

HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT



BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 240

LATENTE SEMANTISCHE ANALYSE ZUR MESSUNG
DER DIVERSITÄT VON FORSCHUNGSGEBIETEN

VON
OLIVER MITESSER

**LATENTE SEMANTISCHE ANALYSE ZUR MESSUNG
DER DIVERSITÄT VON FORSCHUNGSGEBIETEN**

**VON
OLIVER MITESSER**

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 240

Mitesser, Oliver

Latente Semantische Analyse zur Messung der Diversität von Forschungsgebieten / von Oliver Mitesser. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2008. – IX, 81 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 240)

ISSN 14 38-76 62

Abstract:

Vor dem Hintergrund aktueller Forschungspolitik ist die Hypothese einer abnehmenden Forschungsvielfalt plausibel. Gewissheit darüber gewinnt man allerdings erst dann, wenn eine entsprechende quantitative Analyse gelungen ist. In vorliegender Arbeit werden dazu zwei bibliometrische Methoden analysiert: die deterministische und die probabilistische Variante der Latenten Semantischen Analyse (LSA) in Kombination mit der Shannonschen Entropie als Diversitätsmaß. Die beiden statistischen Verfahren zeigen sich grundsätzlich für die Aufgabe geeignet und können verwendet werden, um die Vielfalt in den Forschungszweigen Informationswissenschaft und Elektrochemie zu messen. Im Gegensatz zur ursprünglichen Vermutung ergibt sich allerdings in allen untersuchten Fällen kein Absinken, sondern ein Anstieg bei der zeitlichen Entwicklung der Forschungsvielfalt innerhalb der letzten 20 Jahre.

Diese Veröffentlichung geht zurück auf eine Master-Arbeit im Studiengang Library and Information Science (Master of Arts) an der Humboldt-Universität zu Berlin.

Online-Version: <http://www.ib.hu-berlin.de/~kumlau/handreichungen/h240/>

Inhaltsverzeichnis

Problemstellung	1
1 Einleitung	3
1.1 Vielfalt als Voraussetzung für Entwicklung und Anpassung . . .	3
1.2 Diversitätsmaße	4
1.3 Kozitationsanalyse	6
1.4 Bibliographische Kopplung	7
2 Material und Methoden	11
2.1 Datengrundlage	11
2.1.1 Szientometrie	12
2.1.2 Elektrochemie	14
2.2 (Deterministische) Latente Semantische Analyse	15
2.2.1 Grundprinzip	15
2.2.2 Diversitätsbestimmung	18
2.3 Probabilistische Latente Semantische Analyse	19
2.3.1 Grundprinzip	19
2.3.2 Algorithmus	23
2.3.3 Diversitätsmaß	24
2.4 Implementierung	24
3 Ergebnisse	25
3.1 LSA	25
3.1.1 Grundfragen	25
3.1.2 Szientometrie	36
3.1.3 Elektrochemie	40
3.2 PLSA	43
3.2.1 Grundfragen	43
3.2.2 Szientometrie	59

4 Diskussion	61
4.1 LSA und PLSA	61
4.2 Diversitätsentwicklung	64
4.3 Weitere auswertbare Dokumenteigenschaften	66
4.4 Expertenbefragung	67
Zusammenfassung	68
Literaturverzeichnis	70

Abbildungsverzeichnis

1.1	Beitrag einer einzelnen relativen Häufigkeit zum Diversitätsmaß	5
2.1	Zeitliche Entwicklung der Publikationszahlen in ausgewählten Zeitschriften zur Szientometrie	13
2.2	Zeitliche Entwicklung der Publikationszahlen in ausgewählten Zeitschriften zur Elektrochemie	13
2.3	Schematische Darstellung der Referenz-Dokument-Matrix mit m Spalten (Dokumenten) und n Zeilen (Referenzen).	16
2.4	Schematische Darstellung des der PLSA zu Grunde liegenden statistischen Modells	21
3.1	Diversitätsverteilung für Zufallsmatrizen (LSA)	30
3.2	Diversitätsverteilung für Zufallsspaltenmatrizen (LSA)	33
3.3	Diversität in Abhängigkeit der Anzahl berücksichtigter Eigenwerte für zwei Typen von Zufallsmatrizen (LSA)	34
3.4	Relative Diversität in Abhängigkeit der Anzahl berücksichtigter Eigenwerte für zwei Typen von Zufallsmatrizen (LSA)	35
3.5	Artikelzahl innerhalb der Szientometrie-Jahrgänge	36
3.6	Screepplot zum Szientometrie-Jahrgang 1986	37
3.7	Screepplot zum Szientometrie-Jahrgang 2006	38
3.8	Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie (LSA)	38
3.9	Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie bei konstanter mittlerer Referenzzahl pro Dokument (LSA)	39
3.10	Artikelzahl innerhalb der Elektrochemie-Jahrgänge	40
3.11	Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Elektrochemie (LSA)	41
3.12	Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Elektrochemie bei konstanter mittlerer Referenzzahl pro Dokument (LSA)	42

3.13	Entropie in Abhängigkeit von der Log-Likelihood bei 100 Realisationen des PLSA Algorithmus für die völlig heterogene Bibliographie.	45
3.14	Entropie in Abhängigkeit von der Log-Likelihood bei 100 Realisationen des PLSA Algorithmus für die völlig homogene Bibliographie.	48
3.15	Entropie in Abhängigkeit von der Log-Likelihood bei 100 Realisationen des PLSA Algorithmus für eine schwach gekoppelte Bibliographie.	50
3.16	Diversität in Abhängigkeit der Themenzahl (PLSA)	58
3.17	Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie (PLSA mit 10 Themen)	59
3.18	Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie (PLSA mit 20 Themen)	60
4.1	Schematische Darstellung der LSA in Matrixform	63
4.2	Schematische Darstellung der PLSA in Matrixform	63

Tabellenverzeichnis

2.1	Ausgewertete Zeitschriften zur Szientometrie	12
2.2	Ausgewertete Zeitschriften zur Elektrochemie	14

Problemstellung

Um die Heterogenität oder Vielfalt einer Menge von Objekten – seien es Lebewesen, Gegenstände oder Konzepte – zu bewerten, ist es zunächst notwendig durch die Auswahl relevanter Charakteristika ein Mindestmaß an Vergleichbarkeit zwischen den Objekten herzustellen. Anhand der ausgewählten Eigenschaften können die Objekte dann z.B. paarweise als gleich oder ungleich identifiziert werden. Ein genügend differenziertes Vergleichsmaß erlaubt es auch, Ähnlichkeiten graduell zu quantifizieren. In einem weiteren Schritt muss es dann gelingen, die Bewertung des paarweisen Ähnlichkeitsgrades einzelner Objekte (Mikroebene) auf die Gesamtheit aller Objekte zu übertragen (Makroebene), um schließlich verschiedene Objektmengen miteinander vergleichen zu können. Dazu wird die Verteilung der vorliegenden individuellen Ähnlichkeitsmessungen weiter abstrahiert und jeder Objektmenge eine möglichst einfache und griffige Kenngröße zugeordnet, die schließlich ihre Vielfältigkeit in einem einzelnen Wert charakterisiert.

Für die szientometrische Analyse der Vielfalt in Forschungslandschaften wurden bisher auf der Mikroebene vor allem die Methoden der *Kozitation* und der *Bibliographischen Kopplung* verwendet. Damit konnte die Ähnlichkeit zwischen den wissenschaftlichen Artikeln eines Jahrgangs ausgewählter Zeitschriften bewertet werden. Die darauf aufbauende Klassifikation der Dokumente ließ sich dann nutzen, um die Vielfalt in der entsprechenden Forschungslandschaft zu messen. Als integriertes Maß für die Vielfalt wurde meist die *Shannonsche Entropie* gewählt.

Nachdem die bisherigen Vorgehensweisen methodische Schwierigkeiten aufweisen, sollen in dieser Arbeit alternative Ansätze untersucht werden, die weniger am paarweisen Vergleich der Objekte ansetzen als an der Gesamtstruktur der Objektmenge. Ziel ist es, mit der *Latenten Semantischen Analyse* sowohl in deterministischer als auch probabilistische Variante (LSA bzw. PLSA) vertraut zu werden und deren Potential für die Diversitätsmessung abzuschätzen. Beide Methoden werden bisher erfolgreich beim Information Retrieval eingesetzt. Im Rahmen dieser Arbeit sollen sie implementiert, analysiert und als Ausgangspunkt für die Diversitätsmessung und exemplarische

Untersuchung der Diversitätsentwicklung in den Forschungsbereichen *Szientometrie* und *Elektrochemie* verwendet werden.

Zuvor müssen die notwendigen Daten aus der Online-Datenbank *ISI Web of Science* gewonnen und für die automatisierte Auswertung vorbereitet werden.

Kapitel 1

Einleitung

1.1 Vielfalt als Voraussetzung für Entwicklung und Anpassung

Innerhalb der wissenschaftlichen Forschung entwickeln sich Fachgemeinschaften und Forschungsthemen, die weder personell noch inhaltlich scharf gegeneinander abgrenzbar sind. Thematische Vielfalt wird in Analogie zur Biodiversität dann als Vorteil angesehen, wenn man sich schnell neuen Bedingungen anpassen muss. Nachdem weder voraussehbar ist, welche Fragestellungen zukünftig relevant sein werden, noch, welcher Ansatz am Ende zur Lösung eines gegebenen Problems führen wird, ist thematische Breite günstig für die Fortentwicklung der Wissenschaft, so wie Biodiversität Voraussetzung ist für die Evolution der biologischen Arten. Der thematischen Vielfalt steht die häufig notwendige Konzentration der verfügbaren Mittel entgegen. Evaluationsbasierte Forschungsfinanzierung kann Nischen benachteiligen und eine Tendenz zum Mainstream erzeugen. Stellt sie wirklich – wie die Verfechter der sogenannten Homogenitätshypothese vermuten – eine Gefahr für die Vielfalt der Forschung dar?

Um diese Frage zu beantworten, ist es notwendig, die Forschungsvielfalt von Fachgebieten messbar zu machen und deren Entwicklung über einen relevanten Zeitraum zu verfolgen. Dies ist bisher nicht vollständig gelungen. Meinungen von Wissenschaftlern haben in dieser Frage keine verlässliche Evidenz, weil sie voreingenommen sein können. Einzuschätzen, welche der beantragten Forschungsvorhaben wichtig sind, ist nicht objektiv möglich. Non-konformistische Ansätze können von der Mehrheit einer Fachgemeinschaft als schlechte Wissenschaft wahrgenommen werden. Umgekehrt können Forscher die geringe Anerkennung ihrer Ergebnisse auf deren Spezifität zurückführen, um ihre u.U. mangelnde Qualität (vor sich und anderen) zu verbergen.

Um die Homogenitätshypothese zu testen, müssen Verfahren verwendet werden, die unabhängig davon sind, wie Wissenschaftler das Problem wahrnehmen. Bibliometrische Verfahren bieten sich für die Konstruktion objektiver Maße von Forschungsvielfalt an. Das Diversitätskonzept wurde jedoch bisher in der Wissenschaftsforschung selten verwendet und noch nicht befriedigend operationalisiert. Als Pionier auf diesem Gebiet hat Grupp 1990 einen bibliometrischen Ansatz zu Messung der Forschungsvielfalt vorgeschlagen. Mit dem Ziel, Interdisziplinarität bibliometrisch messbar zu machen, wurde sie von Bordons, Morillo und Gómez (2004) sowie von Rafols und Meyer (2007a, 2007b) auf thematische Diversität zurückgeführt. Eine generelle Diskussion der Anwendung von Diversitätsmaßen in der Wissenschaftsforschung hat Stirling 2007 vorgelegt. Ein von Havemann und Gläser (2007) initiiertes Ver- such zur bibliometrischen Messung von Forschungsvielfalt wird weiter unten näher vorgestellt.

1.2 Diversitätsmaße

Die Bewertung der Heterogenität oder Vielfalt einer Gruppe von Objekten ist nicht nur eine Frage nach der Anzahl der unterschiedlichen Klassen, in die sich die Objekte einteilen lassen. Vielfalt hängt auch davon ab, wie umfangreich die einzelnen Klassen sind und wie unterschiedlich sie untereinander sind.

Die biologische Vielfalt (Diversität) eines Biotops lässt sich grob dadurch beurteilen, dass man die Anzahl von Tier- und Pflanzenarten bestimmt, die dort gemeinsam leben. Größere Artenzahl scheint zunächst größere biologische Diversität zu bedeuten. Wird der Organismenbestand allerdings von einer einzigen Art dominiert, während alle anderen Arten nur sporadisch auftreten, kann ein weiterer, vergleichbarer Lebensraum trotz geringerer Artenzahl einen vielfältigeren Eindruck erwecken, wenn die Individuenzahlen gleichmäßiger auf die vorkommenden Arten verteilt sind. Neben der Artenzahl spielt also bei der Beurteilung auch die Verteilung der Individuen auf die Arten eine Rolle. Diese wird als *Evenness* oder auch als *Balance* der Verteilung bezeichnet (Stirling 2007). Ein Diversitätsmaß, das sowohl Artenzahl und Balance berücksichtigt, ist der mittlere Informationsgehalt der Aussage, dass ein Individuum im Biotop einer bestimmten Art angehört. Sie ergibt sich aus den relativen Häufigkeiten p_k der K Arten im Biotop zu

$$H = - \sum_{k=1}^K p_k \log p_k. \quad (1.1)$$

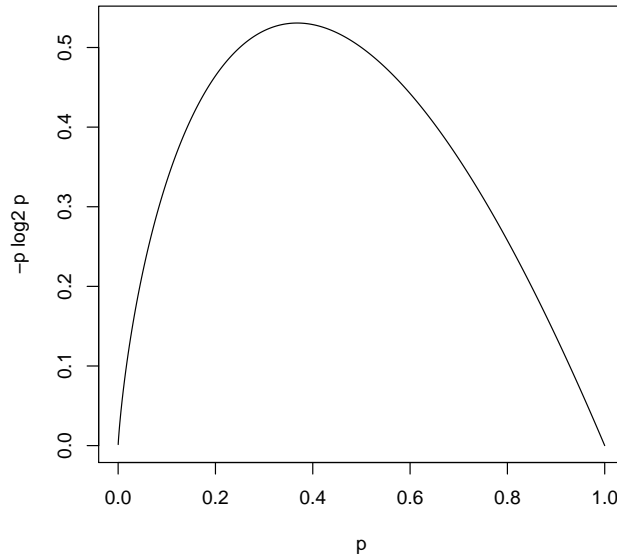


Abbildung 1.1: Beitrag $p \log_2 p$ einer einzelnen relativen Häufigkeit p zum Diversitätsmaß. Der Einfluss ist bei mittleren Anteilen $0 \leq p \leq 1$ am größten.

Die Größe H wird auch als *Shannon-Index* bezeichnet; Gleichung 1.1 ist mit der Boltzmannschen Entropie-Formel identisch.¹ Die Größe H wird maximal, wenn die Individuenzahl für alle Arten gleich ist ($H = -\sum_{k=1}^K \frac{1}{K} \log \frac{1}{K} = \log K$). In diesem Fall wird die Diversität alleine durch die Artenzahl bestimmt. Ein Biotop mit nur einer einzigen Art wird mit minimaler Diversität 0 bewertet.

Ein weiteres Maß für die Diversität, das Balance und Artenzahl berücksichtigt, ist die mittlere Wahrscheinlichkeit, dass zwei zufällig ausgewählte Individuen verschiedenen Arten k und l angehören:²

$$S = \sum_{k,l=1}^K p_k(1 - \delta_{kl})p_l = 1 - \sum_{k=1}^K p_k^2. \quad (1.2)$$

Dieses Maß geht auf Simpson (1949) zurück; $\sum_{k=1}^K p_k^2$ wird auch als *Simpson-Index* bezeichnet. Bei gleichmäßiger Individuenverteilung wird der Simpson-Index minimal und gleich $1/K$. Liegt nur eine Art vor wird $S = 1$.

¹Benutzt man in Gleichung 1.1 den dualen Logarithmus, erhält man den H -Wert in *bit*.

²Kroneckers δ_{kl} ist 1 für $k = l$ und sonst 0.

Neben Artenzahl und Individuenverteilung zwischen den Arten beeinflusst auch die Unterschiedlichkeit der Arten den Eindruck, den man von der Vielfalt eines Lebensraums gewinnt (*Disparity*). Die Formalisierung dieses Aspekts setzt allerdings voraus, dass ein plausibles Abstandsmaß zwischen den Arten existiert. Wenn in einem Waldstück Nadelbäume wachsen, in einem anderen aber ein Mischwald vorliegt, dann wird man den Mischwald auch bei identischer Artenzahl und Balance als vielfältiger einschätzen. Wir gehen damit von der dichotomen Bewertung von Arten – Arten sind entweder *gleich* oder *ungleich* – über zu einer graduellen Unterscheidung, die sich im biologischen Fall mit dem genetischen Verwandtschaftsgrad zwischen den Arten assoziieren lässt. In Gleichung 1.2 muss entsprechend $1 - \delta_{kl}$ durch ein Maß D_{kl} für die Unterschiedlichkeit zweier Arten k und l ersetzt werden, das Werte zwischen 0 und 1 annehmen kann (Shimatani 2001). Man misst dann mit

$$\bar{D} = \sum_{k,l=1}^K p_k D_{kl} p_l \quad (1.3)$$

die mittlere Disparität eines zufälligen Paares.

Die Matrix D_{kl} für die Disparität ($0 \leq D_{kl} \leq 1$) kann durch eine Distanzmatrix d_{kl} ersetzt werden, in der auch Werte > 1 auftreten. Dann erfasst man Diversität durch eine mittlere (taxonomisch oder genetisch definierte) Entfernung \bar{d} zwischen den Individuen des Biotops. Wenn nur eine Art betrachtet wird, kann ihre genetische Diversität durch die mittlere genetische Entfernung zwischen ihren K Individuen gemessen werden, welche i.A. alle genetisch verschieden sind. Dann werden in Gleichung 1.3 die relativen Häufigkeiten alle zu $p_k = 1/K$ und man erhält als Maß ganz einfach³

$$\bar{d} = \frac{1}{K^2} \sum_{k,l=1}^K d_{kl}.$$

1.3 Kozitationsanalyse

Wie gelingt es nun im bibliometrischen Bereich, eine Klassifikation in Analogie zu den biologischen Arten zu gewinnen und daraus ein geeignetes Diversitätsmaß? Eine oft erprobte Methode, die thematische Struktur wissenschaftlicher Zeitschriftenliteratur sichtbar zu machen, ist die auf Marshakova (1973) und auf Small (1973) zurückgehende *Kozitationsanalyse*. Auf Basis

³Die mittlere Entfernung wird eigentlich durch Division mit $K(K-1)$ statt mit K^2 berechnet; K ist aber meist groß genug, so dass dieser Unterschied nicht wesentlich ist, weil $K \approx K-1$.

des *Science Citation Index (SCI)* werden zunächst die für ein Forschungsgebiet relevanten Zeitschriften ausgewählt. In einem Zeitschriftenjahrgang oft zitierte Quellen dienen dann als Symbole für Standardkonzepte (Small 1978). Werden sie auch oft kowitziert, zeigt das ihre thematische Nähe an. Mit dem Salton-Index der Kowitzierung als Ähnlichkeitsmaß werden bei Small und Sweeney (1985) und Small, Sweeney und Greenlee (1985) mittels *single-linkage clustering* Kowitzierungscluster hochzitiertter Referenzen gebildet, wobei die Schwellenwerte von Zitierung und Kowitzierung variiert werden können. Diese Cluster werden dann auf den Jahrgang der zitierenden Aufsätze zurückprojiziert, indem die jeweils ein Cluster zitierenden Arbeiten als eine Forschungsfront angesehen werden.

Ein Versuch, die Größen von Kowitzierungsclustern und von Forschungsfronten in einem Fachgebiet als Ausgangsgrößen für das Entropiemaß der Forschungsvielfalt zu verwenden, zeigt aber schnell, dass diese dafür wenig geeignet sind (Schmidt u. a. 2006). 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 – muss der Zitationsschwellenwert auf ein Minimum herabgesetzt werden. Dadurch kommt aber eine negative Eigenschaft des *single-linkage clustering* zur Wirkung: das *chaining*. Dieser Effekt besteht darin, dass langgezogene Kowitzierungscluster entstehen, deren Enden thematisch wenig miteinander zu tun haben. Ab einem bestimmten Schwellenwert für den Salton-Index der Kowitzierung werden dadurch fast alle zitierten Quellen in einem großen Cluster versammelt.

Dieses negative Ergebnis deutet darauf hin, dass die Klassifizierung der wissenschaftlichen Artikel eines Fachgebiets nach disjunkten Themen – in Analogie zu den disjunkten Arten in einem Biotop – zu grundsätzlichen Schwierigkeiten führt. Auch wenn durch eine andere Clustermethode der *chaining*-Effekt möglicherweise vermeidbar ist, so bleibt die Schwierigkeit, eine Arbeit genau einem thematischen Cluster zuzuordnen.

1.4 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 (Have-mann u. a. 2007). 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 die unmittelbaren geistigen Vorfahren eines Werkes in der Liste der zitierten Quellen aufgeführt. Diese bibliographische Information ist allerdings weitaus unvollständiger als die genetische: keinesfalls ist aus ihr der gesamte geistige Stammbaum ablesbar. 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.

Von einem solchen Unterfangen nahmen Havemann u. a. (2007) wegen des hohen Aufwands zunächst Abstand und versuchten, allein mit den unmittelbaren Vorfahren ein Abstandsmaß zwischen Artikeln zu konstruieren, das 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. Die Zahl gleicher zitierter Quellen kann auch hier 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 $K(K-1)/2$ möglichen Kopplungen zwischen K Artikeln sind realisiert ($< 1\%$). 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 ($> 90\%$), 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 (Havemann u. a. 2007). Dieses Vorgehen ist ganz ähnlich dem von Botafogo, Rivlin und Shneiderman (1992) und von Egge und Rousseau (2003), die ein Maß für die Kompaktheit von Netzwerken aus mittleren Längen kürzester Pfade ableiten.⁴

Havemann u.a. haben 2007 für die elf Jahrgänge 1995–2005 von 14 elektrochemischen Zeitschriften⁵ jeweils die Länge aller kürzesten Pfade in der Hauptkomponente berechnet. Die Entfernung zwischen zwei direkt bibliographisch gekoppelten Artikeln k und l wurde mit $d_{kl} = -\log(J_{kl})$ berechnet. Dabei bezeichnet $J_{kl} \leq 1$ den Jaccard-Index der bibliographischen Kopp-

⁴s. a. den Konferenzbeitrag von Rafols und Meyer (2007a)

⁵Verwendet wurde derselbe Satz von Zeitschriften wie bei Schmidt u. a. (2006) und alle Datensätze vom Dokumenttyp *Article* oder *Letter* aus dem *ISI Web of Science* (WoS).

lung, der als Verhältnis der Länge von Durchschnitt und von Vereinigung der Referenzlisten R_k und R_l der beiden Artikel definiert ist:

$$J_{kl} = \frac{|R_k \cap R_l|}{|R_k \cup R_l|}.$$

Die mittlere Entfernung \bar{d} 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. Diese Tendenz konnte allerdings nicht auf eine sinkende Forschungsvielfalt zurückgeführt werden. Tatsächlich bewegt sich das (geometrische) Mittel der Länge der Referenzlisten der Artikel entgegengesetzt zur mittleren Entfernung in der Hauptkomponente ab 2000/2001 von 18,6 nach oben auf 21,7.⁶ 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.

