 420.

durch $X = U\Lambda^{1/2}V^{T}$. Die r Spalten von U sind die normierten Eigenvektoren der Matrix XX^{T} zu von Null verschiedenen Eigenwerten. Matrix XX^{T} enthält die Kozitationsbeziehungen der m Quellen. Die r Spalten von V sind die normierten Eigenvektoren der Matrix $X^{T}X$ zu von Null verschiedenen Eigenwerten. Matrix $X^{T}X$ enthält die bibliographischen Kopplungen der n Artikel. Die Diagonalmatrix $\Lambda^{1/2}$ enthält die Wurzeln der r Eigenwerte $\lambda_k > 0$, die beiden Matrizen, XX^{T} und $X^{T}X$, gemeinsam sind (wie man leicht zeigen kann).

Man nimmt nun an, dass aus dem Netzwerk r latente Themen extrahierbar sind. Dazu werden in einem linearen Ansatz die n Spaltenvektoren von X nach der r-dimensionalen Orthonormalbasis U entwickelt. Der Beitrag von Thema k zum gesamten Jahrgang ist dann gleich dem Eigenwert λ_k.[27] Die Summe aller Eigenwerte ist gleich dem Quadrat der Frobenius-Norm $|X|_F$ von Matrix X und damit gleich der Zahl der Links im Netzwerk. Die Diversität eines Jahrgangs kann dann aus den relativen Anteilen $p_k = \lambda_k/|X|_F^2$ berechnet werden. Die analoge Rechnung für die Anteile der Themen an den zitierten Quellen führt zum gleichen Ergebnis.[28]

Bei SVD-gestützten Methoden – wie z. B. LSA – wird die Zahl der Dimensionen des Vektorraums künstlich verringert, indem die zu sehr kleinen Eigenwerten gehörenden Eigenvektoren weggelassen werden. Solcherart Dimensionsreduzierung macht die extrahierten Themen für praktische Zwecke übersichtlicher. Wir können auch hier (wie bei der Kozitionsanalyse, s. o.) die kleinen Themen nicht vernachlässigen, wenn wir Vielfalt messen wollen.

6. Erste Ergebnisse

Wir haben für zwei Fachgebiete, für die Elektrochemie – mit dem oben erwähnten Zeitschriftensatz – und für einen Teil der Informationswissenschaft, die SVD-gestützte Extraktion latenter Themen getestet. Die bibliographischen Angaben (inklusive der zitierten Quellen) für 21 Jahrgänge (1986–2006) der folgenden fünf informationswissenschaftlichen Zeitschriften mit hohem Anteil biblio-

27 s. Gl. 2 in: Alter, O. / Brown, P. / Botstein, D., Singular value decomposition for genomewide expression data processing and modeling. – In: Proceedings of the National Academy of Sciences. 97(2000)18, S. 10101 – 10106.

28 Weitere mathematische Einzelheiten der Anwendung der LSA-Methode sind in unserem Beitrag zur COLLNET-Konferenz in Berlin 2008 nachlesbar nachlesbar (s. Fußnote 10), wie auch in der Master-Arbeit von Oliver Mitesser: Mitesser, O., Latente semantische Analyse zur Messung der Diversität von Forschungsgebieten – Methodendiskussion und Anwendungsbeispiel. Master-Arbeit (2008), Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft. http://www.ib.hu-berlin.de/~kumlau/handreichungen/h240

metrischer Aufsätze haben wir aus dem Web of Science heruntergeladen und analysiert:

- *Information Processing & Management,*
- *Journal of the American Society for Information Science (and Technology),*
- *Journal of Documentation,*
- *Journal of Information Science,*
- *Scientometrics.*

Wir können auf die Angabe von Details des Datensatzes und der Berechnung weitgehend verzichten, weil sie in der Masterarbeit von Oliver Mitesser (2008) nachlesbar sind.[29]

Wir haben aus den Jahrgängen jeweils mehrere Stichproben gleicher Größe zufällig ausgewählt und für jeden Jahrgang den Mittelwert und die Standardabweichung (bezüglich der Stichprobenwahl) der Entropie berechnet. Es ergab sich eine deutliche Tendenz zu höheren Entropiewerten (siehe Abbildungen 1 und 3, linke Diagramme). Für den Simpson-Index ergeben sich ganz analoge Bilder. Für die Informationswissenschaft erhöht er sich von 98,35 Prozent auf 98,65Prozent, für die Elektrochemie von 99,71 Prozent auf 99,74 Prozent. Obwohl dies – wie bei der Entropie – nur kleine Differenzen sind, ist doch für beiden Maße ein eindeutiger und bis in die Einzelschritte übereinstimmender Trend unverkennbar.