Ob die Abnahme der mittleren Distanz völlig durch die Zunahme der Kantenzahl erklärt werden kann, ließ sich überprüfen, indem man aus den empirischen Netzwerken Modellgraphen konstruierte, in denen zitierte Quellen zufällig aus den Referenzlisten gelöscht wurden, bis die mittlere Referenzanzahl in allen Doppeljahrgängen gleich war. Die Artikelanzahl 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 eine Zeitreihe vergleichbarer mittlerer Entfernungen \bar{d} zu erhalten, mussten also auch die Artikelanzahl pro Doppeljahrgang durch gleich große Zufallsstichproben normiert werden.

Tatsächlich verschwindet durch diese Prozedur die fallende Tendenz für \bar{d} vollkommen. Beim Messen von Diversität sind zufällige Stichproben von Individuen zulässig, aber zufälliges Streichen von Referenzen 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, das nichts mit Änderungen der Forschungsdiversität zu tun haben muss.

Nachdem die bisherige Suche nach einer robusten Analyse von Forschungsvielfalt nur wenig erfolgreich verlaufen ist, sind neue Ansätze erforderlich. Ein möglicher Kandidat für eine neue Methode ist die *Latente Semantische Analyse*, die im Folgenden sowohl als deterministische als auch probabilistische

⁶Weil die Verteilung der Längen der Referenzlisten schief ist, wurde das geometrische und nicht das arithmetische Mittel als Maßzahl der zentralen Tendenz benutzt.

Variante vorgestellt, in ihren grundlegenden Eigenschaften analysiert und auf aktualisierte und erweiterte Daten aus den Forschungsbereichen *Szientometrie* und *Elektrochemie* angewendet werden soll.

Beide Methoden lassen hoffen, die vorhandenen Schwierigkeiten zu überwinden, und erfüllen einige plausible Anforderungen, die man nach den bisherigen Erfahrungen sinnvollerweise an ein Verfahren zur Diversitätsmessung stellen sollte:

1. Jede Veröffentlichung darf mehreren Themen zugewiesen werden.
2. Jede Referenz (die ebenfalls eine Veröffentlichung ist) darf mehreren Themen zugewiesen werden.
3. Jegliche Information, die in der bibliographischen Kopplung von Veröffentlichungen steckt, und die in der Kozitation von Referenzen enthalten ist, soll für die Auswertung genutzt werden.

Durch diese Anforderungen wird die Asymmetrie zwischen Veröffentlichungen und Referenzen aufgehoben.

Kapitel 2

Material und Methoden

2.1 Datengrundlage

Die Datengrundlage für alle späteren Auswertungen besteht aus den Metadaten zu wissenschaftlichen Veröffentlichungen in ausgewählten Fachzeitschriften der Bereiche *Szientometrie* und *Elektrochemie*. Die verwendeten Metadaten wurden aus der Fachdatenbank *ISI Web of Science* heruntergeladen und liegen als Textdateien vor. Die Datensätze konnten in Paketen zu je 500 Veröffentlichungen erfasst werden (das Maximum für einen einzelnen Download aus der Datenbank). Als Format wurde *Tab-delimited (Windows)* gewählt. Die einzelnen Metadaten-Datensätze entsprechen einzelnen wissenschaftlichen Veröffentlichungen und umfassen jeweils eine Reihe von Feldern wie Autorennamen, Titel, Abstract und andere. Die Datenfelder von größter Bedeutung für diese Arbeit sind die Liste der in der Veröffentlichung zitierten Referenzen und der Dokumententyp. Obwohl die Datensätze mit allen verfügbaren Feldern heruntergeladen wurden, wird in dieser Arbeit nur ein kleiner Teil davon verwendet. Für zukünftige Analysen kann aber auch auf die anderen Eigenschaften der Publikationen zurückgegriffen werden, die in den Metadaten erfasst sind.

Nach dem Download wurden die Textdateien mit den Metadaten zu je 500 Dokumenten nach folgenden Richtlinien weiterverarbeitet, um die spätere Analyse mit der Statistiksoftware *R* vorzubereiten:

- Alle Metadaten zu den Publikationen einer einzelnen Zeitschrift wurden aus verschiedenen Textdateien in eine oder wenige Textdateien mit je maximal 20 MB Datenumfang zusammengefasst. Jede Textdatei enthält eine Kopfzeile mit den Feldkennzeichnern aus dem ISI Web of Science.
- Als Feldtrennzeichen wird der vertikale Strich | verwendet. Stichproben

Tabelle 2.1: Ausgewertete Zeitschriften zur Szientometrie. Das Kürzel in der Spalte Datei liefert in der Form *Kürzel* + „.txt“ die Dateinamen der zugehörigen Textdateien. Hinter dem Zeitschriftentitel ist die Anzahl von Publikationen bzw. von Publikationen vom Typ „Article“ angegeben. (Abkürzungen: JAM = Journal of the American Society; Inf. = Information; J. = Journal)

Zeitschrift	Zeitraum	Datei
Inf. Processing & Management (2887/1757)	1963-2007	IPM62-07
Scientometrics (1683/1442)	1980-2007	STM80-06
J. of Documentation (2500/631)	1973-2007	JOD73-07
J. of Information Science (1801/1182)	1968-2007	JIS68-07
JAM for Inf. Science and Technology (1118/822)	2001-2007	JAS56-07
(zuvor: JAM for Inf. Science (2907/1654))	1970-2000	JAS56-07
(zuvor: American Documentation (795/482))	1956-1969	JAS56-07

zu Beginn haben nahe gelegt, dass dieses Zeichen selbst nicht in den Datensätzen verwendet wird. Diese Vermutung hat sich bestätigt.

- Als Textkennzeichnung wird das Gradzeichen ° verwendet. Nachdem sich erst sehr spät herausgestellt hat, dass dieses Zeichen selbst (äußerst selten) in den Datensätzen erscheint, wurde es vor der Verwendung als Textindikator durch den String *Kringelchen* ersetzt, der nirgends in den Originaldaten vorkommt. Diese Veränderung der Originaldaten hat keinen Effekt auf die Auswertung.

2.1.1 Szientometrie

Für den Bereich Szientometrie wurden fünf Zeitschriften ausgewertet (siehe Tabelle 2.1). Die späteren Analysen basieren nur auf den Veröffentlichungen vom Typ „Article“. Um bei der Auswertung auf einen hinreichend großen minimalen Artikelpool pro Jahrgang (mindestens 100) zurückgreifen zu können, wurde die Auswertung auf den Zeitraum 1986 bis 2007 eingeschränkt. Frühere Jahrgänge enthalten zu wenig Datensätze.

Die zeitliche Entwicklung von Veröffentlichungs- und Referenzanzahlen geben einen groben Überblick über den Datenbestand. Die Anzahl von Veröffentlichungen pro Jahrgang zeigt im Laufe der Zeit einen deutlichen Anstieg (Verdreifachung). Dasselbe gilt für die mittlere Anzahl an Zitationen und die mittlere Anzahl unterschiedlicher zitierter Referenzen (siehe Abbildung 2.1). In den Abbildungen 2.1 sind alle Publikationstypen erfasst (siehe Abbildung 3.5 hinten nur für die Dokumente vom Typ „Article“).

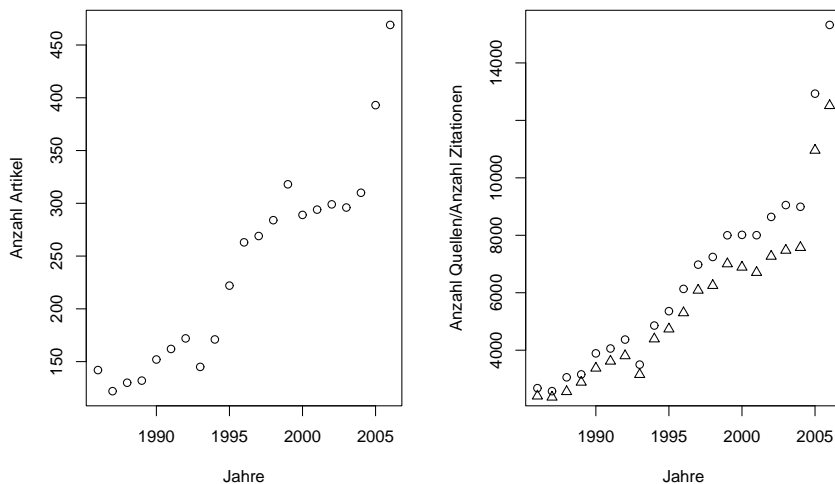


Abbildung 2.1: Zeitliche Entwicklung der Publikationszahlen in den ausgewählten Zeitschriften zur Szientometrie (siehe Tab. 2.1): Anzahl der Veröffentlichungen in einzelnen Jahrgängen (links) und Gesamtzahl der darin enthaltenen Zitationen (rechts, Kreise) bzw. der (unterschiedlichen) Referenzen (rechts, Dreiecke).

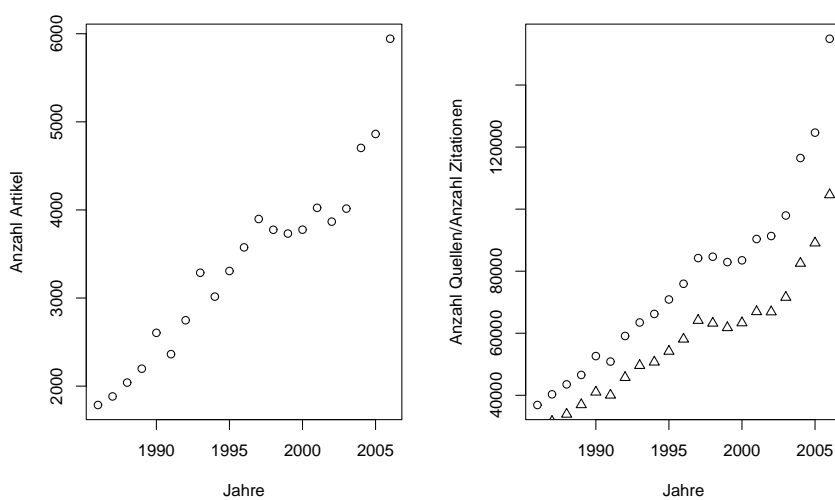


Abbildung 2.2: Zeitliche Entwicklung der Publikationszahlen in den ausgewählten Zeitschriften zur Elektrochemie (siehe Tab. 2.2). Symbole wie oben.

2.1.2 Elektrochemie

Für das Gebiet *Elektrochemie* wurden 14 Zeitschriften ausgewählt (siehe Tabelle 2.2; zur Relevanz der fachlichen Auswahl siehe Schmidt, Gläser, Havemann und Heinz (2006)). Die Jahrgänge von 1986 bis 2006 enthalten deutlich mehr Publikationen als die Zeitschriften zur Szientometrie (> 1000), so dass man auch einen größeren Zeitraum hätte untersuchen können. Die Darstellung wurde aber der unmittelbaren Vergleichbarkeit wegen identisch gewählt. In der späteren Auswertung wurden immer 500 zufällig gewählte Publikationen pro Jahrgang vom Typ „Article“ verwendet. Der relative Anstieg der Gesamtzahlen von Veröffentlichungen in den elektrochemischen Zeitschriften ist mit dem aus der Szientometrie vergleichbar (siehe Abbildung 2.2 für alle Dokumente und Abbildung 3.10 nur für den Typ „Article“).

Tabelle 2.2: Ausgewertete Zeitschriften zur Elektrochemie. Das Kürzel in der Spalte Datei liefert in der Form „ElChemXX“ + *Kürzel* + (evtl. Nummer) + „.txt“ die Dateinamen der zugehörigen Textdateien (siehe auch die Zeilen zum Dateneinlesen im Quellcode im Anhang). Die Aufteilung einer Zeitschrift in mehrere nummerierte Einzeldateien wurde dann vorgenommen, wenn die Dateigröße 20 MB überschritten hätte. Hinter dem Zeitschriftentitel ist in Klammern die Anzahl von Publikationen bzw. von Publikationen vom Typ „Article“ angegeben. (Abkürzungen: J. = Journal; Bioelchem. = Bioelectrochemistry)

Zeitschrift	Zeitraum	Datei
Bioelectrochemistry (659/638)	2000-2007	BIO
(zuvor: Bioelchem. and Bioenergetics (2057/1807))	1975-1999	BAB
Chemical Vapor Deposition (660/568)	1995-2007	CVD
Corrosion Science (5488/5092)	1978-2007	COS
Electroanalysis (4010/3673)	1989-2007	ELA
Electroanalytical Chemistry (26/0)	-	ELC
Electrochimica Acta (13988/13094)	1967-2007	ECA
J. of Applied Electrochemistry (4661/4381)	1974-2007	AEC
J. of Electroanalytical Chemistry (18375/15381)	1964-2007	JEC
J. of Power Sources (8983/8546)	1976-2007	JPS
J. of the Electrochemical Society (53725/25406)	1948-2007	JES
Plating and Surface Finishing (5482/2836)	1977-2007	PSF
Russian J. of Electrochemistry (3222/2913)	1993-2007	RJE
Sensors and Actuators B Chemical (7483/7352)	1990-2007	SAC
Solid State Ionics (10199/9764)	1981-2007	SSI

2.2 (Deterministische) Latente Semantische Analyse

Die *Latente Semantische Analyse (LSA)* wurde in den 80er Jahren unter der Federführung von T. Landauer entwickelt und 1988 in den USA von S. Deerwester, S. Dumais, G. Furnas, R. Harshman, T. Landauer, K. Lochbaum und L. Streeter patentiert (US Patent 4839853). Nach der ersten wissenschaftlichen Veröffentlichung von Deerwester, Dumais, Furnas, Landauer und Harshman (1990) hat sich die Methode international schnell durchgesetzt und wird inzwischen in vielen Bereichen der quantitativen Sozialforschung genutzt (Landauer, McNamara, Dennis und Kintsch 2007).

2.2.1 Grundprinzip

Unter einer Bibliographie B soll im Folgenden eine (beliebig geordnete) Folge von m Dokumenten verstanden werden ($j = 1, \dots, m$). Alle Bibliographien, die in dieser Arbeit betrachtet werden, bestehen aus den Artikeln, die innerhalb eines bestimmten Jahres in einer Reihe von ausgewählten Zeitschriften erschienen sind. Der zur Bibliographie B gehörige Referenzenpool R_B ist die (ebenfalls beliebig geordnete) Menge aller n (unterschiedlichen) Referenzen R_i ($i = 1, \dots, n$), die in mindestens einem der Dokumente der Bibliographie B zitiert werden¹. Ein Dokument lässt sich durch einen Vektor der Länge $n = |R_B|$ repräsentieren. Das Element x_{ij} des Dokumentvektors \vec{x}_j ist 1, wenn im Dokument die entsprechende Referenz R_i vorhanden ist, andernfalls ist das Element des Vektors 0. Damit ist ein Dokument zunächst alleine durch seine Referenzenliste bestimmt.

Die gesamte Bibliographie kann man durch die Folge der Dokumentvektoren \vec{x}_j ($j = 1, \dots, m$) als Referenz-Dokument-Matrix X darstellen. Jedem Dokument der Bibliographie entspricht eine Spalte und jeder Referenz eine Zeile der Matrix X (siehe Abbildung 2.3). X ist eine $n \times m$ Rechtecksmatrix mit n Zeilen und m Spalten. Typischerweise besitzt die Matrix X deutlich mehr Zeilen (Referenzen) als Spalten (Dokumente), so dass hier immer $n \gg m$ gilt. Ziel ist es, aus der Referenz-Dokument-Matrix latente Themen zu extrahieren, die durch verallgemeinerte Dokumente (in Form fiktiver Referenzenlisten mit beliebigen, reellen Einträgen) repräsentiert werden. Die Themenextraktion erfolgt durch Latente Semantische Analyse, deren algorithmischer Kern eine Singulärwertzerlegung (SVD = singular value decomposition) ist. Mit r ($r \leq m \ll n$) soll der Rang der Matrix X bezeichnet werden. Die Indizes i

¹Beachte: Das Symbol R_i wird von jetzt an in Abweichung von der Terminologie in der Einleitung für eine einzelne Referenz und nicht für eine Referenzenliste verwendet.

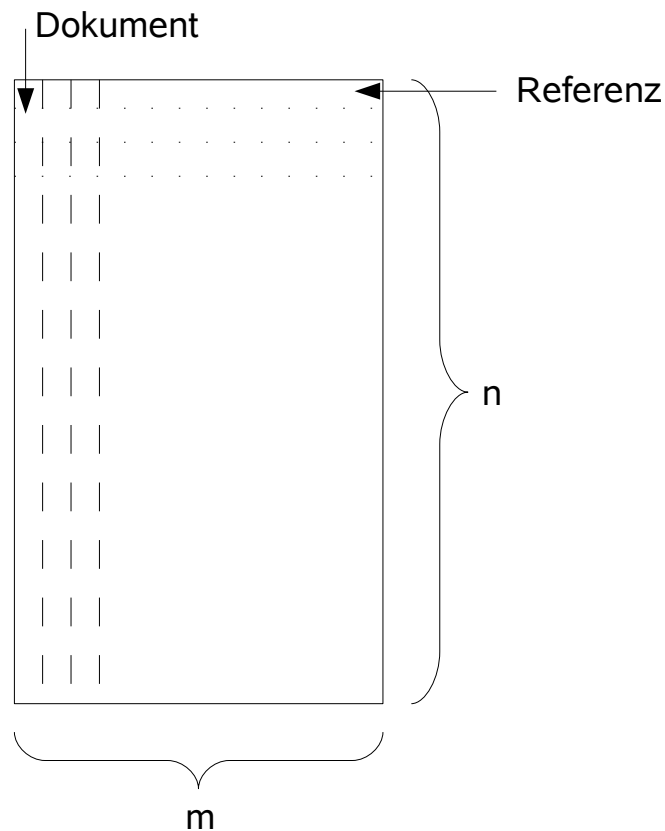


Abbildung 2.3: Schematische Darstellung der Referenz-Dokument-Matrix mit m Spalten (Dokumenten) und n Zeilen (Referenzen).

(für Referenzen), j (für Dokumente) und k (für Themen) laufen im Folgenden immer nach demselben Schema: $i = 1, \dots, n$, $j = 1, \dots, m$ und $k = 1, \dots, r$. Die Singulärwertzerlegung von X lässt sich durch

$$X = U\Lambda^{1/2}V^T \quad (2.1)$$

darstellen (siehe auch Abbildung 4.1 hinten).

Die r Spalten von U ($n \times r$) sind die normierten Eigenvektoren der Matrix XX^T ($n \times n$) zu den von Null verschiedenen Eigenwerten λ_k von XX^T . Die Matrix XX^T enthält die Kozitationsbeziehungen der n Quellen.

Die r Spalten von V ($r \times r$) sind die normierten Eigenvektoren der Matrix $X^T X$ ($m \times m$) zu von Null verschiedenen Eigenwerten. Die Matrix $X^T X$ enthält die bibliographischen Kopplungen der m 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, siehe (Golub und Loan 2007)).

Beispiel:

$$\begin{aligned} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} &= \begin{pmatrix} -0.271 & 0.500 & -0.653 \\ -0.653 & 0.500 & 0.271 \\ -0.653 & -0.500 & 0.271 \\ -0.271 & -0.500 & -0.653 \end{pmatrix} \\ &\times \begin{pmatrix} 1.848 & 0.000 & 0.000 \\ 0.000 & 1.414 & 0.000 \\ 0.000 & 0.000 & 0.765 \end{pmatrix} \\ &\times \begin{pmatrix} -5.00e-01 & 7.07e-01 & -5.00e-01 \\ -7.07e-01 & 3.18e-15 & 7.07e-01 \\ -5.00e-01 & -7.07e-01 & -5.00e-01 \end{pmatrix} \end{aligned}$$

Man nimmt nun an, dass aus dem Netzwerk r unabhängige latente Themen extrahierbar sind. Dazu werden in einem linearen Ansatz die m Spaltenvektoren \vec{x}_j von X nach der r -dimensionalen Orthonormalbasis U entwickelt:

$$\vec{x}_j = \sum_{k=1}^r \vec{u}_k a_{jk}. \quad (2.2)$$

Gleichung 2.2 kann kompakt als $X = UA^T$ geschrieben werden. Die m Spalten von A^T enthalten die Koordinaten der m Artikel in der neuen Basis U . An dieser Stelle wird die zentrale Motivation für die Verwendung der SVD deutlich: die Transformation der Referenz-Dokument-Matrix in eine linear unabhängige Form.

Beispiel:

$$\vec{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.271 & 0.500 & -0.653 \\ -0.653 & 0.500 & 0.271 \\ -0.653 & -0.500 & 0.271 \\ -0.271 & -0.500 & -0.653 \end{pmatrix} \times \begin{pmatrix} 1.848 \cdot -5.00e-01 \\ 1.414 \cdot -7.07e-01 \\ 0.765 \cdot -5.00e-01 \end{pmatrix} \quad (2.3)$$

Ganz analog geht man bei den Quellen vor und erhält $X^T = VB^T$ bzw. $X = BV^T$. Hier ist die Basis V und die Spalten von B^T enthalten die Koordinaten der n Quellen. Der Vergleich mit Gleichung 2.1 ergibt dann für die neuen Koordinaten der Quellen $B = U\Lambda^{1/2}$ und für die der Artikel $A^T = \Lambda^{1/2}V^T$

bzw. $A = V\Lambda^{1/2}$. Letzteres sieht in Koordinatenschreibweise folgendermaßen aus:

$$a_{jk} = \sum_{l=1}^r v_{jl} \delta_{lk} \lambda_k^{1/2} = v_{jk} \lambda_k^{1/2}. \quad (2.4)$$

Die r Koordinaten in jeder der m Zeilen von A geben an, wie groß die Komponenten des jeweiligen Artikels j in die Richtungen der r Themen sind. Im Allgemeinen sind die a_{jk} nicht immer ≥ 0 . Das Vorzeichen dreht sich, wenn der Eigenvektor k seinen Richtungssinn ändert (der nicht festgelegt ist). Als Anteil des Themas k am Artikel j wird jedoch nicht a_{jk} angesetzt, sondern a_{jk}^2 (Alter, Brown und Botstein 2000). Dann spielt das Vorzeichen keine Rolle mehr. Die Summe der Themenanteile eines Artikels a_{jk}^2 ist gleich der euklidischen Norm des Artikel-Vektors \vec{x}_j , die sich beim Wechsel zur neuen Basis U nicht ändert:

$$\sum_{k=1}^r a_{jk}^2 = |\vec{x}_j| = \sum_{i=1}^n x_{ij}^2. \quad (2.5)$$

Weil X binär ist, gilt andererseits

$$\sum_{i=1}^n x_{ij}^2 = \sum_{i=1}^n x_{ij} = |R_{\vec{x}_j}|. \quad (2.6)$$

Die Themenanteile eines Artikels j summieren sich also zur Länge seiner Referenzliste $|R_{\vec{x}_j}|$ auf. Wir können dann für jeden Artikel j aus den relativen Themenanteilen $p_k = a_{jk}^2 / |R_{\vec{x}_j}|$ sofort seine thematische Vielfalt bestimmen, indem wir mit den p_k seine Entropie oder seinen Simpson-Index berechnen (Gleichungen 1.1 und 1.2, S. 4). Nachdem in dieser Arbeit Diversität nur mit Hilfe der Entropieformel gemessen wird, werden die Begriffe „Diversität“ und „Entropie“ im Folgenden synonym verwendet.

2.2.2 Diversitätsbestimmung

Summiert man die Beiträge zweier Artikel, $j = 1$ und $j = 2$, zu einem Thema k , dann ist die Summe offenbar mit $a_{1k}^2 + a_{2k}^2$ anzusetzen. Die relativen Anteile werden dann $p_k = (a_{1k}^2 + a_{2k}^2) / (|R_{\vec{x}_1}| + |R_{\vec{x}_2}|)$. Summiert man so über alle m Artikel und setzt für a_{jk} den Ausdruck aus Gleichung 2.4 ein, erhält man

$$\sum_{j=1}^m a_{jk}^2 = \lambda_k \sum_{j=1}^m v_{jk}^2 = \lambda_k, \quad (2.7)$$

weil die euklidische Norm der Spaltenvektoren von V gleich 1 ist. Der Beitrag von Thema k zum gesamten Jahrgang ist also gleich dem Eigenwert λ_k . Die

Summe aller Eigenwerte ist gleich dem Quadrat der Frobenius-Norm von Matrix X und damit gleich der Zahl der Links im Netzwerk:

$$|X|_{\mathbb{F}}^2 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}. \quad (2.8)$$

Die Diversität eines Jahrgangs kann dann aus den relativen Anteilen $p_k = \lambda_k/|X|_{\mathbb{F}}^2$ berechnet werden. Diese Vorgehensweise wird beispielsweise auch in der Biologie verwendet, um die Komplexität genetischer Daten zu quantifizieren (Alter, Brown und Botstein 2000). Die analoge Rechnung für die Anteile der Themen an den zitierten Quellen führt zum gleichen Ergebnis.

Bei vielen SVD-gestützten Methoden – wie z.B. der LSA – wird die Zahl der Dimensionen des Vektorraums künstlich verringert, indem die sehr kleinen Eigenwerten entsprechenden Eigenvektoren weggelassen werden (Bortz 2005). Solcherart Dimensionsreduzierung macht die extrahierten Themen für praktische Zwecke übersichtlicher. Aber auch hier sollten (wie bei der Koziationsanalyse, siehe oben) die kleinen Themen nicht vernachlässigt werden, wenn man Vielfalt messen will.

2.3 Probabilistische Latente Semantische Analyse

Im Jahr 1999 hat Hofmann eine statistische Variante der Latenten Semantischen Analyse vorgeschlagen, die *Probabilistische Latente Semantische Analyse (PLSA)* als Realisierung eines Aspekte-Modells (Hofmann 1999b; Hofmann 1999a; Hagenaars und McCutcheon 2002).

2.3.1 Grundprinzip

Die PLSA beruht auf einem konkreten statistischen Modell, das durch Zufallsexperimente realisiert werden kann. Die wiederholte Realisation soll die beobachtete Bibliographie verglichen mit allen anderen Bibliographien, die das Experiment als mögliche Ergebnisse liefern kann, mit größtmöglicher Wahrscheinlichkeit reproduzieren. Für eine eingängige Darstellung bietet es sich an, die Bibliographie nicht mehr durch die Matrix X zu repräsentieren, sondern durch eine Menge Y_B , die aus denjenigen Paaren (\vec{x}_j, R_i) – oder kurz (j, i) – besteht, für die gilt: $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$ mit der Einschränkung $R_i = 1$ in \vec{x}_j . Damit besteht die Menge Y_B einfach aus allen Dokument-Referenz-Paaren, die beobachtet wurden.

Im Allgemeinen wird durch ein bestimmtes Zufallsexperiment allerdings eine andere Bibliographie-Menge $Y = \{(j, i)\}$ erzeugt mit derselben Anzahl

an Dokument-Referenz-Paaren wie Y_B (d.h. nicht alle Kombinationen von i und j kommen vor, manche aber evtl. sogar mehrfach). Eine Vektordarstellung wäre an dieser Stelle recht umständlich. Von nun an ist mit „Dokument j “ nicht mehr die konkrete Realisation \vec{x}_j mit den beobachteten Referenzen gemeint, sondern das Konzept „Dokument j “, das mit einer spezifischen Verteilung von latenten Themen assoziiert ist². Die latenten Themen T_k , $k = 1, \dots, K$ sind die verborgenen Klassifizierungsvariablen des Modells. Diese sollen im Folgenden charakterisiert werden. Der Index k wird dazu nun – anders als im letzten Abschnitt – für die latenten Themen verwendet. Es wird angenommen, dass ein Thema T_k aus dem Themenpool mit einer bestimmten Wahrscheinlichkeit $P(k|j)$ auftritt, wenn ein Dokument j vorliegt (bedingte Wahrscheinlichkeit). Auf ähnliche Weise sind die Referenzen an die Themen gekoppelt. $P(i|k)$ ist die Wahrscheinlichkeit, mit der eine Referenz R_i realisiert wird, wenn ein Thema T_k vorliegt (siehe Abbildung 2.4). In seiner konstruktiven Interpretation lässt sich ein bekanntes (und durch die beiden Matrizen $P(i|k)$ und $P(k|j)$ parametrisiertes) PLSA-Modell dann damit wie folgt beschreiben (Hofmann 1999b):

1. Wähle zuerst ein Dokument j' gemäß der Wahrscheinlichkeiten $P(j)$ ($j = 1, \dots, m$) aus dem Dokumentenpool (zur Bedeutung der Dokumentwahrscheinlichkeit $P(j)$ später). Dies entspricht einem Würfelwurf mit einem m -seitigen Würfel, bei dem die Seiten nicht zwangsläufig alle gleich wahrscheinlich sind. (Beachte nochmals, dass mit „Dokument j “ nicht mehr die konkrete Realisation \vec{x}_j mit den beobachteten Referenzen gemeint ist, sondern die bedingten Wahrscheinlichkeiten $P(k|j)$, $k = 1, \dots, K$, die in der statistischen Perspektive einen Dokumenttyp charakterisieren.)
2. Wähle nun für das Dokument j' ein Thema T'_k gemäß der Wahrscheinlichkeiten $P(k|j')$, $k = 1, \dots, K$. Dies entspricht einem Würfelwurf mit einem K -seitigen Würfel.
3. Das gewählte Thema T'_k liefert Wahrscheinlichkeiten $P(i|k')$ dafür, dass die einzelnen Referenzen R_i im Dokument j' auftreten. Mit Hilfe dieser Wahrscheinlichkeiten wird schließlich eine Referenz i' gezogen.

Auf diese Weise konstruiert man ein Paar (j', i') . Die Paarkonstruktion wird sooft wiederholt, wie es Dokument-Referenz-Paare in der beobachteten Bibliographie gibt. Auf diese Weise wird eine vollständige Bibliographie erzeugt.

²Der Begriff „Konzept“ wird hier nicht im mathematischen Sinn der Verbandstheorie sondern umgangssprachlich verwendet.

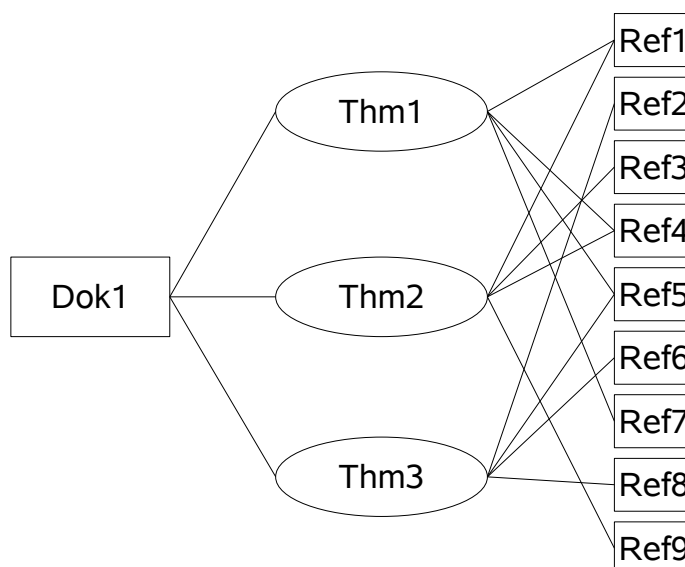


Abbildung 2.4: Schematische Darstellung des der PLSA zu Grunde liegenden statistischen Modells. Zwischen die messbaren Ebenen von Dokumenten und Referenzen wird die latente Ebene der Themen eingefügt.

Neben den Matrizen $P(i|k)$ und $P(k|j)$ sind dazu noch die zusätzlichen freien Parameter $P(j)$ erforderlich, die auch vorgegeben sein müssen.

Bei der analytischen Vorgehensweise ist – im Gegensatz zur oben dargestellten konstruktiven Interpretation – die Parametrisierung des Modells unbekannt und soll mit Hilfe einer vorliegenden Bibliographie geschätzt werden. Das Kriterium mit dem aus der Menge der möglichen Modelle das beste ausgewählt wird, ist die Wahrscheinlichkeit, mit der die beobachtete Bibliographie in der Menge der möglichen, wie oben konstruierten Bibliographien auftaucht. Die Modellparameter sollen so gewählt werden, dass diese Größe maximiert wird.

Es muss also die Wahrscheinlichkeit bestimmt werden, mit der eine bestimmte Bibliographie, die nach der beschriebenen Methode konstruiert wird, auftritt.

Diese Wahrscheinlichkeit ergibt sich als das Produkt der Wahrscheinlichkeiten $P(i \wedge j)$, für die tatsächlich in der Bibliographie beobachteten Paare. Wegen der Kopplung der Referenzen an die Dokumente über die Themen

sind diese Wahrscheinlichkeiten im Allgemeinen unterschiedlich.

$$P(B) = \prod_{(j,i) \in Y_B} P(i \wedge j) = \prod_{i,j}^{n,m} P(i \wedge j)^{x_{ij}} \quad (2.9)$$

Die Wahrscheinlichkeit $P(i \wedge j)$ für die gleichzeitige Beobachtung von Dokument j und Referenz i lässt sich auf die bedingte Wahrscheinlichkeit zurück führen, mit der die Referenzen bei gegebenen Dokumenten auftreten.

$$P(i \wedge j) = P(j)P(i|j) \quad (2.10)$$

Nun kommen die Themen ins Spiel. Sie werden quasi zwischen Dokumente und Referenzen geschoben.

$$P(i|j) = \sum_{k=1}^K P(i|k)P(k|j) \quad (2.11)$$

Dabei wird - wie bereits bei der Beschreibung der generativen Modellinterpretation unausgesprochen unterstellt - bedingte stochastische Unabhängigkeit vorausgesetzt. Dies bedeutet, dass die Kopplung von Referenzen und Dokumenten alleine über die Themen vermittelt wird, darüber hinaus wird Unabhängigkeit gefordert (Fahrmeir, Hamerle und Tutz 1996).

Für $P(B)$ ergibt sich also insgesamt

$$P(B) = \prod_{i,j}^{n,m} P(i \wedge j)^{x_{ij}} = \prod_{i,j}^{n,m} [P(j) \sum_{k=1}^K P(i|k)P(k|j)]^{x_{ij}} \quad (2.12)$$

Vor der eigentlichen Analyse sind die bedingten Wahrscheinlichkeiten $P(i|k)$ und $P(k|j)$ nicht bekannt. Dasselbe gilt für $P(j)$. Ziel der Analyse ist es, diese Parameter zu schätzen. Als Kriterium dafür dient die Maximierung von $P(B)$. $P(B)$ ist die Wahrscheinlichkeit, dass die vorliegende Bibliographie (unter allen möglichen Bibliographien) tatsächlich auftritt. In der Sprache der LSA werden alle denkbaren alternativen Bibliographien durch bestimmte Matrizen X repräsentiert, nämlich diejenigen mit derselben Anzahl von Einträgen mit Wert 1. Bei bekannten Modellparametern lassen sich diese Matrizen über die korrespondierende Bibliographie und Gleichung (2.12) als mehr oder weniger wahrscheinlich bewerten.

Die Tatsache, dass die Parameter $P(j)$ als Faktoren abgespalten werden können, lässt erahnen, dass die Optimierung der beiden Matrizen mit den bedingten Wahrscheinlichkeiten völlig unabhängig von der Optimierung der Dokumentwahrscheinlichkeiten erfolgen kann. Insbesondere kann diese zweite Optimierung ganz unterbleiben, wenn man nicht am Ergebnis interessiert ist.

Insgesamt ist noch zu beachten, dass die Themenwahl insofern als vollständig betrachtet wird, als die Nebenbedingungen

$$\sum_{k=1}^K P(k|j) = 1 \quad (2.13)$$

$$\sum_{i=1}^n P(i|k) = 1 \quad (2.14)$$

erfüllt sein müssen. D.h. insbesondere, dass alle Themen so auf jedes Dokument aufgeteilt werden, dass sie in der Summe 1 ergeben. Ganz analog werden alle Referenzen anteilig jedem der Themen zugeordnet.

Die Anzahl der Themen muss vorgegeben werden und stellt einen zusätzlichen Parameter des Modells dar, der einer Optimierung nach geeigneten Kriterien unterworfen werden kann. Vor diesem Hintergrund besteht der Kern des Algorithmus in einem Maximum-Likelihood Verfahren.

Neben der hierarchischen Verwendung der Themen in der Darstellung der bedingten Wahrscheinlichkeit $P(i|j) = \sum_{k=1}^K P(i|k)P(k|j)$ ist auch eine völlig symmetrische Herleitung basierend auf $P(i \wedge j) = \sum_{k=1}^K P(k)P(i|k)P(j|k)$ möglich (Hofmann 1999b). Diese Variante wird hier allerdings nicht vorgestellt. Beachten sollte man allerdings, dass sie eine andere Parametrisierung des Modells verwendet.

2.3.2 Algorithmus

Typischerweise wird nicht die Gesamtwahrscheinlichkeit selbst maximiert, sondern deren Logarithmus. Wegen der Monotonie der Logarithmusfunktion wird die Lage des Maximums dabei nicht verändert. So entledigt man sich der Produkte und des Exponenten in Gleichung 2.12).

$$\log(P(B)) = \sum_{j=1, i=1}^{m, n} x_{ij} \log \sum_{k=1}^K P(i|k)P(k|j) + \sum_{j=1, i=1}^{m, n} x_{ij} \log P(j) \quad (2.15)$$

Die Standardvorgehensweise zur Maximierung der Log-Likelihood Funktion besteht darin, die Nebenbedingungen durch Lagrange-Multiplikatoren zu berücksichtigen, die Zielfunktionen abzuleiten, die Ableitungen gleich Null zu setzen und die entstandenen Gleichungssysteme per Fixpunktiteration zu lösen. Nach (Kaban 2004) oder (Chien, Wu und Wu 2005) erhält man dabei folgenden Algorithmus, wenn man wie oben für die konstruktive Beschreibung des Modells die asymmetrische Formulierung der PLSA zu Grunde legt:

$$P(i|k)' = P(i|k) \sum_{j=1}^m \frac{x_{ij}}{\sum_{k'=1}^K P(i|k')P(k'|j)} P(k|j) \quad (2.16)$$

$$P(i|k)'' = \frac{P(i|k)'}{\sum_{i'=1}^n P(i'|k)'} \quad (2.17)$$

$$P(k|j)' = P(k|j) \sum_{i=1}^n \frac{x_{ij}}{\sum_{k'=1}^K P(i|k')P(k'|j)} P(j|k) \quad (2.18)$$

$$P(k|j)'' = \frac{P(k|j)'}{\sum_{k'=1}^K P(k'|j)'} \quad (2.19)$$

Für alle i , j und k muss solange iteriert werden, bis Konvergenz erreicht ist. Im Nenner der Terme könnte der Wert 0 vorkommen. Hier kann man sich durch das Aufaddieren eines kleinen Wertes behelfen. Die Iteration muss außerdem initialisiert werden. Dazu können zwei normierte Zufallsmatrizen verwendet werden. Der Algorithmus garantiert nur für die Konvergenz in ein lokales Maximum. Die bisherige Praxis lässt allerdings hoffen, dass fast immer das globale Maximum gefunden wird (Hofmann 1999a), andernfalls kann z.B. nach wiederholter Anwendung des Iterationsverfahrens aus der Menge der lokalen Maxima das globale heraus gesucht werden.

Alternativ zur asymmetrischen Ableitung des Algorithmus kann man auch von der symmetrischen Formulierung des Problems ausgehen und erhält dann einen anderen Algorithmus für die alternative Parametrisierung (Hofmann 1999b).

2.3.3 Diversitätsmaß

Die Bedeutung eines Themas innerhalb der Bibliographie lässt sich nun bestimmen als

$$p_k = \frac{\sum_{j=1}^m P(k|j)}{\sum_{j=1}^m \sum_{k'=1}^K P(k'|j)} \quad (2.20)$$

Daraus ergibt sich als Diversitätsmaß wie in Gleichung (1.1)

$$H = - \sum_{k=1}^K p_k \log p_k. \quad (2.21)$$

2.4 Implementierung

Alle Berechnungen wurden in der kostenfrei verfügbaren Statistiksprache *R* implementiert (R Development Core Team 2008). Für die SVD konnte das Paket *corpco* eingebunden werden. Die PLSA musste elementar implementiert werden. Die Ausführung der Programme erfolgte im Batchbetrieb auf einem Rechner an der HU-Berlin. Dazu wurden die Standard-UNIX Programme *SSH* und *NOHUP* eingesetzt.

Kapitel 3

Ergebnisse

3.1 LSA

Um sich mit den grundlegenden Eigenschaften der LSA vertraut zu machen, bietet es sich an, einige einfache numerische Beispiele zu studieren.

3.1.1 Grundfragen

Dabei sollen die folgenden Fragen beantwortet werden: Wie wird durch die LSA mit anschließender Diversitätsberechnung die Vielfalt innerhalb einer völlig heterogenen, völlig homogenen und verschiedenen zufälligen Bibliographien bewertet? In denjenigen Fällen, in denen die Antwort aus der Theorie eindeutig vorhersehbar ist, dient die Bewertung lediglich dazu zu bestätigen, dass die Implementierung korrekt ausgeführt wurde.

Bewertung einer völlig heterogenen Bibliographie

Zunächst wird eine völlig heterogene Bibliographie betrachtet, die nur aus linear unabhängigen Dokumentvektoren besteht, im folgenden Beispiel aus drei Dokumenten mit je drei unterschiedlichen Referenzen. Mit X wird wie in Abschnitt 2.2 die Referenz-Dokument-Matrix bezeichnet, U , Λ und V sind

die Matrizen der SVD-Zerlegung¹.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad U = \begin{pmatrix} -0.577 & 0.000 & 0.000 \\ -0.577 & 0.000 & 0.000 \\ -0.577 & 0.000 & 0.000 \\ 0.000 & 0.000 & -0.577 \\ 0.000 & 0.000 & -0.577 \\ 0.000 & 0.000 & -0.577 \\ 0.000 & -0.577 & 0.000 \\ 0.000 & -0.577 & 0.000 \\ 0.000 & -0.577 & 0.000 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 1.73 & 0.00 & 0.00 \\ 0.00 & 1.73 & 0.00 \\ 0.00 & 0.00 & 1.73 \end{pmatrix} \quad V = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

$$H = 1.58 \quad H_{max} = 1.58$$

Alle Eigenwerte sind identisch. Wie zu erwarten wird in diesem Fall maximale Diversität $H = H_{max}$ erreicht. Die Transformation von X nach U besteht (neben Umordnung) lediglich in der Normierung der Dokumentvektoren. Die drei Einträge von V , die ungleich 0 sind, stellen die Koordinaten der Dokumente im transformierten Raum dar.

Bewertung einer völlig homogenen Bibliographie

Das entgegengesetzte Extrem erhält man bei einer völlig homogenen Bibliographie, die nur aus identischen Dokumenten besteht. Hier werden drei Dokumente mit drei identischen Referenzen gewählt.

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad U = \begin{pmatrix} -5.77e-01 & -8.16e-01 & 5.48e-17 \\ -5.77e-01 & 4.08e-01 & -7.07e-01 \\ -5.77e-01 & 4.08e-01 & 7.07e-01 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \end{pmatrix}$$

¹In der Darstellung der Matrizen wird immer die wissenschaftliche Zahlenschreibweise verwendet. Werte werden nicht gerundet, sondern mgl. so präsentiert wie berechnet.

$$\Lambda = \begin{pmatrix} 3.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 7.45e-17 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 1.46e-33 \end{pmatrix} V = \begin{pmatrix} -0.577 & -0.816 & 0.000 \\ -0.577 & 0.408 & 0.707 \\ -0.577 & 0.408 & -0.707 \end{pmatrix}$$

$$H = 6.8e-32 \quad H_{max} = 1.58$$

Nachdem die Dokumente linear abhängig sind, ist der Rang der Referenz-Dokument-Matrix gleich 1 und der transformierte Raum nur eindimensional. Der Algorithmus liefert wegen der begrenzten Rechengenauigkeit allerdings verschwindend kleine Reste in den anderen Dimensionen, so dass die resultierenden Matrizen in maximal möglicher Größe dargestellt werden müssen. Dies hat aber keinen Effekt auf das Gesamtergebnis. Sowohl bei U als auch bei V ist nur die erste Spalte relevant.

Wie erwartet erhält man hier für das Diversitätsmaß einen verschwindend geringen Wert bei 0. Nur ein Eigenwert ist tatsächlich von 0 verschieden. Die Matrix der Eigenwerte lässt sich auf den Skalar größer 0 reduzieren.

Bewertung einer schwach gekoppelten Bibliographie

Konstruiert man eine schwach gekoppelte Bibliographie, in der ausgehend vom ersten Beispiel zwei weitere Referenzen hinzugefügt werden, erhält man z.B.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} U = \begin{pmatrix} 0.1432 & -0.3135 & -0.4971 \\ 0.1432 & -0.3135 & -0.4971 \\ 0.4649 & -0.4874 & -0.0985 \\ 0.3217 & -0.1740 & 0.3986 \\ 0.3217 & -0.1740 & 0.3986 \\ 0.5797 & 0.2169 & 0.1774 \\ 0.2580 & 0.3909 & -0.2212 \\ 0.2580 & 0.3909 & -0.2212 \\ 0.2580 & 0.3909 & -0.2212 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.29 & 0.00 & 0.00 \\ 0.00 & 1.89 & 0.00 \\ 0.00 & 0.00 & 1.48 \end{pmatrix} V = \begin{pmatrix} 0.328 & -0.591 & -0.737 \\ 0.737 & -0.328 & 0.591 \\ 0.591 & 0.737 & -0.328 \end{pmatrix}$$

$$H = 1.5; \quad H_{max} = 1.58$$

Die Diversität fällt dabei nur wenig unter den Maximalwert ab. Nachdem auch die empirischen Bibliographien auf schwach gekoppelten Referenzlisten beruhen, wird man bei realen Daten Diversitätswerte nahe am Maximum erhalten.