Als nächstes testeten wir, ob die SVD-Extraktion latenter Themen genauso sensibel gegenüber der Tendenz zu längeren Referenzenlisten ist, wie unsere oben beschriebene Methode. Für diesen Zweck konstruierten wir ein Modell-Netzwerk für jeden Jahrgang beider Zeitschriftensätze, indem wir zufällig Zitationslinks zwischen Artikeln und Quellen beseitigten, bis wir eine mittlere Zahl von 15 Referenzen pro Artikel erreichten.

Diese Reduktion des Netzwerkes verändert die Entropiewerte für jeden Jahrgang, aber die Tendenz zu höheren Werten wird dadurch nicht beeinträchtigt, wie die Abbildungen 2 und 4 zeigen.[30] Sogar die Schritte von einem Jahrgang zu nächsten ändern sich nicht sehr, lediglich die Standardabweichungen vergrößern sich.

29 s. vorige Fußnote

30 Die trivialen Diagramme auf den rechten Seiten werden angezeigt, um die Reduktionsprozedur zu überwachen.

Abbildung 1. *Zeitreihe 1986–2006 der mittleren Entropie und der mittleren Referenzenzahl pro Artikel in fünf informationswissenschaftlichen Journalen. Entropie-Mittelwerte (Maximum* $\log_2 100 = 6{,}64$*) und Standardabweichung (als Fehlerbalken) sind für jeweils 50 Stichproben von 100 zufällig ausgewählten Artikeln berechnet.*

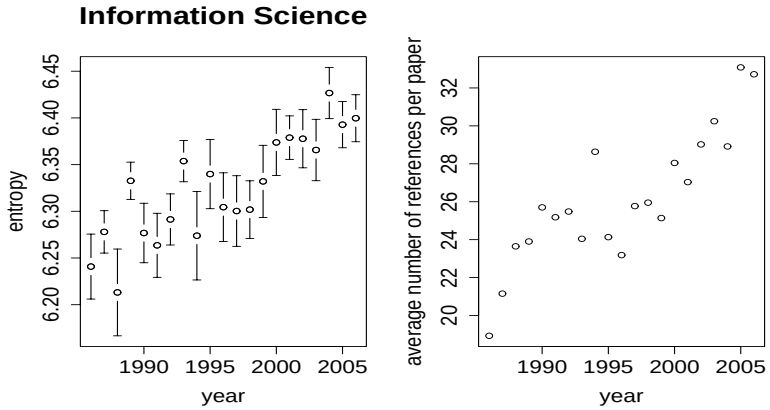


Abbildung 2. *Zeitreihe 1986–2006 mittlerer Entropie in einem Modell konstruiert aus fünf informationswissenschaftlichen Journalen, wobei die mittlere Referenzenzahl pro Artikel durch zufälliges Streichen auf 15 reduziert wurde (siehe a. Abb. 1).*

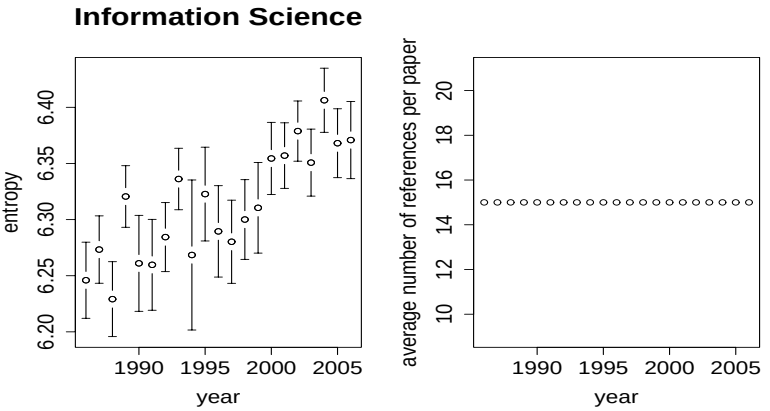


Abbildung 3. *Zeitreihe 1986–2006 der mittleren Entropie und der mittleren Referenzenzahl pro Artikel in 13 elektrochemischen Journalen. Entropie-Mittelwerte (Maximum* $\log_2 500 = 8.97$*) und Standardabweichung (als Fehlerbalken) sind für jeweils 50 Stichproben von 500 zufällig ausgewählten Artikeln berechnet.*