Anders als im Fall linear unabhängiger Dokumente besitzen die Dokumente im (dreidimensionalen) transformierten Raum nicht vernachlässigbare Koordinatenwerte in alle Richtungen.

Nullzeile hinzufügen

Das Anfügen einer einzelnen Nullzeile an die Referenz-Dokument-Matrix verändert das Ergebnis erwartungsgemäß nicht.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0.1432 & -0.3135 & -0.4971 \\ 0.1432 & -0.3135 & -0.4971 \\ 0.4649 & -0.4874 & -0.0985 \\ 0.3217 & -0.1740 & 0.3986 \\ 0.3217 & -0.1740 & 0.3986 \\ 0.5797 & 0.2169 & 0.1774 \\ 0.2580 & 0.3909 & -0.2212 \\ 0.2580 & 0.3909 & -0.2212 \\ 0.2580 & 0.3909 & -0.2212 \\ 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.29 & 0.00 & 0.00 \\ 0.00 & 1.89 & 0.00 \\ 0.00 & 0.00 & 1.48 \end{pmatrix} \quad V = \begin{pmatrix} 0.328 & -0.591 & -0.737 \\ 0.737 & -0.328 & 0.591 \\ 0.591 & 0.737 & -0.328 \end{pmatrix}$$

$$H = 1.5 H_{max} = 1.58$$

Nullspalte hinzufügen

Jede zusätzliche Spalte in der Matrix (d.h. jedes weitere Dokument) muss die maximale Diversität erhöhen, wenn man davon ausgeht, dass die maximal mögliche Anzahl nicht verschwindender Eigenwerte der Anzahl der Dokumente entspricht.

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0.1432 & -0.3135 & -0.4971 & -0.1002 \\ 0.1432 & -0.3135 & -0.4971 & 0.4414 \\ 0.4649 & -0.4874 & -0.0985 & -0.3412 \\ 0.3217 & -0.1740 & 0.3986 & 0.7604 \\ 0.3217 & -0.1740 & 0.3986 & -0.2396 \\ 0.5797 & 0.2169 & 0.1774 & -0.1797 \\ 0.2580 & 0.3909 & -0.2212 & 0.0599 \\ 0.2580 & 0.3909 & -0.2212 & 0.0599 \\ 0.2580 & 0.3909 & -0.2212 & 0.0599 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.29 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.89 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.48 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00e + 00 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.328 & -0.591 & -0.737 & 0.000 \\ 0.737 & -0.328 & 0.591 & 0.000 \\ 0.591 & 0.737 & -0.328 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

$$H = 1.5; H_{max} = 2$$

Durch eine Nullspalte (d.h. ein Dokument ohne Referenzen) ändern sich die Eigenwerte nicht. Es entsteht allerdings ein weiterer Wert sehr nahe an 0.²

Identische Spalte hinzufügen

Nun fügen wir der Bibliographie eine Kopie eines bereits vorhandenen Dokuments hinzu (Verdopplung einer bereits vorhandenen Spalte).

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad U = \begin{pmatrix} -0.0197 & 0.2889 & 0.5310 & -0.6616 \\ -0.0197 & 0.2889 & 0.5310 & 0.0569 \\ -0.1272 & 0.6528 & 0.1454 & 0.6047 \\ -0.1076 & 0.3638 & -0.3857 & -0.1292 \\ -0.1076 & 0.3638 & -0.3857 & -0.1292 \\ -0.5683 & 0.1693 & -0.2510 & -0.3463 \\ -0.4608 & -0.1945 & 0.1347 & 0.1154 \\ -0.4608 & -0.1945 & 0.1347 & 0.1154 \\ -0.4608 & -0.1945 & 0.1347 & 0.1154 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.91e+00 & 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 2.06e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 1.51e+00 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 & 1.29e-16 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.0572 & 0.5963 & 0.8007 & 0.0000 \\ -0.3130 & 0.7509 & -0.5815 & 0.0000 \\ -0.6704 & -0.2007 & 0.1016 & 0.7071 \\ -0.6704 & -0.2007 & 0.1016 & -0.7071 \end{pmatrix}$$

$$H = 1.39; H_{max} = 2$$

Dabei nähern sich die nicht verschwindenden Eigenwerte an und der Absolutwert der Diversität sinkt. Tendenziell nähert sich die Struktur der Matrix

²Ein Eigenwert 0 resultiert anteilig in $p = 0$. Dies würde eigentlich zu einer Singularität bei der Berechnung der Entropie führen, wenn der Logarithmus gebildet wird. Setzt man die Berechnung allerdings bei 0 stetig fort ($\lim(x \log x) \rightarrow 0$ für $x \rightarrow 0$), verschwindet das Problem (siehe Abbildung 1.1). Numerisch tritt das Problem deswegen nicht auf, weil Werte, die hier als $0.00e+00$ erscheinen, im Arbeitsspeicher sehr kleinen positiven Werten entsprechen, die mit der gewählten Stellenanzahl nicht mehr darstellbar sind.

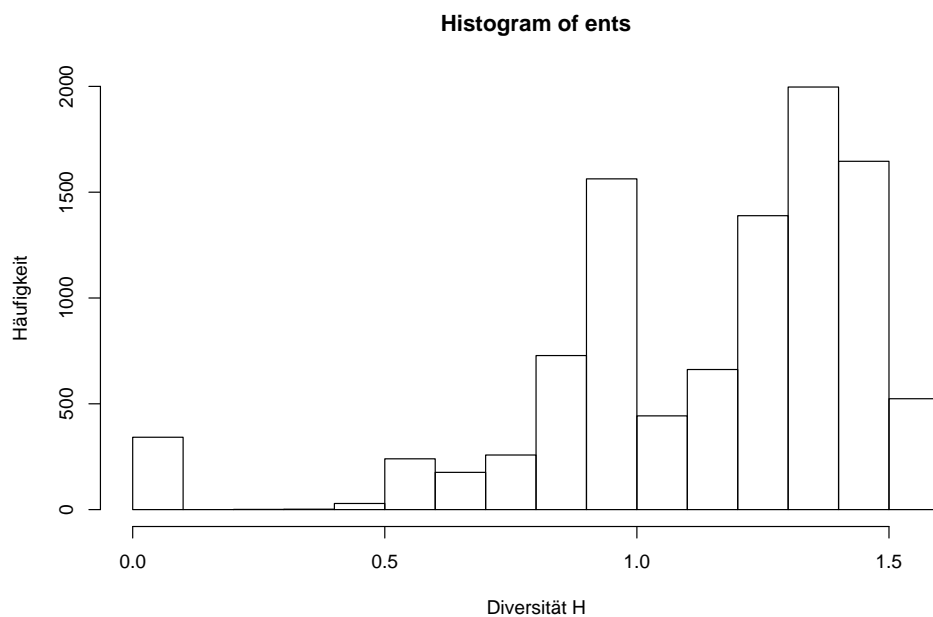


Abbildung 3.1: Diversitätsverteilung für Zufallsmatrizen (LSA). Es werden 10000 Matrizen (14 Spalten, 50 Zeilen) erzeugt, in denen die einzelnen Elemente mit Wahrscheinlichkeit 0.2 gleich 1 sind.

durch identische Spalten der homogenen Bibliographie, die mit Diversität 0 bewertet wird. Wegen der erhöhten maximalen Diversität sinkt die relative Diversität umso mehr.

Matrix mit Zufallselementen

Als nächstes soll eine Zufallsmatrix betrachtet werden, deren Elemente unabhängig voneinander mit konstanter Wahrscheinlichkeit (hier 0.2) realisiert werden. Die folgenden Matrizen sind ein Beispiel für eine mögliche Realisation eines solchen Systems. Die Diversitätswerte werden allerdings als Mit-

telwerte für mehrere Realisationen angegeben.

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0.0 & 0.0 & 1.0 \\ -0.5 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ -0.5 & 0.0 & 0.0 \\ -0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ -0.5 & 0.0 & 0.0 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad V = \begin{pmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\bar{H} = 1.14; \sigma_H = 0.33; H_{max} = 1.58$$

Die mittlere Diversität \bar{H} liegt (bei 10000 Wiederholungen) deutlich unter der maximalen Diversität. Die Standardabweichung von 0.33 liefert bei 10000 Wiederholungen einen Standardfehler für die Schätzung des Mittelwerts von nur 0.0033. Eine genauere Analyse zeigt allerdings, dass die Häufigkeitsverteilung der Diversitätswerte, wie sie im Experiment durch unterschiedliche Zufallsmatrizen realisiert werden, nicht unimodal ist (siehe Abbildung 3.1). Die diskrete Struktur der Matrix, in der nur 0 und 1 Einträge erlaubt sind, scheint sich auf die erlaubten Diversitätswerte zu übertragen. Auch diese zeigen ein spezifisches Muster, in dem ganze Wertebereiche ausgespart bleiben.

Zufallsspalten mit zwei zufälligen Einträgen pro Spalte

Fordert man für das nächste Beispiel zusätzlich genau zwei zufällig positionierte Einträge pro Spalte (d.h. zwei Referenzen pro Dokument), wird beim entsprechenden Zufallsexperiment (10000 Wiederholungen) im Vergleich zu oben nur noch eine Teilmenge der Dokumente realisiert. Die folgenden Ma-

trizen sind wieder nur exemplarisch zu verstehen.

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 7.85e-17 & 0.00e+00 & 1.36e-16 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ 0.00e+00 & -7.07e-01 & 0.00e+00 \\ 0.00e+00 & -7.07e-01 & 0.00e+00 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ -8.16e-01 & 0.00e+00 & -1.61e-16 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ -4.08e-01 & 0.00e+00 & -7.07e-01 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ -4.08e-01 & 0.00e+00 & 7.07e-01 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 1.73 & 0.00 & 0.00 \\ 0.00 & 1.41 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix} \quad V = \begin{pmatrix} -0.707 & 0.000 & 0.707 \\ -0.707 & 0.000 & -0.707 \\ 0.000 & -1.000 & 0.000 \end{pmatrix}$$

$$\bar{H} = 1.41; \sigma_H = 0.178; H_{max} = 1.58$$

Gegenüber den vorherigen Zufallsmatrizen nimmt die Vielfalt zu. Dies ist auf den ersten Blick überraschen, wenn man die scheinbare Einschränkung der Heterogenität durch die Spaltenbedingung bedenkt. Die Erklärung findet man durch einen Blick auf die Verteilung der Diversitätswerte (siehe Abbildung 3.2). Obwohl auch hier der minimale Diversitätswert von 0 realisiert werden kann, wenn drei identische Spalten vorliegen, tritt dieser Fall seltener auf als zuvor. Zudem gibt es weniger Möglichkeiten für andere Konstellationen mit kleiner Entropie.

Zufallsspalten mit drei zufälligen Einträgen pro Spalte

Es folgt eine weitere Realisation des Zufallsexperiments, diesmal allerdings mit genau drei Referenzen pro Dokument.

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad U = \begin{pmatrix} -1.32e-16 & -9.00e-18 & 2.74e-16 \\ -7.17e-01 & -1.58e-01 & -3.72e-17 \\ -5.25e-01 & 4.32e-01 & 1.61e-16 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ -2.62e-01 & 2.16e-01 & -7.07e-01 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ -1.92e-01 & -5.90e-01 & -6.19e-17 \\ 0.00e+00 & 0.00e+00 & 0.00e+00 \\ -1.92e-01 & -5.90e-01 & -6.19e-17 \\ -2.62e-01 & 2.16e-01 & 7.07e-01 \end{pmatrix}$$

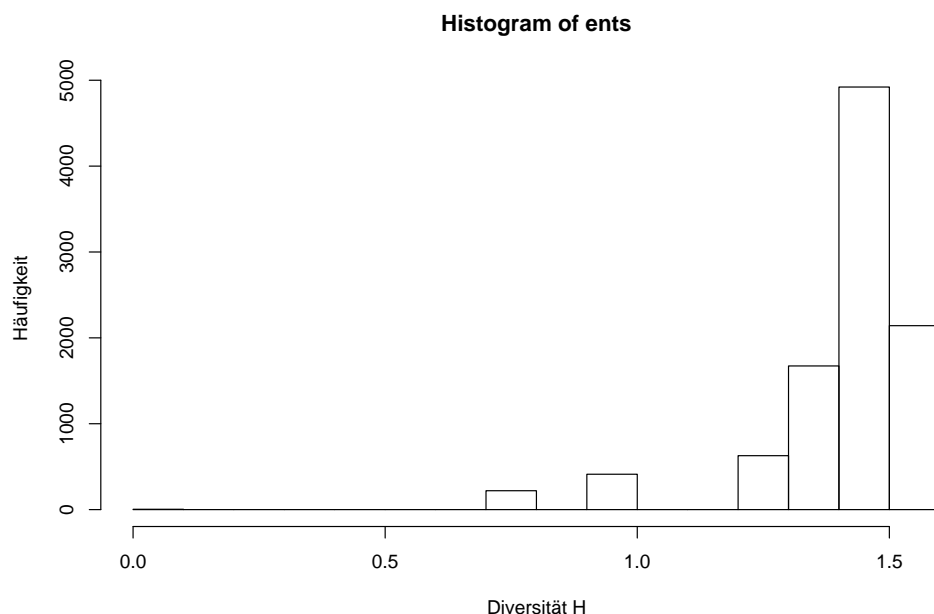


Abbildung 3.2: Diversitätsverteilung für Zufallsspaltenmatrizen (LSA). Es werden 10000 Matrizen (14 Spalten, 50 Zeilen) erzeugt, in denen jede Spalte genau zwei 1-Elemente besitzt. Diese werden zufällig auf die möglichen Plätze verteilt.

$$\Lambda = \begin{pmatrix} 2.39 & 0.00 & 0.00 \\ 0.00 & 1.51 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix} \quad V = \begin{pmatrix} -6.28e-01 & 3.25e-01 & -7.07e-01 \\ -4.60e-01 & -8.88e-01 & -1.12e-17 \\ -6.28e-01 & 3.25e-01 & 7.07e-01 \end{pmatrix}$$

$$\bar{H} = 1.37; \sigma_H = 0.158; H_{max} = 1.58$$

Im Vergleich zum Szenario mit nur zwei Einträgen pro Spalte sinkt die Diversität. Diese Tendenz setzt sich fort, wenn die Anzahl von vorgegebenen Referenzen pro Spalte weiter erhöht wird, weil die Dokumente dann zwangsläufig in den zitierten Referenzen ähnlicher werden. Dies gilt es im Kopf zu behalten, wenn man später die Diversitätswerte nahe am theoretischen Maximum für die realen Beobachtungsdaten interpretieren will.

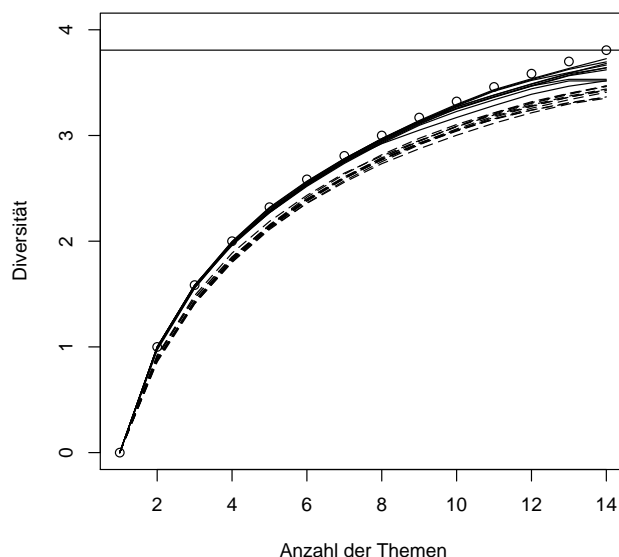


Abbildung 3.3: Diversität in Abhängigkeit der Anzahl berücksichtigter Eigenwerte (=Themen) für zwei Typen von Zufallsmatrizen ($m = 14$; $n = 50$); Durchgezogene Linien: 10 Wiederholungen mit genau zwei zufälligen Einträgen pro Spalte; gestrichelte Linien: 10 Wiederholungen mit genau zehn zufälligen Einträgen pro Spalte.

Anzahl der Eigenwerte

Bei der Anwendung der LSA im Bereich des Information Retrievals werden typischerweise nur die größten Eigenwerte berücksichtigt³. Die Zahl der Dimensionen des Vektorraums wird künstlich verringert, indem die sehr kleinen Eigenwerte und die zugehörigen Eigenvektoren weggelassen werden. Eine derartige Dimensionsreduzierung macht die extrahierten Themen für praktische Zwecke übersichtlicher. Für die Diversität könnte dies allerdings zu einem verzerrten Bild führen, wenn systematisch die kleinen Themen ignoriert werden würden.

Abbildungen 3.3 und 3.4 zeigen die Abhängigkeit der Diversität von der Anzahl der berücksichtigten Eigenwerte. Zwei Typen von Zufallsmatrizen

³Für die Wahl des geeigneten Eigenwertbereichs gibt es verschiedene Heuristiken (Bortz 2005)

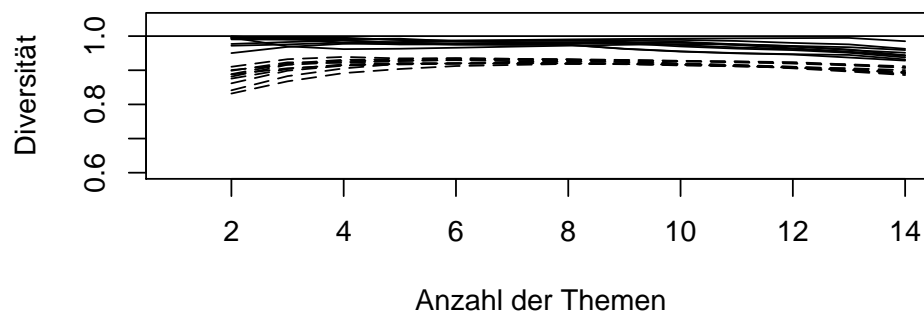


Abbildung 3.4: Relative Diversität in Abhängigkeit der Anzahl berücksichtigter Eigenwerte (=Themen) für zwei Typen von Zufallsmatrizen ($m = 14$; $n = 50$); Durchgezogene Linien: 10 Wiederholungen mit genau zwei zufälligen Einträgen pro Spalte; gestrichelte Linien: 10 Wiederholungen mit genau zehn zufälligen Einträgen pro Spalte.

werden in je zehn Realisationen untersucht. Bei jeder Themenzahl lassen sich die beiden Typen gut unterscheiden. Innerhalb eines Typs gibt es allerdings Überschneidungen im Diversitätsverlauf, d.h. je nach Themenzahl wird mal die eine Matrix, mal die andere als diverser bewertet. Bei relativer Betrachtung scheint dann größtmögliche Trennschärfe erreicht zu werden, wenn die Zahl der Themen an den beiden Enden des zulässigen Bereichs liegt (siehe Abbildung 3.4).

3.1.2 Szientometrie

Die LSA soll nun auf die Datensammlung zur Szientometrie angewendet werden. Untersucht werden 21 Jahrgänge (1986–2006) von fünf szientometrischen Zeitschriften (siehe Tabelle 2.1). Aus jedem Jahrgang werden 50 zufällige Stichproben gleichen Stichprobenumfangs (100 Artikel) ausgewertet. Es werden nur Veröffentlichungen vom Typ „Article“ verwendet. Durch die Standardisierung der ausgewerteten Artikelmenen wird sicher gestellt, dass mögliche Effekte nicht durch eine systematische Variation in den Artikelzahlen hervorgerufen werden (siehe Abbildung 3.5).

Eigenwertverteilung

Bei der Datenauswertung liegen nur äußerst selten lineare Abhängigkeiten innerhalb der Stichproben vor. Dementsprechend erhält man pro Stichprobe, die mit LSA untersucht wird, durchweg 100 Eigenwerte. Abbildungen 3.6 und 3.7 zeigen je drei zufällige Beispiele von (der Größe nach geordneten und auf

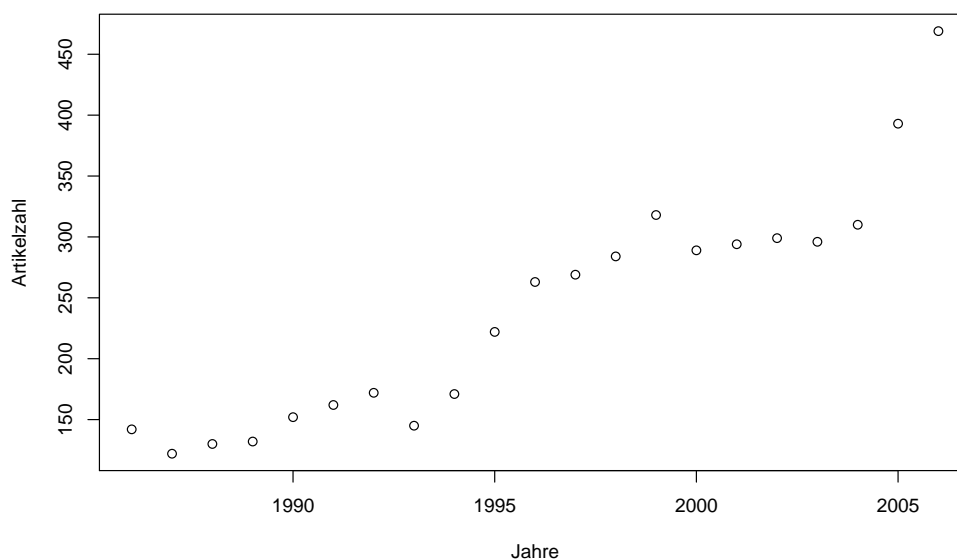


Abbildung 3.5: Artikelzahl innerhalb der Szientometrie-Jahrgänge.

den größten Wert normierten) Eigenwerten (Screeplots⁴), wie sie aus der LSA der Szientometriedaten resultieren. Die erste Abbildung repräsentiert die Situation zu Beginn des analysierten Zeitraums (1986), die zweite Abbildung das Ende (2006). Bereits auf den ersten Blick erkennt man deutlich, dass sich der Bereich der großen Eigenwerte im Laufe der Zeit verbreitert hat. Wenn dies kein zufälliger Effekt in den ausgewählten Beispielen ist, muss sich dies in einer Zunahme des Diversitätsmaßes niederschlagen.

Logarithmische Transformationen der Eigenwertdiagramme zeigen, dass die Verteilung der Eigenwerte keinem Potenzgesetz folgt und damit nicht skalenfrei ist.

Diversitätsentwicklung

Für jeden Jahrgang werden Mittelwert und Standardabweichung aus den 50 Diversitätsberechnungen bestimmt und dargestellt (Abbildung 3.8). Bei relativ kleiner Effektgröße findet man im Zeitraum von 1986 bis 2006 eine deutliche Tendenz zu höheren Diversitätswerten (von 94% bis hin zu 96% des Maximalwertes), die die Variabilität innerhalb der Stichproben übersteigt. Gleichzeitig steigt auch die mittlere Anzahl von Referenzen pro Artikel. Bisher wurde nur die Anzahl von Artikeln pro Stichprobe auf 100 normiert, dagegen nicht die Anzahl von Zitationen. Es liegt nahe zu vermuten, dass diese statistische Eigenschaft für den Trend verantwortlich ist.

Eine zusätzliche Modellrechnung (Abbildung 3.9) mit konstanter mittlerer Referenzanzahl zeigt jedoch, dass der Effekt erhalten bleibt, auch wenn so-

⁴ *scree* engl. für Schutthalde; gemeint ist der rechte Bereich mit den kleinen Eigenwerten, die bei einer Faktorenanalyse weggelassen werden (Bortz 2005).

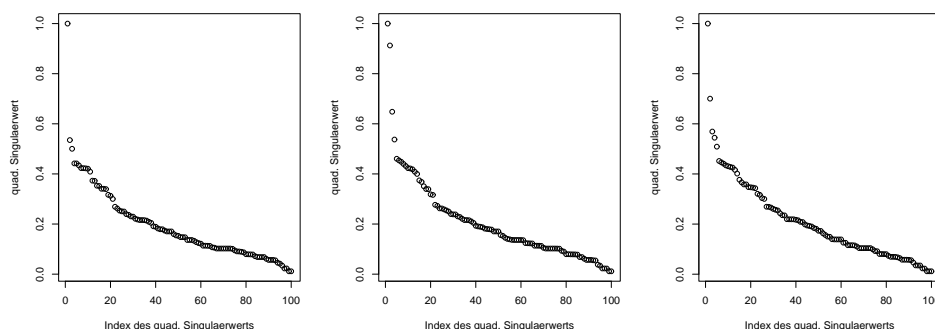


Abbildung 3.6: Screeplot zum Szientometrie-Jahrgang 1986: Diagramm der (normierten) Eigenwerte in abnehmender relativer Größe.

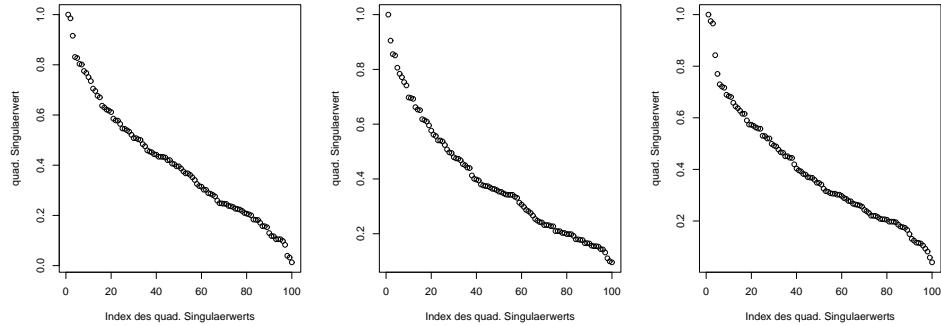


Abbildung 3.7: Screeplot zum Szientometrie-Jahrgang 2006: Diagramm der (normierten) Eigenwerte in abnehmender relativer Größe.

lange Referenzen zufällig gelöscht werden, bis den Artikeln im Mittel genau 15 Quellen bleiben.

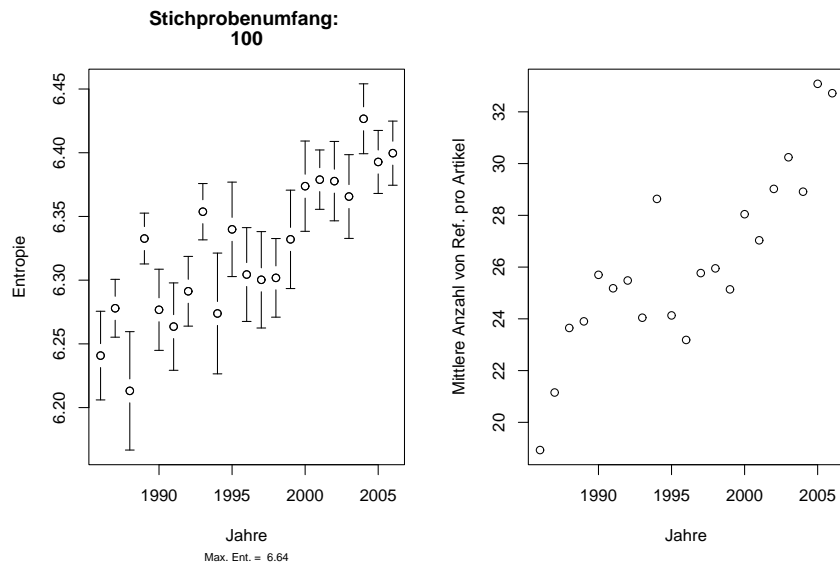


Abbildung 3.8: Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie (LSA). Aus dem Dokumentenpool werden pro Jahrgang 50 Stichproben mit je 100 Artikeln gezogen und deren Diversität berechnet. Die Fehlerbalken entsprechen der Standardabweichung in diesen 50 Wiederholungen (links). Rechts: Mittlere Anzahl von Referenzen pro Artikel.

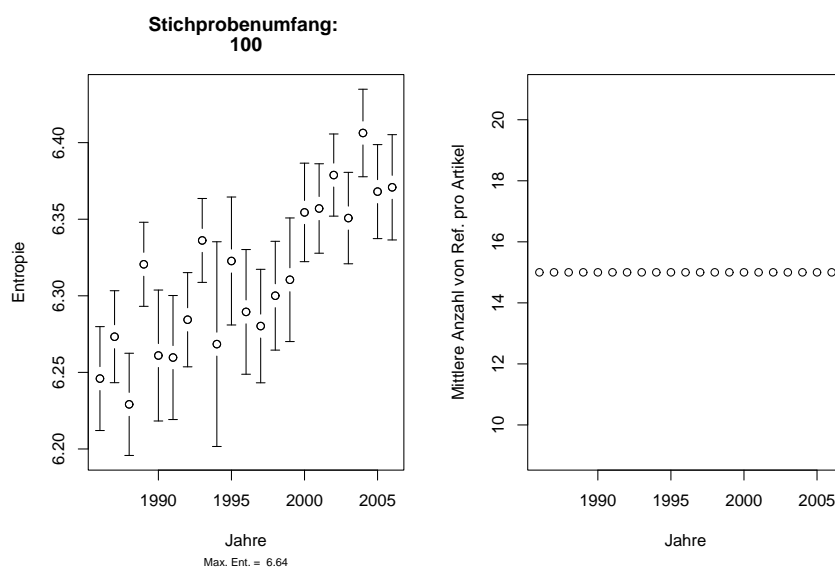


Abbildung 3.9: Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie bei konstanter mittlerer Referenzzahl pro Dokument (LSA). Aus dem Dokumentenpool werden pro Jahrgang 50 Stichproben mit je 100 Artikeln gezogen und deren Diversität berechnet. Dabei werden solange zufällig Einträge aus der Referenz-Dokument-Matrix gelöscht, bis im Mittel 15 Referenzen pro Dokument verbleiben. Die Fehlerbalken entsprechen der Standardabweichung in den 50 Wiederholungen (links). Rechts: Mittlere Anzahl von Referenzen pro Dokument.

3.1.3 Elektrochemie

Für die Auswertung der Zeitschriften aus dem Bereich Elektrochemie kann man ganz analog vorgehen. Auch hier wird der Zeitraum von 1986 bis 2006 untersucht. Der Auswertung liegen 14 elektrochemische Zeitschriften zu Grunde (siehe Tabelle 2.2). Aus jedem Jahrgang werden 50 zufällige Stichproben gleichen Stichprobenumfangs (500 Veröffentlichungen vom Typ „Article“) ausgewertet. Die Standardisierung des Stichprobenumfangs stellt wieder sicher, dass mögliche Effekte nicht durch eine systematische Variation in den Artikelzahlen hervorgerufen werden (siehe Abbildung 3.10).

Abbildung 3.10 zeigt, dass der Stichprobenumfang noch deutlich weiter gesteigert werden könnte. Entsprechende Versuche (Stichprobenumfang 1000) haben allerdings stets zu Programmabstürzen wegen Speichermangel geführt. Eine Vergrößerung des Arbeitsspeichers kann dieses Problem sicher beheben.

Eigenwertverteilung

Die Entwicklung der Eigenwertverteilung von 1986 bis 2006 ist mit der Situation in der Szientometrie vergleichbar. Wenige zu Beginn ausgeprägte Ei-

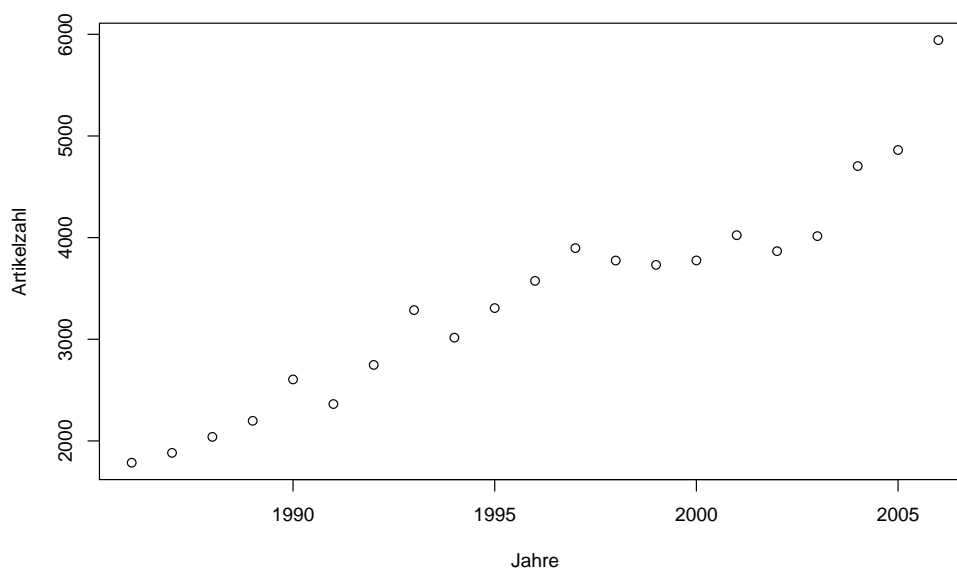


Abbildung 3.10: Artikelzahl innerhalb der Elektrochemie-Jahrgänge.

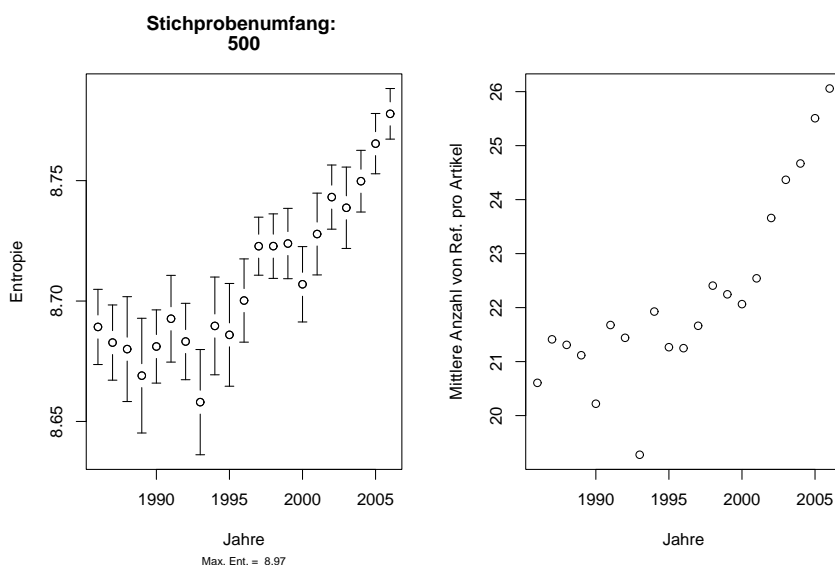


Abbildung 3.11: Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Elektrochemie (LSA). Aus dem Dokumentenpool werden pro Jahrgang 50 Stichproben mit je 500 Artikeln gezogen und deren Diversität berechnet. Die Fehlerbalken entsprechen der Standardabweichung in den 50 Wiederholungen (links). Rechts: Mittlere Anzahl von Referenzen pro Dokument.

genwerte werden im Laufe der Zeit durch eine breitere Struktur ersetzt (hier ohne Abbildungen).

Diversitätsentwicklung

Auch in der Elektrochemie ist ein klarer Trend hin zu größerer Diversität zu erkennen bei gleichzeitigem Anstieg der mittleren Referenzzahl pro Artikel (Abbildung 3.11). Die Werte steigen von 97% auf 98% verglichen mit dem erreichbaren Diversitätsmaximum. Der Anstieg setzt allerdings erst um 1995 ein. Fixiert man in einem weiteren Experiment die mittlere Referenzzahl pro Artikel bleibt die Tendenz erhalten (Abbildung 3.12).

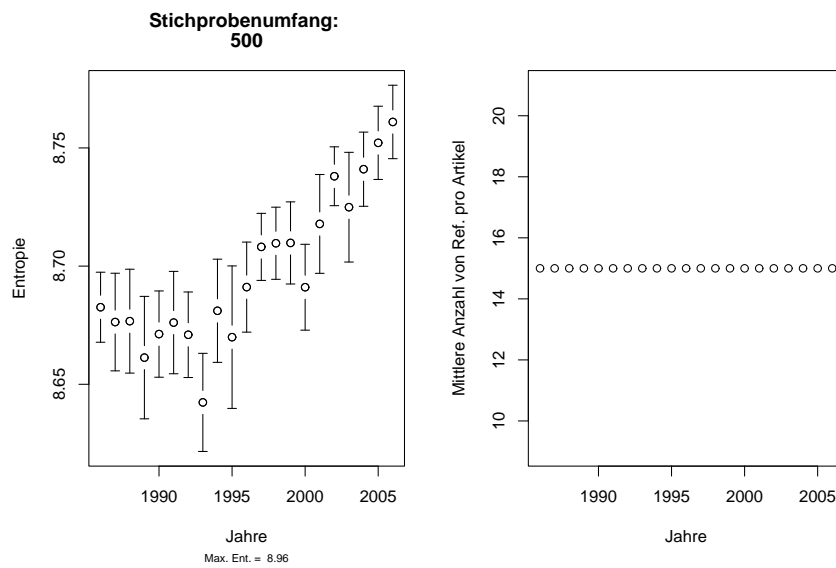


Abbildung 3.12: Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Elektrochemie bei konstanter mittlerer Referenzanzahl pro Dokument (LSA). Aus dem Dokumentenpool werden pro Jahrgang 50 Stichproben mit je 500 Artikeln gezogen und deren Diversität berechnet. Dabei werden solange zufällig Einträge aus der Referenz-Dokument-Matrix gelöscht, bis im Mittel 15 Referenzen pro Dokument verbleiben. Die Fehlerbalken entsprechen der Standardabweichung in den 50 Wiederholungen (links). Rechts: Mittlere Anzahl von Referenzen pro Dokument.

3.2 PLSA

Auch für die PLSA sollen zunächst einige numerische Beispiele untersucht werden, um mit den grundlegenden Eigenschaften der Methode vertraut zu werden.

3.2.1 Grundfragen

Wie im Abschnitt 3.1 zur LSA werden zuerst die folgenden einfachen Fragen anhand überschaubarer fiktiver Datensätze beantwortet: Wie wird durch die PLSA mit anschließender Diversitätsberechnung die Vielfalt innerhalb einer völlig heterogenen, völlig homogenen und verschiedenen zufälligen Bibliographien bewertet? In denjenigen Fällen, in denen die Antwort aus der Theorie eindeutig vorhersehbar ist, dient die Bewertung lediglich dazu, zu bestätigen, dass die Implementierung korrekt ausgeführt wurde.

Maximale Themenzahl

Ein entscheidender Unterschied zur LSA tritt sofort zu Tage, wenn man versucht ganz analog vorzugehen, und die Themenzahl K – wie zuvor die Anzahl der berücksichtigten Eigenwerte – maximal, also gleich der Anzahl der Dokumente m in der Bibliographie wählt. Bei $K = m$ Themen liefert die PLSA immer maximale Diversität $H = \log_2 K$. Dabei erreicht die Methode größtmögliche Trennschärfe bei der Themenextraktion, indem sie den Dokumenten alle möglichen Themen (bis auf Vertauschung) eineindeutig zuordnet.

Dies kann man verstehen, wenn man sich die konstruktive Interpretation der PLSA ins Gedächtnis ruft. Bei $K = m$ besteht die Möglichkeit, jedem Dokument vollständig und genau ein Thema zuzuordnen, und diesem Thema dann gleichmäßig verteilt alle Referenzen, die beim korrespondierenden Dokument gefunden werden. Damit werden alle $P(i|j) = \frac{1}{|\bar{x}_j|}^5$, also exakt gleich ihrer relativen Häufigkeit im fraglichen Dokument. Damit kann es bei der Suche nach maximaler Wahrscheinlichkeit kein besseres Modell als dieses geben (evtl. aber noch andere mit derselben Wahrscheinlichkeit).

Dieser Befund bedeutet, dass man bei der Verwendung von PLSA zur Untersuchung der zeitlichen Entwicklung von Diversität in Bibliographien immer gezwungen ist, eine Themenzahl unter dem Maximum vorzugeben. Es wird damit die Frage aufgeworfen, ob für die Diversitätsuntersuchung eine besonders günstige Vorgabe für die Themenzahl gefunden werden kann (siehe

⁵Mit der Vektornorm ist hier die Summe der Elemente gemeint, d.h. die Anzahl von Referenzen im Dokument.

dazu den Abschnitt Diskussion). Für die folgenden fiktiven Beispiele mit Dokumentenzahlen $m = 3$ oder $m = 4$ wird immer Themenzahl $K = 2$ ($< m$) gewählt.

Bewertung einer völlig heterogenen Bibliographie

Zunächst betrachten wir drei Dokumente mit jeweils drei unterschiedlichen Referenzen. Die Referenz-Dokument-Matrix heißt wie bisher X . Mit $P1$ und $P2$ werden die Parametermatrizen $P(i|k)$ und $P(k|j)$ der PLSA bezeichnet (siehe Abschnitt 2.3).

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 1.60e - 11 & 1.00e + 00 & 1.00e + 00 \\ 1.00e + 00 & 1.91e - 12 & 2.49e - 09 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 6.56e - 13 & 3.33e - 01 \\ 6.57e - 13 & 3.33e - 01 \\ 6.54e - 13 & 3.33e - 01 \\ 1.67e - 01 & 5.46e - 14 \\ 1.67e - 01 & 1.46e - 13 \\ 1.67e - 01 & 7.25e - 14 \\ 1.67e - 01 & 3.95e - 10 \\ 1.67e - 01 & 3.61e - 10 \\ 1.67e - 01 & 2.69e - 10 \end{pmatrix}$$

$$H = 0.918; H_{max} = 1; L_0 = -20.3$$

Die Matrix $P2$ ist wie gefordert in allen drei Spalten, also für alle drei Dokumente auf 1 normiert. Dies gilt nicht für die beiden Zeilen, also die beiden Themen. In der ersten (und ganz analog in der zweiten) Zeile lässt sich die Bedeutung von Thema 1 (bzw. Thema 2) ablesen. Thema 2 wird vollständig dem ersten Dokument zugeordnet, Thema 1 gleichermaßen den anderen beiden Dokumenten. Aus der relativen Bedeutung der Themen (2/3 bzw. 1/3)

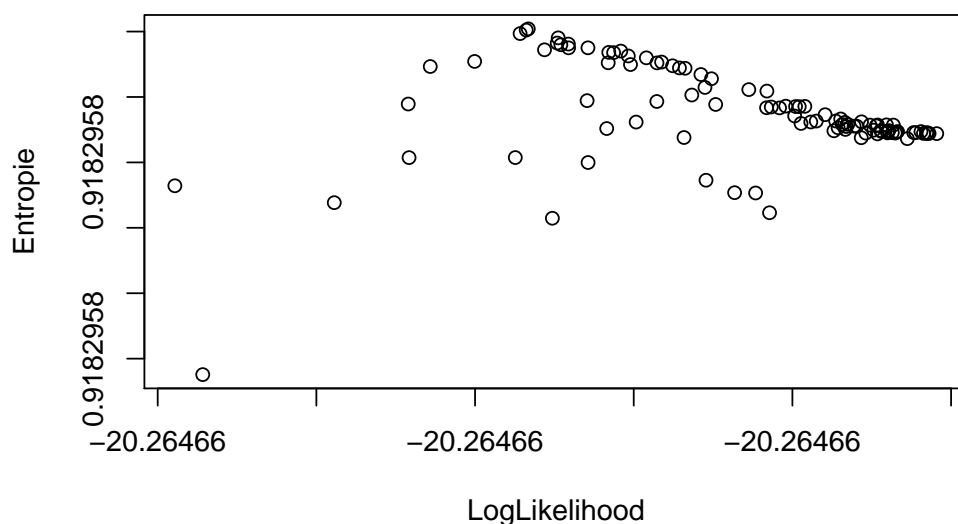


Abbildung 3.13: Entropie in Abhängigkeit von der Log-Likelihood bei 100 Realisationen des PLSA Algorithmus für die völlig heterogene Bibliographie. Die Variabilität ist sowohl in Entropie- als auch in Log-Likelihood-Richtung äußerst gering und liegt unterhalb der Genauigkeit der Achsenbeschriftung.

ergibt sich der Diversitätswert H . Die Größe L_0 bezeichnet das erreichte Maximum der Log-Likelihood-Funktion.

Bei eingeschränkter Themenzahl kann offensichtlich auch eine Bibliographie mit drei linear unabhängige Dokumente wie hier nicht mehr maximale Diversität erreichen, wenn die Maximierung der Beobachtungswahrscheinlichkeit gefordert wird. Zwei Dokumente werden vollständig dem einen, das verbleibende Dokument dem zweiten Thema zugeordnet. Dafür gibt es in diesem Fall sechs äquivalente Möglichkeiten (Thema 1: Dok1 oder Dok2 oder Dok3 oder Dok1+Dok2 oder Dok1+Dok3 oder Dok2+Dok3; Thema 2 komplementär), die sich beim Austauschen von Thema 1 und 2 auf 3 Möglichkeiten reduzieren. Welche davon der PLSA-Algorithmus liefert, hängt von den gewählten Anfangsbedingungen ab. Dies zeigt eine Wiederholung der Analyse mit variierten Anfangsbedingungen.

Man könnte evtl. vermuten, dass neben der stark asymmetrischen The-

menzuordnung, wie sie von der PLSA geliefert wird, auch andere Möglichkeiten, die Themen auf die Dokumente zu verteilen, eine kaum geringere Beobachtungswahrscheinlichkeit liefern. Für den folgenden naheliegenden Fall ist dies allerdings nicht richtig.

$$P2 = \begin{pmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{pmatrix}$$

$$L_0 = -29.3$$

Abbildung 3.13 zeigt, dass sich die Log-Likelihood nicht verändert, wenn man das Iterationsverfahren mit unterschiedlichen zufälligen Anfangsbedingungen wiederholt. Dasselbe gilt für die Diversität. Lediglich die Themenzuordnungen sind durch entsprechende Permutationen in den Spalten der Matrizen $P2$ zufällig vertauscht.