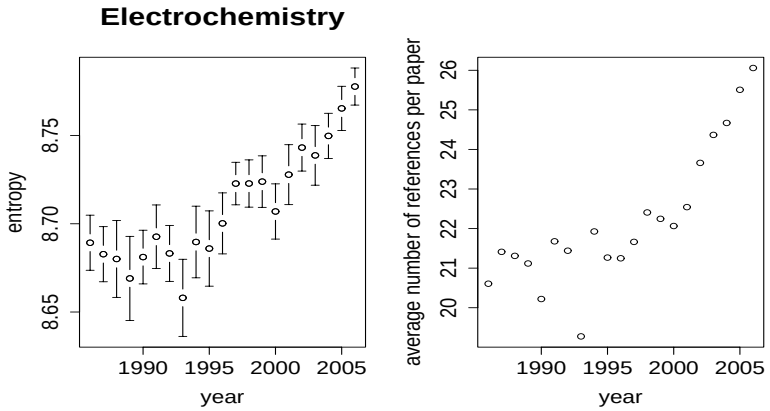
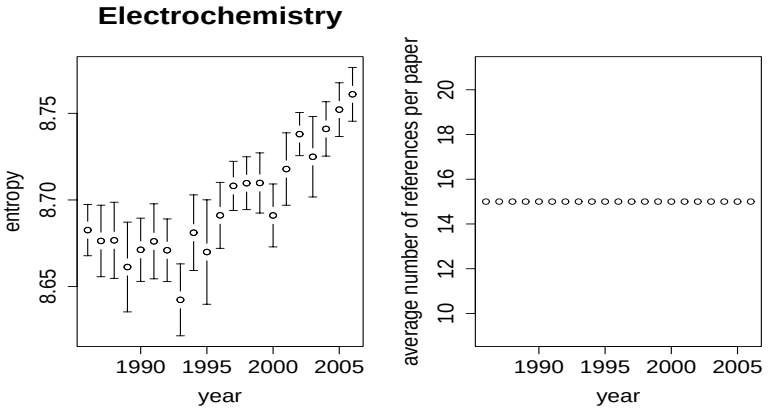


Abbildung 4. *Zeitreihe 1986–2006 mittlerer Entropie in einem Modell konstruiert aus 13 elektrochemischen Journalen, wobei die mittlere Referenzenzahl pro Artikel durch zufälliges Streichen auf 15 reduziert wurde (siehe auch Abb. 3).*



5. Diskussion

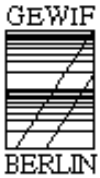
Wir können bislang nicht erklären, warum in beiden Forschungsfeldern zwischen 1986 und 2006 die Entropie latenter Themen ihr jeweiliges theoretisches Maximum anstrebt. In der Informationswissenschaft starten wir bei 94 Prozent des Maximalwertes und enden mit ungefähr 96 Prozent. In der Elektrochemie haben wir im gleichen Zeitraum einen Anstieg von 97 Prozent auf 98 Prozent. Offenbar kann sich dieser Trend nicht so fortsetzen und muß sich in den folgenden Jahren zumindest verlangsamen.

Wenn der Anstieg der Entropiewerte kein Artefakt ist, d. h. tatsächlich eine steigende Forschnungsvielfalt anzeigt – wie kann dann dieser Anstieg erklärt werden? Es könnte zur Erklärung eine allen Forschnungsfeldern inhärente Tendenz zur Diversifikation angenommen werden. Würde sich das bestätigen, könnte die Homogeniserungsthese nur durch den Vergleich von in Ländern und Fachgebieten unterschiedlich starken Trends getestet werden, was wir auch als nächstes vorhaben.

Die deutliche Tendenz zu längeren Referenzenlisten in beiden untersuchten Fachgebieten verdient eine weitere Untersuchung. Um zu prüfen, ob SVD-basierte Entropiemaße latenter Themen nicht durch Änderungen im Zitationverhalten beeinflusst werden, werden wir analysieren, welcherart Quellen jetzt mehr zitiert werden als früher.

Die von uns gefundene deutliche Tendenz zu höheren Entropiewerten sollte – so sie steigende Diversität anzeigt – auch durch die gewöhnliche Latente Semantische Analyse (LSA) der bipartiten Netzwerke von Artikeln und den in ihnen verwendeten Termen bestätigt werden. Unsere Ergebnisse können auch durch eine LSA-Variante, der Probabilistischen Latenten Semantischen Analyse (PLSA) getestet werden.

Gesellschaft für
Wissenschaftsforschung



Werner Ebeling
Heinrich Parthey (Hrsg.)

Selbstorganisation in Wissenschaft und Technik

Wissenschaftsforschung
Jahrbuch 2008

Sonderdruck

Mit Beiträgen von:
Werner Ebeling • Klaus Fischer
Klaus Fuchs-Kittowski • Jochen Gläser
Frank Havemann • Michael Heinz
Karlheinz Lüdtke • Oliver Mitesser
Heinrich Parthey • Andrea Scharnhorst

Wissenschaftsforschung
Jahrbuch **2008**

Bibliographische Informationen Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN 978-3-86573-45-9

Druck und Bindung: Schaltungsdienst Lange o.H.G., Berlin

Printed in Germany

38,00 Euro