Dass die PLSA für die drei linear unabhängigen Dokumentvektoren nicht maximale Diversität liefert, ist insbesondere ein Effekt des kleinen Datensatzes. Die Dokumente aus heterogenen Dokumentmengen werden von der PLSA möglichst vollständig und gleichmäßig auf die Themen aufgeteilt. Bei drei Dokumenten und zwei Themen kann so eine Ungleichverteilung entstehen, die sofort verschwindet, wenn man einen vierten linear unabhängigen Dokumentvektor hinzufügt. Man erhält dann maximale Diversität $H = 1$.

Bewertung einer völlig homogenen Bibliographie

Als nächstes wird eine völlig homogene Bibliographie untersucht, die aus drei identischen Dokumenten mit je drei Referenzen besteht.

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 0.583 & 0.583 & 0.583 \\ 0.417 & 0.417 & 0.417 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 5.98e - 02 & 7.15e - 01 \\ 3.81e - 01 & 2.67e - 01 \\ 5.60e - 01 & 1.71e - 02 \\ 2.59e - 17 & 9.53e - 17 \\ 6.53e - 17 & 2.68e - 17 \\ 2.80e - 17 & 9.16e - 17 \\ 5.42e - 17 & 4.61e - 17 \\ 6.15e - 17 & 3.35e - 17 \\ 3.85e - 17 & 7.35e - 17 \end{pmatrix}$$

$$H = 0.98; H_{max} = 1; L_0 = -14.3$$

Die identischen Dokumente einer völlig homogenen Bibliothek werden nicht alle einem Thema zugeordnet, sondern jeweils in identischer Weise beiden. U.U. bewertet das stochastische Modell aber auch andere Zuordnungen nicht deutlich anders. Diese Vermutung bestätigt sich sofort, wenn man das Iterationsverfahren mit zufälligen Anfangsbedingungen wiederholt (siehe Abbildung 3.14). Obwohl die Log-Likelihood kaum schwankt, findet man eine Vielzahl unterschiedlicher Entropiewerte, die aus unterschiedlichsten Dokument-Themen-Matrizen ($P2$) resultieren. Diese Matrizen stimmen alle in ihren Spalten überein, können aber gleichzeitig in den Zeilen stark variieren. Dies bedeutet, dass die PLSA eine extrem homogene Bibliographie nur schwer in deren Vielfalt bewerten kann. Selbst unterschiedlichste Themenzuordnungen der Dokumente (mit entsprechend unterschiedlicher Diversität) erscheinen nahezu gleich wahrscheinlich, wenn sie mit einer geeigneten Verteilung der Referenzen auf die Themen einhergehen.

Entsprechend kritisch ist das oben genannte einzelne Resultat ($H = 0.98$) für die Diversität in der Beispielrealisation zu sehen. Es wird allerdings nicht zufällig gewonnen, sondern nach wiederholter Anwendung des Iterationsverfahrens (100 Wiederholungen) durch Auswahl desjenigen Ergebnisses mit maximalem Log-Likelihood-Wert. Auch diese Vorgehensweise – die im Folgenden immer angewendet wird – liefert hier allerdings keine Garantie für das Auffinden eines eindeutigen globalen Maximums.

Das globale Maximum im Wahrscheinlichkeitsraum muss sich nicht zwingend mit einem einzelnen Diversitätswert assoziieren lassen. Dies ist dann der Fall, wenn unterschiedliche Parametermatrizen (wie hier) zur selben Wahrscheinlichkeit führen aber gleichzeitig zu verschiedenen Diversitätswerten. Eine kurze Rechnung zeigt, dass auch eine extreme Themenverteilung mit vollständiger Zuordnung von Thema 1 zu allen drei Dokumenten (und Thema 2 zu keinem) denselben Log-Likelihood-Wert liefern kann wie in Abbildung 3.14 abzulesen, wenn die Themen-Referenz Verteilungen geeignet gewählt werden. Im Rahmen dieser Untersuchung ist dieses Phänomen jedoch nur für extrem ho-

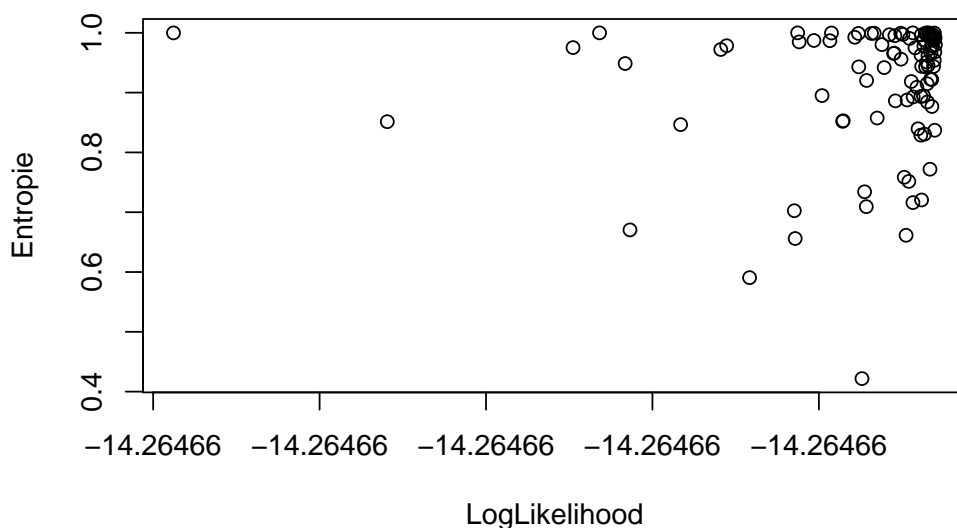


Abbildung 3.14: Entropie in Abhängigkeit von der Log-Likelihood bei 100 Realisationen des PLSA Algorithmus für die völlig homogene Bibliographie. Die Variabilität ist in Entropie-Richtung extrem hoch, während in Log-Likelihood-Richtung kaum Schwankungen auftreten (unterhalb der Genauigkeit der Achsenbeschriftung).

homogene Bibliographien aufgetreten, wie sie so unter den empirischen Daten nicht zu finden sind.

Bewertung einer schwach gekoppelten Bibliographie

Wie verhält es sich mit einer schwach gekoppelten Bibliographie, in der das gleichzeitige Auftreten von Referenzen in verschiedenen Dokumenten vorhanden, aber selten ist? Dazu wird die folgende Referenz-Dokument-Matrix

untersucht:

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 1.00e + 00 & 1.00e + 00 & 1.79e - 09 \\ 1.98e - 09 & 7.74e - 07 & 1.00e + 00 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 1.43e - 01 & 1.09e - 10 \\ 1.43e - 01 & 1.54e - 10 \\ 2.86e - 01 & 6.15e - 11 \\ 1.43e - 01 & 1.48e - 25 \\ 1.43e - 01 & 1.87e - 24 \\ 1.43e - 01 & 2.50e - 01 \\ 1.25e - 35 & 2.50e - 01 \\ 1.78e - 35 & 2.50e - 01 \\ 4.86e - 35 & 2.50e - 01 \end{pmatrix}$$

$$H = 0.918; H_{max} = 1; L_0 = -25.7$$

Obwohl hier Dokumente mit gemeinsamen Referenzen vorliegen, ergibt sich kein Unterschied zum Fall mit linear unabhängigen Dokumentvektoren.

Der Blick auf die Variabilität des Ergebnisses mit unterschiedlichen Anfangsbedingungen (Abbildung 3.15) bestätigt die Vermutung aus dem letzten Abschnitt, dass man im Allgemeinen keine Konvergenz des Algorithmus in das globale Maximum erwarten kann. Für zufällige Anfangsbedingungen finden sich hier einige wenige lokale Maxima, ganz anders als im Fall der völlig heterogenen Bibliographie, von der sich dieses Beispiel nur in zwei Matrixelementen unterscheidet (vgl. Abbildung 3.13). Es ist also in jedem Fall erforderlich das Iterationsverfahren der PLSA zu wiederholen und aus den lokalen Maxima das Ergebnis mit höchster Log-Likelihood auszuwählen.

Bisher gibt es keine Anzeichen dafür, dass gerade globale Maxima durch das Iterationsverfahren mit zufälligen Anfangsbedingungen selten erreicht würden, so dass man sich zunächst mit moderaten Wiederholungszahlen begnügen kann (für die folgenden Ergebnisse stets 100).

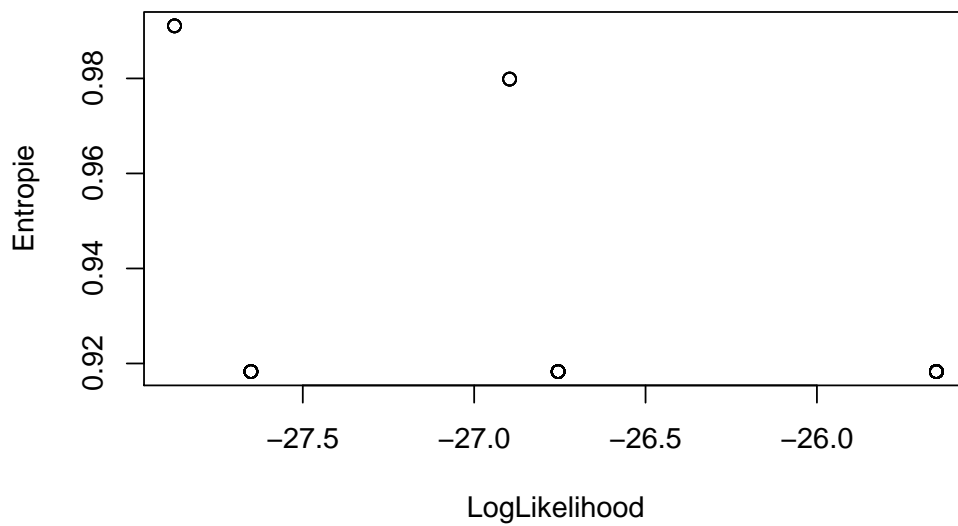


Abbildung 3.15: Entropie in Abhängigkeit von der Log-Likelihood bei 100 Realisationen des PLSA Algorithmus für eine schwach gekoppelte Bibliographie. Variabilität ist sowohl in Entropie- als auch in Log-Likelihood-Richtung zu beobachten.

Nullzeile hinzufügen

Im nächsten Beispiel wird eine Nullzeile an die schwach gekoppelte Bibliographie angefügt.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 1.33e - 35 & 2.43e - 06 & 1.00e + 00 \\ 1.00e + 00 & 1.00e + 00 & 1.03e - 38 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 1.78e - 37 & 1.43e - 01 \\ 1.63e - 37 & 1.43e - 01 \\ 3.52e - 25 & 2.86e - 01 \\ 8.23e - 20 & 1.43e - 01 \\ 1.58e - 19 & 1.43e - 01 \\ 2.50e - 01 & 1.43e - 01 \\ 2.50e - 01 & 2.82e - 48 \\ 2.50e - 01 & 3.32e - 48 \\ 2.50e - 01 & 1.18e - 48 \\ 1.87e - 37 & 3.66e - 35 \end{pmatrix}$$

$$H = 0.918; H_{max} = 1; L_0 = -25.7$$

Eine Nullzeile verändert das vorherige Ergebnis nicht. Die zusätzliche Referenz, die nicht in den Dokumenten vorkommt, erhält für beide Themen die Wahrscheinlichkeit 0. Auf diese Weise wird auch dieselbe Log-Likelihood wie zuvor erreicht.

Nullspalte hinzufügen

Nun wird der schwach gekoppelten Bibliographie ein Dokument ohne Referenzen hinzugefügt.

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 4.52e - 20 & 1.41e - 06 & 1.00e + 00 & 3.92e - 01 \\ 1.00e + 00 & 1.00e + 00 & 1.46e - 16 & 6.08e - 01 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 5.15e - 25 & 1.43e - 01 \\ 1.31e - 24 & 1.43e - 01 \\ 1.40e - 17 & 2.86e - 01 \\ 8.03e - 16 & 1.43e - 01 \\ 6.49e - 17 & 1.43e - 01 \\ 2.50e - 01 & 1.43e - 01 \\ 2.50e - 01 & 1.20e - 24 \\ 2.50e - 01 & 1.37e - 24 \\ 2.50e - 01 & 1.17e - 24 \end{pmatrix}$$

$$H = 0.932; H_{max} = 1; L_0 = -25.7$$

Beim Hinzufügen der Nullspalte tritt zum ersten Mal der Fall auf, dass neben der vollständigen Identifikation einzelner Dokumente mit einzelnen Themen gleichzeitig auch gemischte Zuordnungen entstehen. Das Nulldokument wird zu 39% mit Thema 1 und zu 61% mit Thema 2 identifiziert. Wieder bleibt die Log-Likelihood gegenüber der einfachen schwach gekoppelten Bibliographie unverändert. Die Diversität hat aber zugenommen. Die PLSA fordert die Randbedingung, dass jedem Dokument ein vollständiger Themenpool zugeordnet werden muss (Summation über alle Themen für jedes Dokument gleich 1). Dies gilt auch für ein Nulldokument. Eine Veränderung in der Diversität ist also nicht überraschend. Offensichtlich gelingt es, die zusätzlichen Parameter so zu wählen, dass sich dabei keine Verschlechterung der Log-Likelihood einstellt.

Identische Spalte hinzufügen

Im folgenden Szenario wird die schwach gekoppelte Bibliographie um einen bereits vorhandenen Dokumentvektor erweitert.

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 4.14e - 09 & 1.91e - 19 & 1.00e + 00 & 1.00e + 00 \\ 1.00e + 00 & 1.00e + 00 & 4.37e - 20 & 1.58e - 20 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 4.99e - 10 & 1.43e - 01 \\ 4.89e - 10 & 1.43e - 01 \\ 7.02e - 15 & 2.86e - 01 \\ 1.04e - 59 & 1.43e - 01 \\ 9.05e - 61 & 1.43e - 01 \\ 2.50e - 01 & 1.43e - 01 \\ 2.50e - 01 & 7.13e - 40 \\ 2.50e - 01 & 5.83e - 40 \\ 2.50e - 01 & 5.26e - 40 \end{pmatrix}$$

$$H = 1; H_{max} = 1; L_0 = -33.7$$

Die identischen Dokumente werden beide vollständig mit dem ersten Thema identifiziert. Die Referenzzuordnung zu diesem Thema liefert gleiche Wahrscheinlichkeiten an den Stellen, in denen tatsächlich eine Referenz beobachtet wird, und sonst 0. Auch für das andere Thema verschwinden die Matrixelemente für diejenigen Referenzen, die in den beiden anderen korrespondierenden Dokumenten nicht vorliegen. An den anderen Stellen (mit 1-Elementen) folgen die Wahrscheinlichkeiten der Referenzen ihren relativen Häufigkeiten in den beiden Dokumenten. So liefert die PLSA ein stimmiges Bild bei der Themenextraktion.

Man erhält allerdings wegen der Symmetrie in der Thema-Dokument-Matrix maximale Diversität. Im Vergleich zur LSA hat sich die Diversitätstendenz bei der Erweiterung der schwach gekoppelten Bibliographie mit einem identischen Vektor umgekehrt. Für die LSA lag eine Reduzierung vor, hier findet man eine Erhöhung, die aber bei ähnlichen Szenarien nicht im Allgemeinen zu erwarten ist.

Zufallsmatrix

Wir betrachten nun wieder eine Matrix, deren Einträge elementweise ausgewürfelt werden. Mit Wahrscheinlichkeit 0.2 erscheint eine 1 und entsprechend mit Wahrscheinlichkeit 0.8 eine 0. Im Mittel erwartet man also bei 30 Einträgen sechsmal die 1.

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 5.32e - 35 & 1.00e + 00 & 1.00e + 00 \\ 1.00e + 00 & 2.55e - 12 & 3.33e - 09 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 3.28e - 12 & 1.49e - 15 \\ 3.02e - 12 & 2.82e - 15 \\ 6.67e - 01 & 3.74e - 10 \\ 2.54e - 12 & 4.97e - 15 \\ 6.30e - 13 & 5.00e - 01 \\ 2.31e - 12 & 5.87e - 15 \\ 6.54e - 13 & 5.00e - 01 \\ 1.83e - 12 & 7.49e - 15 \\ 3.33e - 01 & 1.40e - 09 \\ 6.01e - 13 & 1.05e - 14 \end{pmatrix}$$

$$\bar{H} = 0.928; \sigma_H = 0.0924; H_{max} = 1; \bar{L}_0 = -9.44$$

Das Experiment wird 500 mal wiederholt. Die oben angegebenen Matrizen repräsentieren eine einzelne dieser Systemrealisationen. Die mittlere Diversität ist bei äußerst geringer Variabilität zwischen den Systemrealisationen vergleichbar mit dem Fall der schwach gekoppelten Bibliographie.

Zufallsspalten mit zwei Einträgen

Im nächsten Experiment werden die Freiheitsgrade des Systems eingeschränkt. Nun fordern wir genau zwei Einträge pro Spalte der Matrix an zufälligen

Plätzen. Hier muss die Matrix also genau sechsmal die 1 enthalten.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 1.85e - 16 & 1.00e + 00 & 1.00e + 00 \\ 1.00e + 00 & 3.33e - 09 & 3.94e - 15 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 4.16e - 17 & 5.00e - 01 \\ 6.37e - 15 & 1.37e - 14 \\ 3.52e - 17 & 5.00e - 01 \\ 2.50e - 01 & 9.29e - 16 \\ 8.60e - 15 & 1.11e - 14 \\ 2.50e - 01 & 1.12e - 09 \\ 8.75e - 15 & 1.09e - 14 \\ 2.50e - 01 & 1.63e - 15 \\ 2.50e - 01 & 7.21e - 10 \\ 9.37e - 15 & 9.92e - 15 \end{pmatrix}$$

$$\bar{H} = 0.934; \sigma_H = 0.0322; H_{max} = 1; \bar{L}_0 = -8.2$$

Im Vergleich zum vorherigen System gibt es eine sehr kleine Erhöhung, die sich statistisch allerdings erst nachweisen lässt, wenn man die Überschneidung der beiden Entropieverteilungen direkt betrachtet. Weil hier keine Verteilungsannahmen gemacht werden können, kann man alleine aus Mittelwert und Standardabweichung nur wenig schließen.

Die beobachtete Tendenz stimmt mit den Resultaten bei der LSA überein. Die Effektgröße ist aber deutlich kleiner.

Zufallsspalten mit drei Einträgen

Nun wird derselbe Versuch mit genau drei Einträgen pro Spalte wiederholt.

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 1.00e+00 & 2.04e-20 & 2.50e-09 \\ 5.82e-31 & 1.00e+00 & 1.00e+00 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 1.27e-23 & 9.48e-24 \\ 6.64e-24 & 1.36e-23 \\ 3.33e-01 & 3.91e-36 \\ 3.33e-01 & 4.86e-36 \\ 3.33e-01 & 2.34e-36 \\ 8.40e-26 & 1.67e-01 \\ 2.07e-23 & 1.21e-24 \\ 1.41e-12 & 3.33e-01 \\ 6.23e-10 & 1.67e-01 \\ 1.41e-12 & 3.33e-01 \end{pmatrix}$$

$$\bar{H} = 0.922; \sigma_H = 0.0160; H_{max} = 1; \bar{L}_0 = -17.2$$

Hier ergibt sich gegenüber dem letzten Experiment (wie auch im Fall der LSA) ein leichter Abfall.

Größere Matrizen

Die bisherigen Beispiele mit geringer Dokumenten- und Themenzahl könnten den Eindruck erwecken, dass die Zuordnung von Dokumenten und Themen vorwiegend dichotom (0 oder 1) erfolgt. Größere Zufallsmatrizen wie das

folgende Beispiel zeigen allerdings, dass dies nicht der Fall ist.

$$X = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$P2 = \begin{pmatrix} 0 & 0 & 0.5 & 1 & 0.24 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.62 & 0 & 0.58 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0.33 & 0 \\ 1 & 0 & 0 & 0 & 0.14 & 0 & 0 & 0.33 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0.67 & 0.33 & 1 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0.42 & 0 & 0.33 & 0 \end{pmatrix}$$

$$P1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0.43 & 0 \\ 0.53 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.6 \\ 0.22 & 0.26 & 0 & 0 & 0 & 0 \\ 0 & 0.15 & 0 & 0 & 0 & 0 \\ 0 & 0.15 & 0.33 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0.25 & 0.14 & 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0.67 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.57 & 0 \end{pmatrix}$$

$$H = 2.54; H_{max} = 2.58; L_0 = -68.1$$

Die Matrixelemente werden hier der übersichtlicheren Darstellung wegen auf zwei Nachkommastellen gerundet.

Variation der Themenzahl

Schließlich soll noch der Effekt der Themenzahl auf die Entropie untersucht werden. Wie für die LSA werden zwei Matrixtypen analysiert mit zwei bzw.

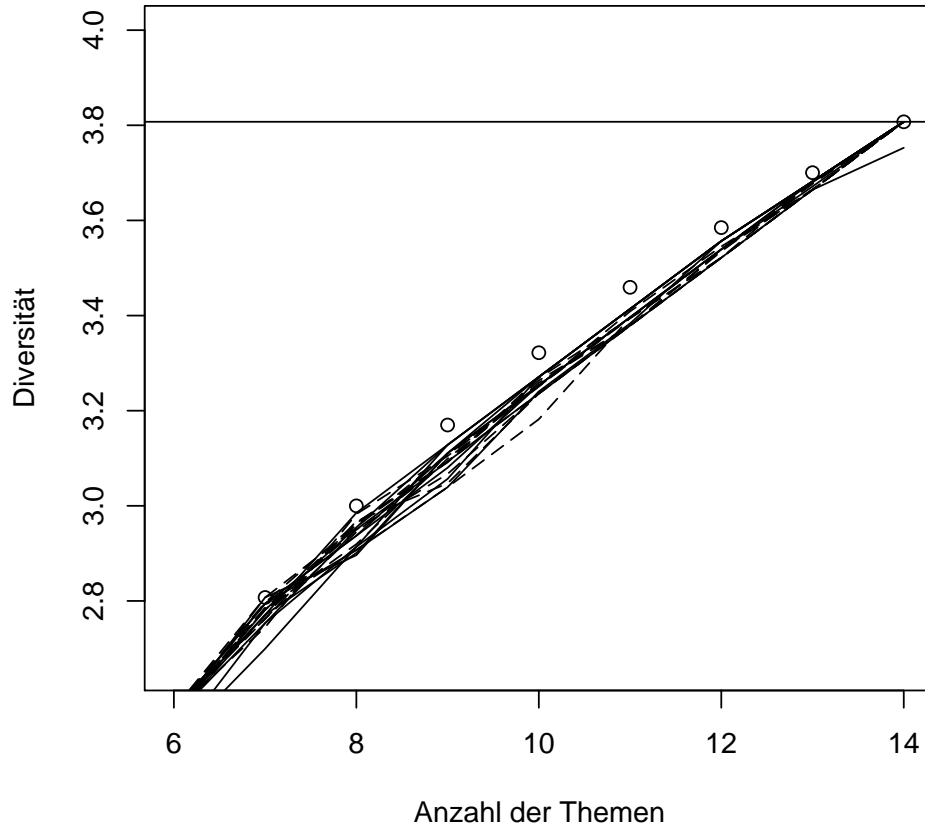


Abbildung 3.16: Diversität in Abhängigkeit der Themenzahl (PLSA). 14 Dokumente, 50 potentielle Referenzen insgesamt ($m = 14$, $n = 50$). Durchgezogene Linie: 2 Referenzen pro Spalte, 10 Wiederholungen; gestrichelte Linie: 10 Referenzen pro Spalte, 10 Wiederholungen. Kreise: maximale Diversität.

mit zehn 1-Elementen pro Spalte bei einer Gesamtgröße von 50×14 . Anders als für die LSA lässt die PLSA keine Identifikation der beiden verwendeten Matrixtypen anhand ihrer Diversitätswerte zu (siehe Abbildung 3.16). Trotz der Strukturähnlichkeit erscheinen die Matrizen für die PLSA sehr heterogen. Die verschiedenen 1-Elementzahlen in den Spalten können keine Unterschiede in der Diversitätsbewertung durch die PLSA hervorzurufen.

3.2.2 Szientometrie

Die bisherige Analyse der Diversitätsberechnungen auf Basis einer PLSA Klassifikation lässt noch kein endgültiges Urteil über die Eignung der Methode zu. Dennoch sollen zumindest exemplarisch einige empirische Datensätze damit ausgewertet werden.

Durch die erforderlichen Wiederholungen des Iterationsverfahrens übersteigt der Rechenaufwand der PLSA den der LSA bei der hier verwendeten Implementierung deutlich. Eine typische PLSA-Auswertung der Szientometriedaten für eine vorgegebene Themenzahl dauert einige Tage auf einem aktuell handelsüblichen PC. Deswegen bleibt für diese Arbeit nur die Möglichkeit, die szientometrische Bibliographie, für einige wenige Themenzahlen zu untersuchen. Gewählt werden 10 (Abbildung 3.17) und 20 Themen (Abbildung 3.18). Die Stichproben werden in ähnlicher Weise gezogen wie zuvor für die LSA, 20 mal 100 Artikel pro Jahrgang. (Die reduzierte Zahl von Wiederho-

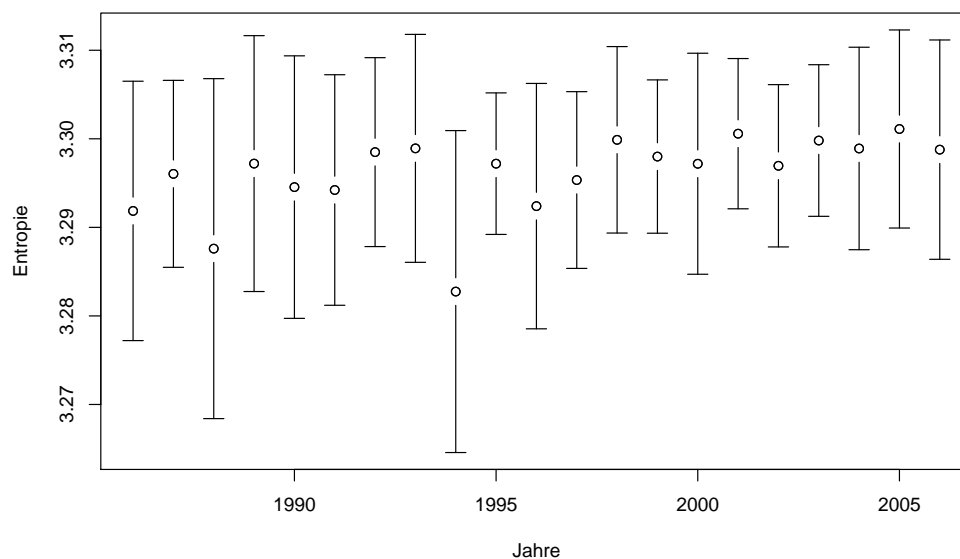


Abbildung 3.17: Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie (PLSA mit 10 Themen). Aus dem Dokumentenpool werden pro Jahrgang 20 Stichproben mit je 100 Artikeln gezogen und deren Diversität durch die PLSA berechnet. Die Fehlerbalken entsprechen der Standardabweichung in den 20 Wiederholungen.

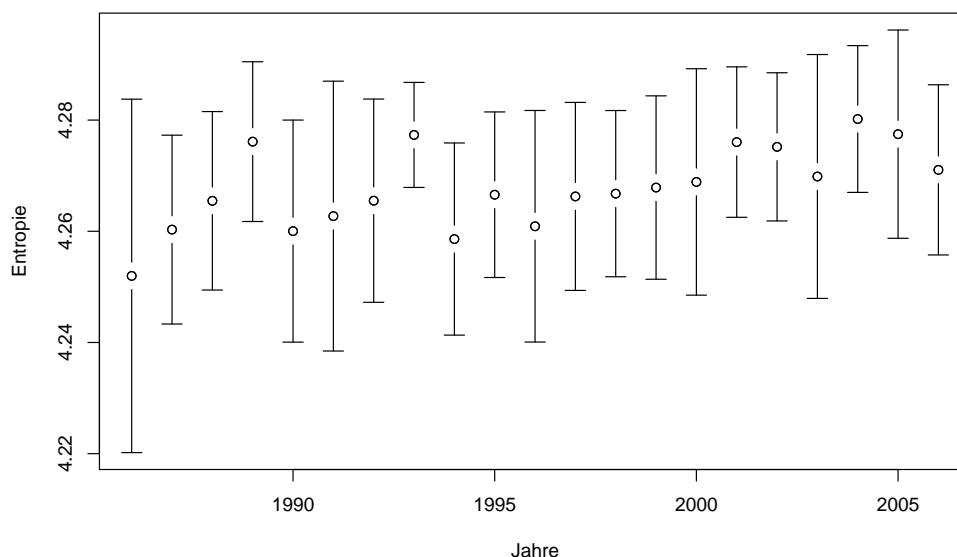


Abbildung 3.18: Zeitliche Entwicklung der Diversität in den ausgewählten Zeitschriftenjahrgängen zur Szientometrie (PLSA mit 20 Themen). Aus dem Dokumentenpool werden pro Jahrgang 20 Stichproben mit je 100 Artikeln gezogen und deren Diversität durch die PLSA berechnet. Die Fehlerbalken entsprechen der Standardabweichung in den 20 Wiederholungen.

lungen ist wiederum der Rechenzeit geschuldet.)

In beiden Analysen kann man einen äußerst schwachen Trend hin zu größeren Diversitätswerten vermuten. Dies wird durch einen F-Test auf linearen Anstieg bestätigt. In beiden Fällen liefert der Signifikanztest kleinste p -Werte ($p < 0.0001$). Gleichzeitig sind die Effektgröße und der Anteil der durch eine lineare Regression erklärten Varianz – wie nach den Abbildungen zu erwarten ist – sehr klein. Die Rechnungen können auch genutzt werden, um zu bestätigen, dass die wiederholten Iterationsverfahren für die empirischen Daten stets eine überschaubare Menge an lokalen Maxima liefern und Mehrdeutigkeiten wie bei der homogenen Bibliographie nicht auftreten.

Dieses Resultat, welches in seiner Tendenz mit den Ergebnissen der LSA übereinstimmt, ermutigt in Zukunft noch größere Themenzahlen zu untersuchen und auch die Datensätze zur Elektrochemie mit einzubeziehen.

Kapitel 4

Diskussion

4.1 LSA und PLSA

Den Diversitätsbestimmungen in dieser Arbeit liegen zwei klassifikatorische Methoden zu Grunde, die LSA und die PLSA. Anders als beispielsweise die Analyse zur Spezialisierung in biologischen Nahrungsnetzen, die sich in einfachen Fällen als bipartite Netzwerke darstellen lassen (Blüthgen, Menzel und Blüthgen 2006; Blüthgen, Menzel, Hovestadt, Fiala und Blüthgen 2007; Bascompte, Jordano und Olesen 2006), kann die Diversitätsanalyse hier nicht im Originalraum der Daten vorgenommen werden. Während in Nahrungsnetzen die zentrale Klassifikation, nämlich die Zuordnung von Individuen zu Arten, bereits von vornherein (z.B. anhand von physiologischen Kriterien) vorgenommen werden kann, findet die elementare Datenerhebung in der vorliegenden szientometrischen Untersuchung nur auf dem untergeordneten Individuenniveau der einzelnen Publikation statt. Die „Artzuordnung“ der Publikationen muss dann indirekt in Form der Themenzuordnung durch automatische Klassifikationsverfahren wie die LSA und PLSA geleistet werden. Ein wesentlicher Unterschied zur biologischen Analogie besteht in der Aufhebung der eindeutigen Zuordnung. Während ein Organismus klassisch genau mit einer biologischen Art identifiziert wird, kann eine Publikation anteilig mehreren Themen zugeordnet werden. Auch in der Biologie wurden bereits ähnliche, erweiterte Konzepte zur Diversitätsmessung entwickelt (Shimatani 2001, und dort zitierte Quellen).¹

Zunächst wurde die LSA bezüglich ihrer Eignung für die Diversitätsmessung untersucht. Die Anwendung auf verschiedene fiktive Testbibliographien hat die erwarteten, plausiblen Ergebnisse geliefert und gezeigt, dass für die

¹Siehe auch im Abschnitt Einleitung.

empirischen Daten mit Diversitätswerten nahe des Maximums zu rechnen ist. Nicht völlig zwingend erscheint die Verwendung der Eigenwerte zur Berechnung der Diversität, nachdem in der SVD zunächst nur die Singulärwerte auftreten. Die Spiegelung der Eigenvektoren darf allerdings keinen Effekt auf die Diversitätsberechnung haben. Damit ist es plausibel, die Quadrate der transformierten Koordinaten und dementsprechend die Eigenwerte als Quadrate der Singulärwerte zu verwenden.

Auch die PLSA wurde mit den einfach strukturierten Testbibliographien untersucht. Dabei hat man gelegentlich Ergebnisse erhalten, die auf den ersten Blick unerwartet schienen. Nach eingehenderen Überlegungen haben allerdings auch diese Klassifikationsresultate vor dem Hintergrund des generativen Wahrscheinlichkeitsmodells ein in sich stimmiges Bild ergeben. Die Testdaten wurden so gewählt, dass ein einfacher Vergleich zwischen LSA und PLSA möglich wurde. Es ist allerdings zu befürchten, dass dadurch für die Wirkungsweise der PLSA – anders als für die der LSA – noch kein vollständig repräsentatives Bild vorliegt. Die bisherige Untersuchung ist nicht erschöpfend. Sie hat weitere Fragen aufgeworfen, die in der Zukunft noch beantwortet werden müssen.

Gerade für sehr homogene Bibliographien scheint die PLSA kaum Trennschärfe bei der Themenextraktion zu besitzen. Liegt dies an den spezifischen Testdaten, oder gilt dies allgemein? Hier sind weitere Experimente nötig. Was passiert, wenn man die Homogenität allmählich zerstört? Die PLSA scheint tendenziell Diversitätswerte zu liefern, die noch näher am Maximum liegen als bei der LSA. Welche minimalen Werte können mit Hilfe der PLSA überhaupt identifiziert werden? Welche Strukturen haben die zugehörigen Matrizen? Unter Umständen kann man an dieser Stelle auch durch analytische Verfahren weiterkommen.

Nachdem die empirischen Daten stark heterogene und schwach gekoppelte Strukturen aufweisen und gleichzeitig in großer Menge vorliegen, sind die Antworten auf obige Fragen u.U. nicht relevant für die Analyse der erhobenen Daten, würden aber sicherlich das Verständnis für die Methode ein weiteres Stück voranbringen.

Die Bedeutung der Themenzahl muss ebenfalls noch geklärt werden. Es liegt nahe die PLSA für verschiedene Themenzahlen durchzuführen und dann diejenige Anzahl an Themen auszuwählen, die maximale Log-Likelihood liefert. Voraussetzung ist, dass ein eindeutiges Maximum im Innern der Menge der erzeugten Wahrscheinlichkeiten existiert.

Offensichtlich ist insgesamt: Unterschiedliche Klassifikationsverfahren liefern unterschiedliche Klassifikationen für dieselbe Menge, selbst wenn die Methoden verwandt sind, wie LSA und PLSA. Beide Verfahren besitzen eine

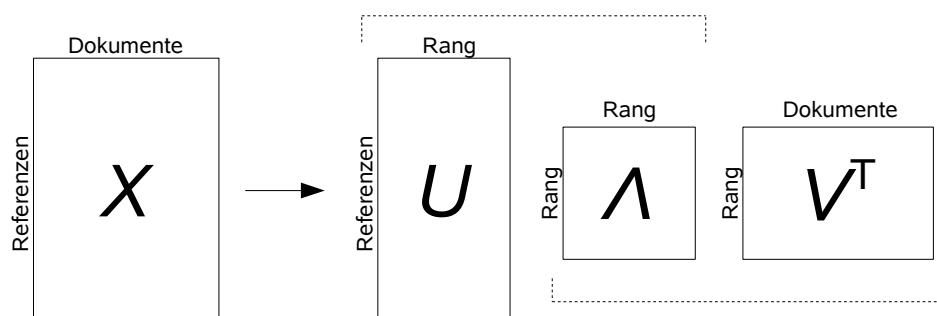


Abbildung 4.1: Schematische Darstellung der LSA in Matrixform.

eingängige Matrixrepräsentation (siehe Abbildungen 4.1 und 4.2). Die LSA wurde im Kern als Matrixfaktorisierung hergeleitet (siehe Abschnitt 3.1 und Abbildung 4.1). Sie zerlegt die Koordinaten der Dokumente im Referenzraum in linear unabhängige Vektoren und unterstellt dabei ein lineares Modell (Golub und Loan 2007), das sich zunächst als ausreichende Näherung für die wahren Verhältnisse erwiesen hat. Um die PLSA in Matrixschreibweise darzustellen und in Analogie zur LSA zu setzen, bietet sich die symmetrische Repräsentation des Modells an (Hofmann 1999b). Aus Abbildung 4.2 wird dann deutlich, dass die Themenverteilung in der PLSA die Rolle der Eigenwerte in der LSA übernimmt und die beiden Matrizen der bedingten Wahrscheinlichkeiten von Dokumenten und Referenzen die Rollen der linken bzw. rechten Eigenvektoren übernehmen.

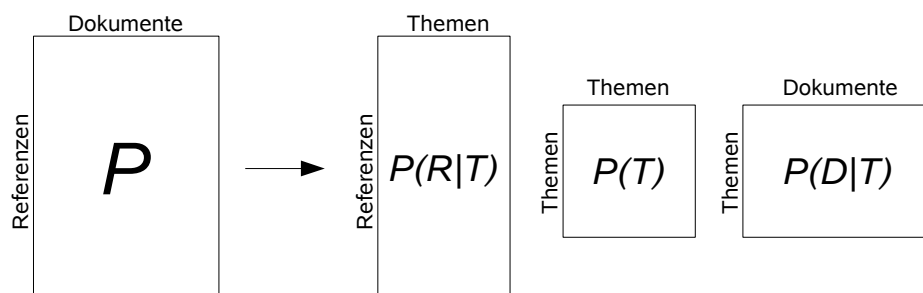


Abbildung 4.2: Schematische Darstellung der PLSA in Matrixform. Die Matrix P enthält die Wahrscheinlichkeiten $P(i \wedge j)$. Die Matrizen $P(R|T)$ und $P(D|T)$ enthalten die bedingten Wahrscheinlichkeiten $P(i|k)$ und $P(j|k)$. $P(T)$ ist eine Diagonalmatrix mit den Wahrscheinlichkeiten $P(k)$ auf der Hauptdiagonalen.

Die PLSA liefert für alle Themenzahlen unabhängig voneinander ein optimales Paar an Parametermatrizen. Auch die LSA gehorcht einem Maximierungsprinzip. Mit jedem der Eigenvektoren, geordnet nach fallenden Eigenwerten, wird stets der größtmögliche Anteil an verbliebener Restvarianz im Sinne der linearen Regression erklärt (Bortz 2005). Die grundlegenden Unterschiede zwischen den beiden Optimierungsprinzipien werden bei Hofmann (1999b) erläutert.

Die methodische Weiterentwicklung der Latenten Semantischen Analyse setzt hauptsächlich bei der PLSA an. Nachdem man erkannt hat, dass die Themenverteilung in den Dokumenten nicht zufriedenstellend repräsentiert wird, verwendet man die Dirichlet-Verteilung als zusätzliche a priori Verteilungsannahme (Blei, Ng und Jordan 2003). Man spricht dann von *Latent Dirichlet Allocation (LDA)*. Dieses Verfahren wurde bereits erfolgreich zur Untersuchung der zeitlichen Entwicklung von Themen in Bibliographien eingesetzt (Griffiths und Steyvers 2004). Die Analyse von Themendynamiken hat durch die Konzentration auf den statistischen Bereich insgesamt eine Reihe neuer Impulse erhalten und beginnt sich im Moment zu entfalten (Xing 2005; Blei und Lafferty 2006; Newman, Chemudugunta, Smyth und Steyvers 2006; Wallach 2006).

4.2 Diversitätsentwicklung

Ein zentrales Ergebnis dieser Arbeit hat die Anwendung der LSA auf umfangreiche Datensätze aus den Bereichen *Szientometrie* und *Elektrochemie* geliefert. Die Analysen konnten in beiden Bereichen keinen Anhaltspunkt für die Homogenisierungshypothese liefern. Es scheint sogar der umgekehrte Trend vorzuliegen. Die Vielfalt in den betrachteten Forschungslandschaften steigt an und nähert sich dem mit den genutzten Methoden nachweisbaren Maximum.

Die bisherigen Resultate legen nahe, die weitere Vorgehensweise zunächst auf die LSA zu konzentrieren und zum einen die Kausalanalyse voranzutreiben und zum anderen die Auswertung zu differenzieren.

Was ist die Ursache für die beobachtete steigende Tendenz der Diversität nahe am Maximum? Bisher ist diese Frage noch nicht geklärt. Einige weitere Randomisierungsexperimente könnten dabei aufschlussreich sein. Obwohl gezeigt werden konnte, dass die Tendenz auch bei konstanter Artikelzahl pro Jahrgang und konstanter mittlerer Referenzanzahl pro Artikel erhalten bleibt, könnte es weitere Eigenschaften der Bibliographie geben, die lediglich statistischer Natur sind und u.U. eine Erklärung für die Beobachtungen liefern. Dies gilt z.B. für die mittlere Anzahl *unterschiedlicher* Referenzen im

Dokumentenpool, die anders als die mittlere Gesamtzahl von Zitationen pro Artikel bisher noch nicht normiert wurde. Findet man hier eine systematische Entwicklung im Laufe der Zeit, so sollte überprüft werden ob das Gesamtergebnis auch erhalten bleibt, wenn man die Tendenz mit einem geeigneten Zufallsexperiment zerstört.

Für eine gegebene Dokumentenzahl und Referenzenverteilung ist die Menge der möglichen, resultierenden Bibliographien begrenzt. Nur ein geringer Teil der Referenzen kommt mindestens zweimal vor, die meisten Quellen werden nur in einem Dokument zitiert. Welche Variabilität lässt sich überhaupt in der Diversität erzeugen, wenn man zufällige Paare von Referenzen in unterschiedlichen Dokumenten vertauscht (und dabei Mehrfachzitationen derselben Referenz innerhalb desselben Dokuments verbietet)? Die Antwort auf diese Frage könnte eine neue Referenzdiversität liefern, die evtl. die Nähe der bisher bestimmten Diversitätswerte zum Maximum relativiert.² In jedem Fall wird sie zu einem besseren Verständnis der Methode führen. Sollte die beobachtete Tendenz aller geforderter Kritik standhalten, könnte sie die Hypothese von Krohn und Küppers (1989) bestätigen, die in der Autonomie der Wissenschaften paradoxerweise eine Ursache für Themenkonzentration sehen. Diese Vereinheitlichungstendenz könnte, so die Autoren, durch eine stärkere Konkurrenz um Mittel aufgebrochen werden.

Nachdem die Themenvielfalt auf der Ebene der einzelnen Artikel gemessen werden kann und sowohl Autoren als auch deren Nationalitäten in den Metadaten verfügbar sind, bietet es sich an, die Diversitätsanalyse auch nach Ländern zu differenzieren. Dazu bestimmt man die länderspezifischen Diversitätsbeiträge der einzelnen Themen, indem man jedes Dokument mit dem Anteil der landeseigenen Autoren an der gesamten Autorenzahl gewichtet. Interessant wäre dann die Frage, ob es zwischen den Ländern Unterschiede in der Diversitätsentwicklung gibt.

Bisher wurde nur die *Shannonsche Entropie* als Diversitätsmaß getestet. Unter Umständen liefert der in der Einleitung erwähnte *Simpson-Index* eine bessere Auflösung der Diversitätswerte. Dies lässt sich in den bisherigen Auswertungen einfach ergänzen.

Spezielle Aufmerksamkeit sollte auch der Rolle des Stichprobenumfangs geschenkt werden. Unter Umständen kann durch Extrapolation eine Unabhängigkeit des Diversitätsmaßes von der Stichprobengröße und damit die direkte Vergleichbarkeit zwischen Datensammlung unterschiedlichen Umfangs (wie z.B. zwischen Elektrochemie und Szientometrie) erreicht werden. Dudok de Wit (1999) hat eine Methode entwickelt, den Effekt der endlichen

²Referenzdiversität ist hier im Sinne von Vergleichsdiversität zu verstehen. Es ist also nicht eine Referenz in einem Dokument gemeint.

Stichprobengröße auf die Entropie zu korrigieren.

4.3 Weitere auswertbare Dokumenteigenschaften

Eine andere bisher ungeklärte Frage ist ebenfalls von zentraler Bedeutung. Genügen die durch die Referenzlisten gegebenen Informationen, um die Dokumente hinreichend zu charakterisieren, so dass eine Themenzuordnung grundsätzlich überhaupt möglich ist? Diese Frage lässt sich auf unterschiedliche Weise beantworten. Zum einen kann man weitere formale Dokumenteigenschaften neben den Referenzlisten als formale Klassifikationsgrundlage bemühen, oder man nimmt eine intellektuelle Interpretation der gelieferten formalen Klassifikation vor und prüft sie so auf Plausibilität.

Mit Hilfe von LSA und PLSA lassen sich neben den Referenzlisten auch andere Dokumenteigenschaften auswerten. Die ISI Web of Science Metadaten enthalten eine Reihe weiterer Informationen, wie Titel, Stichworte und Abstracts zu den Veröffentlichungen. Nach entsprechenden Vorbereitungen zur Standardisierung wie dem Ausschluss von Stoppwörtern oder der linguistischen Vorverarbeitung zur Reduzierung auf Wortstämme können Begriffe aus Titel, Stichworten oder Abstracts ganz ähnlich wie die Referenzlisten analysiert werden (Quesada 2007). Neben Einzelworten könnten auch Worttupel verwendet werden. Der anschließende Vergleich mit der Klassifizierung auf Grundlage der Referenzlisten wird zeigen, ob ein stabiles Phänomen gemessen wurde, d.h. ob es überhaupt möglich ist auf Basis von Referenzlisten auf Zusammenhänge zwischen Artikeln zu schließen, oder ob die angenommene Repräsentation eines Artikels durch die Referenzliste zu lückenhaft ist. Bestenfalls liefern hier unterschiedliche Ansätze eine ähnliche Klassifikation, so dass man schließen kann, dass die unterschiedlichen Vorgehensweisen zur Auswertung dieselben Konzepte erfassen. Ist dies nicht der Fall, so liegt es am nächsten, die unterschiedlichen Kriterien zu kombinieren (White und Griffith 1981; Braam, Moed und van Raan 1991).

All diese Ansätze lassen sich vollständig auf der formalen Ebene durchführen. Der entscheidende Nachweis, dass die Methode eine sinnvolle Klassifikation liefert, gelingt allerdings erst, wenn die Klassifikation einer intellektuellen Begutachtung unterworfen wird.

4.4 Expertenbefragung

Dazu müssen die formal gewonnenen Themen Experten aus der Szientometrie und der Elektrochemie vorgelegt und durch diese interpretiert werden. Dies ist allerdings erst möglich, wenn aus der großen Zahl von Stichproben, die für die Auswertungen innerhalb und über die Jahrgänge hinweg verwendet wurden, eine einzelne repräsentative Analyse zur Vorlage bei den Experten ausgewählt wurde. Dafür sollte die thematische Ähnlichkeit zwischen den Stichproben noch näher untersucht werden, um dann eine geeignete Auswahl zu treffen oder eine zusätzliche Analyse auf Grundlage aller verfügbarer Daten (nicht nur derer in den Stichproben) zu wählen.

Alle bisher erwähnten Methoden setzen bei einer formalen Analyse an, erfordern im Nachhinein allerdings eine intellektuelle Interpretation. Dies gilt für viele Verfahren des maschinellen Lernens, die in den Bereich des unüberwachten Lernens fallen, wie die typischen Clusterverfahren in ihren zahlreichen Varianten oder bestimmte Typen neuronaler Netze (Witten und Frank 2005; Bishop 2006).

Grundsätzlich könnte man auch umgekehrt vorgehen und die intellektuelle Erschließung an den Anfang setzen. Ein Experte identifiziert die Themenbereiche seines Arbeitsgebietes und nimmt eine thematische Einordnung für ausgewählte Publikationen vor. Die große Schwäche dieser Methode liegt in der Subjektivität, der die Analyse durch die persönliche Beurteilung unterworfen ist. Dem könnte man zum einen dadurch begegnen, dass man die auf ein Minimum an Datensätzen begrenzte Erschließung mit verschiedenen Personen wiederholt und als Grundlage für ein überwachttes maschinelles Lernverfahren nutzt. Durch die Expertenklassifikation entstehen Testdatensätze, die dann z.B. durch ein neuronales Netz oder ein statistisches Modell gelernt und verallgemeinert werden. Auf diese Weise können umfangreiche Datenbestände klassifiziert und einer Diversitätsbestimmung zugänglich gemacht werden. Zum andern könnte die Expertenbefragung auch zwischen zwei formalen Analysen vorgenommen werden. Zu Beginn stünde dann eine grobe Vorklassifikation, die Vollständigkeit garantiert. Die Expertenbefragung sorgt anschließend für Interpretierbarkeit, macht aber an manchen Stellen u.U. eine Veränderung der Klassifikation notwendig. Diese wird dann von einem zweiten formalen Verfahren durchgeführt.

Zusammenfassung

Für diese Arbeit wurden umfangreiche Metadaten zu wissenschaftlichen Veröffentlichungen in den Forschungsbereichen *Szientometrie* und *Elektrochemie* aus der Online-Datenbank *ISI Web of Science* gewonnen, zu zwei entsprechenden Bibliographien zusammengefasst und für die automatisierte Auswertung vorbereitet. Zur Auswertung der Metadatenansammlungen wurden mit Hilfe der Statistiksoftware *R* zwei Methoden implementiert, analysiert und deren klassifikatorische Ergebnisse als Ausgangspunkt für eine Diversitätsmessung und zur Untersuchung der Diversitätsentwicklung verwendet.

Die *Latente Semantische Analyse* basiert im Kern auf einer Singulärwertzerlegung und führt die vektorielle Darstellung einer Bibliographie in eine normierte, linear unabhängige Darstellung über, aus der latente Themen und insbesondere deren relative Bedeutung abgelesen werden können. Die Gewichtung der latenten Themen wird dann zur Berechnung der Diversität in Form der *Shannonsche Entropie* genutzt. Damit erhält man eine einzelne Kenngröße für jede Artikelsammlung, die es ermöglicht verschiedene Sammlungen, insbesondere die zeitliche Entwicklung von Zeitschriftenjahrgängen zu vergleichen.

Einfache Experimente mit fiktiven Daten haben zunächst gezeigt, dass die Analyse plausible Ergebnisse liefert. Die erzielten Diversitätswerte liegen allerdings für typische Bibliographien nahe am möglichen Maximum, so dass nur eine geringe Trennschärfe erreicht werden kann. Dennoch konnte sowohl im Bereich der *Szientometrie* als auch für die *Elektrochemie* eine steigende Tendenz nachgewiesen werden, mit der sich die Diversität dem Maximum annähert.

Die *Probabilistische Latente Semantische Analyse* basiert auf einem generativen statistischen Modell und liefert eine Wahrscheinlichkeitsverteilung für die in den Daten enthaltenen Themen. Diese Wahrscheinlichkeitsverteilung dient als Grundlage für die Diversitätsberechnung. Die Methode wurde an denselben Testbibliographien wie die LSA erprobt. Die Anwendung auf fiktive Daten hat teilweise überraschende Ergebnisse geliefert, die in sich allerdings ein stimmiges Bild ergeben haben. Die Tendenz zu Diversitätswerten

nahe am Maximum ist bei der PLSA noch ausgeprägter als bei der LSA. Für sehr homogene Bibliographien, gelingt die Zuordnung eines eindeutigen Diversitätswertes nur schwer. Unterschiedliche Diversitätswerte erscheinen in homogenen Bibliographien als gleich wahrscheinlich. Es konnte noch nicht geklärt werden, welches Spektrum an Diversität überhaupt von der PLSA geliefert werden kann. Dies wirft neue Fragen auf, die in Zukunft untersucht werden müssen. Die Analyse der Methode ist an dieser Stelle noch nicht abgeschlossen.

Für die empirischen Daten zur *Szientometrie* konnte von der PLSA ein äußerst schwacher Anstieg der Diversität nachgewiesen werden. Die Untersuchung hat sich wegen der erforderlichen Rechenzeit allerdings nur auf kleine Themenmengen erstreckt und muss auf größere Themenzahlen sowie die Daten zur Elektrochemie ausgeweitet werden.

Literatur

- Alter, O., P. Brown und D. Botstein (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the USA* 97(18), S. 10101–10106.
- Bascompte, J., P. Jordano und J. Olesen (2006). Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science* 312, S. 431–433.
- Bishop, C. (2006). *Pattern recognition and machine learning (Information Science and Statistics)*. Berlin: Springer.
- Blei, D. und J. Lafferty (2006). Dynamic topic models. In: W. W. Cohen und A. Moore (Hrsg.), *In Proceedings of the 23rd International Conference on Machine Learning*, Volume 148 of *ACM International Conference Proceeding Series*, S. 113–120. ACM.
- Blei, D., A. Ng und M. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, S. 993–1022.
- Blüthgen, N., F. Menzel und N. Blüthgen (2006). Measuring specialization in species interaction networks. *BMC Ecology* 6(9).
- Blüthgen, N., F. Menzel, T. Hovestadt, B. Fiala und N. Blüthgen (2007). Specialization, constraints, and conflicting interests in mutualistic networks. *Current Biology* 17, S. 341–346.
- Bordons, M., F. Morillo und I. Gómez (2004). *Handbook of quantitative science and technology research*, Chapter Analysis of cross-disciplinary research through bibliometric tools, S. 437–456. Dordrecht: Kluwer.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Botafogo, R., E. Rivlin und B. Shneiderman (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems (TOIS)* 10(2), S. 142–180.

- Braam, R., H. Moed und F. van Raan (1991). Mapping of science by combined cocitation and word analysis. I. Structural aspects. *Journal of the American Society of Information Science* 42(4), S. 233–251.
- Chien, J., M. Wu und C. Wu (2005). Bayesian learning for latent semantic analysis. Vortrag Interspeech 2005 - Eurospeech; Lisabon, Portugal. (Vortrag; Insitutshomepage: <http://speech.csie.ntnu.edu.tw>).
- Deerwester, S., S. Dumais, G. Furnas, T. Landauer und R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), S. 391–407.
- Dudok de Wit, T. (1999). When do finite sample effects significantly affect entropy estimates? *The European Physical Journal B* 11, S. 513–516.
- Egghe, L. und R. Rousseau (2003). BRS-compactness in networks: Theoretical considerations related to cohesion in citation graphs, collaboration networks and the internet. *Mathematical and Computer Modelling* 37(7-8), S. 879–899.
- Fahrmeir, L., A. Hamerle und G. Tutz (1996). *Multivariate statistische Verfahren* (2. überarb. Aufl.). Berlin: Walter de Gruyter.
- Golub, G. H. und C. F. V. Loan (2007). *Matrix computations* (3. Aufl.). Johns Hopkins Studies in the Mathematical Sciences. New Delhi, India: Hindustan Book Agency.
- Griffiths, T. und M. Steyvers (2004, April). Finding scientific topics. *Proceedings of the National Academy of Sciences of the USA* 101 Suppl 1, S. 5228–5235.
- Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. *Scientometrics* 18(3-4), S. 219–239.
- Hagenaars, J. und A. McCutcheon (Hrsg.) (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.
- Havemann, F., M. Heinz, M. Schmidt und J. Gläser (2007). Measuring diversity of research in bibliographic-coupling networks. In: D. Torres-Salinas und H. Moed (Hrsg.), *Proceedings of ISSI 2007*, Volume 2, Madrid, S. 860–861. (Poster Abstract).
- Hofmann, T. (1999a). Probabilistic latent semantic analysis. In: *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, UAI'99*, Stockholm.
- Hofmann, T. (1999b). Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, S. 50–57.

- Kaban, A. (2004). Lecture 5: Probabilistic latent semantic analysis. University of Birmingham. (Vorlesung; Instituts-Homepage: <http://www.cs.bham.ac.uk>).
- Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation* 14, S. 10–25.
- Krohn, W. und G. Küppers (1989). *Die Selbstorganisation der Wissenschaft*. Frankfurt: Suhrkamp.
- Landauer, K., D. McNamara, S. Dennis und W. Kintsch (Hrsg.) (2007). *Handbook of latent semantic analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Marshakova, I. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2– Informatsionnye Protsessy I Sistemy* 2(6), S. 3–8.
- Newman, D., C. Chemudugunta, P. Smyth und M. Steyvers (2006). Analyzing entities and topics in news articles using statistical topic models. In: *ISI*, San Diego, CA, USA, S. 93–104.
- Quesada, J. (2007). *Handbook of latent semantic analysis*, Chapter Creating your own LSA space, S. 71–88. Lawrence Erlbaum Associates.
- R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Rafols, I. und M. Meyer (2007a). Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. In: D. Torres-Salinas und H. Moed (Hrsg.), *Proceedings of ISSI 2007*, Volume 2, Madrid, S. 631–637.
- Rafols, I. und M. Meyer (2007b). How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics* 70(3), S. 633–650.
- Schmidt, M., J. Gläser, F. Havemann und M. Heinz (2006). A methodological study for measuring the diversity of science. In: *Proceedings des International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, 10-12 May 2006*, Nancy, S. 129–137.
- Shimatani, K. (2001). On the measurement of species diversity incorporating species differences. *Oikos* 93(1), S. 135–147.
- Simpson, E. (1949). Measurement of diversity. *Nature* 163(4148), S. 688.

- Small, H. (1973). Cocitation in scientific literature: a new measure of relationship between two documents. *Journal of the American Society for Information Science* 24, S. 265–269.
- Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science* 8(3), S. 327–340.
- Small, H. und E. Sweeney (1985). Clustering the Science Citation Index using cocitations. *Scientometrics* 7(3-6), S. 391–409.
- Small, H., E. Sweeney und E. Greenlee (1985). Clustering the Science Citation Index using cocitations II: Mapping science. *Scientometrics* 8(5-6), S. 321–340.
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface* 4(15), S. 707–719.
- Wallach, H. (2006). Topic modeling: beyond bag-of-words. In: *ICML*, S. 977–984.
- White, H. und B. Griffith (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society of Information Science* 32(5), S. 163–171.
- Witten, I. und E. Frank (2005). *Data mining: Practical machine learning tools and techniques* (2. Aufl.). San Francisco, CA: Morgan Kaufmann.
- Xing, E. (2005). On topic evolution. *School of Computer Science, Carnegie Mellon University*, S. 5–115.

Anhang A

Listing 1: Quellcode für LSA-Auswertung (exemplarisch für Abbildung 3.11)

```
1 # Bibliothek fuer schnelle SVD laden;
2 # muss zuvor einmal heruntergeladen werden,
3 # da nicht in R-Standardinstallation enthalten.
4 library(corpcor)
5
6 # Lokalen Datenpfad definieren;
7 # muss auf anderen Rechnern entsprechend angepasst werden.
8 # st="I:\\Daten\\Bibliothek\\Masterarbeit\\data\\"
9 st="/home/mitesser/data/szientometrie/"
10
11 # Auswahl des Datensets
12 # 1 = Szientometrie
13 # 2 = Elektrochemie
14 SelDataSet=2
15
16 # Dateinamen, die geladen werden sollen, definieren (Dateien muessen im eben
17 # definierten Verzeichnis liegen)
18 if (SelDataSet==1){
19   st="/home/mitesser/data/szientometrie/"
20   datapaths=c(
21     "STM80-06.txt",
22     "IPM62-07.txt",
23     "JOD73-07.txt",
24     "JIS68-07.txt",
25     "JAS56-07.txt"
26   )
27 }
28 if (SelDataSet==2){
29   st="/home/mitesser/data/elektrochemie/"
30   datapaths=c(
31     "ElChemXXBAB.txt",
32     "ElChemXXBIO.txt",
33     "ElChemXXCVD.txt",
34     "ElChemXXCOS.txt",
35     "ElChemXXELA.txt",
36     "ElChemXXELC.txt",
37     "ElChemXXECA1.txt",
38     "ElChemXXECA2.txt",
39     "ElChemXXAEC.txt",
40     "ElChemXXJEC1.txt",
41     "ElChemXXJEC2.txt",
42     "ElChemXXJEC3.txt",
43     "ElChemXXJES1.txt",
44     "ElChemXXJES2.txt",
45     "ElChemXXJES3.txt",
46     "ElChemXXJES4.txt",
47     "ElChemXXJES5.txt",
48     "ElChemXXPSF.txt",
49     "ElChemXXRJE.txt",
50     "ElChemXXSAC.txt",
51     "ElChemXXSSI1.txt",
52     "ElChemXXSSI2.txt",
53     "ElChemXXJPS1.txt",
54     "ElChemXXJPS2.txt"
55   )
56 }
57
58
59 # Anzahl der zu verarbeitenden Dateien bestimmen
60 SetNumber=length(datapaths)
61
62 # Variable zur Datenaufnahme leeren (falls vorher besetzt)
```

```

63 z=c()
64 zhlp=c()
65
66 # Auswahl der Spalten in den Textdateien, die verwendet werden sollen.
67 # Die anderen Spalten brauchen nicht im Arbeitsspeicher gehalten werden
68 fff=c(9,17,26,2,16)
69
70 # Wiederhole fuer jede einzulesende Datei
71 for (k in 1:SetNumber){
72
73 # Gib Dateinamen aus
74 print(datapaths[k])
75 # Lese eine einzelne Textdatei ein. Spezielles Datenformat erforderlich: mit Kopfzeile
76
77 # Trennung der Felder mit "/", Textindikator " "
78 zhlp=read.table(paste(st,datapaths[k],sep=""),sep=" ",header=T,as.is=T,quote=" ")
79 # Kette Datensatze aneinander, aber nur die Spalten fff
80 z=rbind(z,zhlp[,fff])
81 #z=rbind(z,data.frame(DT=zhlp$DT,NR=zhlp$NR,PY=zhlp$PY,AU=zhlp$AU,CR=zhlp$CR))
82 }
83
84 # Reduziere Menge der Datensatze auf "Artikel", d.h. entferne alle Dokumente,
85 # die nicht Artikel sind, und entferne Artikel ohne Referenzen
86 z=z[z$DT=="Article" & z$NR>0,]
87
88 # Initialisiere Vektoren zur Aufnahme der Entropie- und Jahresdaten
89 # Vec: Vektor zur Aufnahme der Entropiewerte aus den einzelnen Stichproben (Achtung
90 # mehrere Stichproben pro Jahrgang)
91 # Yvec: Jahreszahlen
92 # Rvec: Vektor zur Aufnahme der mittleren Anzahl von Referenzen in einer Stichprobe pro
93 # Artikel (mehrfache Referenzen zaehlen mehrfach)
94 # Uvec: Vektor zur Aufnahme der mittleren Anzahl von unterschiedlichen Referenzen in
95 # einer Stichprobe pro Artikel (mehrfache Referenzen zaehlen einfach)
96 Vec=c(); Yvec=c(); Rvec=c(); Uvec=c()
97
98 # Festlegung der zu analysierenden Zeitspanne
99 timespan=1986:2006
100
101 # Umfang und Anzahl der Stichproben, die aus einem Jahrgang gezogen werden
102 # SampleN: Umfang einer Stichprobe (soviele Artikel werden gezogen)
103 # SampleS: Anzahl der Stichproben (so oft wird gezogen)
104 # SampleRef: nicht verwendet
105 # Achtung: Anzahl der Artikel in einem Jahrgang muss groesser sein als SampleN
106 SampleN=500
107 SampleS=50
108 #SampleRef=SampleN*15
109
110 # Lege Datei fest, die Histogramme der Referenzenverteilung in einem Jahrgang aufnimmt
111 #jpeg(filename = paste(st,"images\\EntHistTLMH80Sel","-",SampleN,"-",SampleS,".jpg",sep
112 # "="), width = 640, height = 1000, pointsize = 12, quality = 100, bg = "white")
113 #par(mfrow=c(4,3))
114
115 # Analyse der Jahrgaenge
116 for (k in (timespan)){
117
118 # postscript("/home/mitesser/images/ScreeAlleLCHEM. -",SampleN,"-",SampleS,"-",k,".eps
119 # ",sep=""), width = 8, height = 4)
120 # par(mfrow=c(1,5))
121
122 # Waehle alle Artikel aus Jahrgang k aus
123 z0=z[z$PY==k,]
124 print(paste("nr.of art.in",k,":",length(z0$AU)))
125
126 #rnksum=0
127
128 # Wiederhole Stichproben
129 for (huz in (1:SampleS)){
130
131 # Ziehe Stichprobe von Artikeln aus dem gesamten Jahrgang
132 z1=z0[sample(1:length(z0$AU),SampleN),]
133
134 # Bestimme Anzahl der Dokumenten im aktuellen Jahrgang
135 # Hier sollte SampleN rauskommen
136 n=length(z1$AU);# print(c("number of articles: ",n))
137
138 # Erzeuge Dataframe mit aufgetrennten Referenzenlisten
139 # RefList: Dataframe zur Aufnahme der einzelnen Referenzenstrings aus allen Artikeln
140 # (evtl. doppelte Referenzen!)
141 RefList=c();
142 # Wiederhole fuer alle Artikel
143 for (i in 1:n) {
144 # Referenzenfeld ($SCR) wird bei den Strichpunkten zertrennt
145 y=split(z1[i,]$SCR,split=";");

```

```

139   # Kette Referenzen aneinander
140   RefList=c(RefList,y)
141 }
142
143 # Verwandte Dataframe in Vektor mit allen im Datensatz verwendeten Referenzen
144 TList=c(); for (i in 1:n) TList=c(TList,RefList[[i]])
145
146 # Sublist: im Moment nicht verwendet.
147 #SubList=TList[sample(1:length(TList),SampleRef)]
148
149 # Erzeuge Referenzenvektor, der alle verwendeten Referenzen
150 # genau einmal enthaelt.
151 # DList: Vektor mit Referenzen ohne Duplikate
152 DList=sort(unique(TList))
153 #DList=DList[sample(1:length(DList),SampleRef)]
154
155 # Bestimme Anzahl der unterschiedlichen Referenzen
156 N0=length(DList)
157
158 # Lege Refenz-Dokumentmatrix an. Dokumente sind hier noch Zeilen.
159 # m: Matrix mit Artikel Daten
160 m=matrix(rep(0,n*N0),nrow=n,ncol=N0)
161 for (i in 1:n) m[i,which(DList %in% RefList[[i]])]=1
162
163 # transponiere Matrix, so dass ein Dokument einer Spalte entspricht
164 m=t(m)
165
166 # Bestimme zufaellig Indexmenge der zu loeschenden Matrixeintraege
167 # (Normierung)
168 #Nullen=sample(1:sum(m),sum(m)-SampleRef)
169
170 # Uebertrage Indexmenge auf Matrix
171 # h=which(m==1,arr.ind=T)
172
173 # Loesche Eintraege aus Matrix
174 # m[cbind(h[Nullen,1],h[Nullen,2])]=0
175
176 # entferne 0 Zeilen aus der Matrix
177 # sollte es eigentlich gar nicht mehr geben
178 m=m[apply(m,MARGIN=1,sum)!=0,]
179
180 # Kern der Berechnung: Fuehre Singulaerwertzerlegung durch
181 g=fast.svd(m)
182 # if (huz < 6) plot(g$d^2,xlab="Index des quad. Singulaerwerts", ylab="quad.
183   Singulaerwert")
184 #rnksum=rnksum+sum(g$d>0.0000001)
185
186 # Bestimme Summe der Singulaerwerte, noch ueberprufen: Vorzeichen
187 # SVSum: Summe der Singulaerwerte
188 SVSum=sum(g$d^2)
189
190 # Bestimme Entropie aus Singulaerwerten
191 # ent: Entropie
192 ent=- sum((g$d^2/SVSum)*log2(g$d^2/SVSum))
193
194 # Maximale Entropie
195 # entmax: maximale Entropie
196 entmax=log2(length(g$d))
197
198 # Haenge aktuellen Entropiewert an Entropienvektor an, und aktuellen
199 # Jahreswert an Jahresvektor
200 #Evec=c(Evec,ent/entmax)
201 Evec=c(Evec,ent); Yvec=c(Yvec,k); Rvec=c(Rvec,sum(m)/SampleN); Uvec=c(Uvec,length(m)
202   [,1])/SampleN)
203
204 # grafische Ausgabe
205 # plot(g$d^2,main=datapaths[k],sub=paste("Spektrum fr ",n,"Doks, Entropie = ",format
206   (ent,digits=3)),ylab="Eigenwertwert",xlab="Index")
207 }
208 # hist(z0$NR, main=paste(k," Mittl. Rang: ", rnksum/SampleS), sub=paste("Mittl. Anz.
209   Ref.:", mean(z0$NR)), xlab="Referenzenanzahl")
210 #dev.off()
211 }
212
213 # Initialisiere Vektoren fuer mittlere Entropie (Mittelwerte aus Stichproben!) und
214 # zugehoerige Standardabweichungen
215 mEvec=c(); sEvec=c(); mRvec=c(); mUvec=c()
216
217 # Bestimme Mittelwert und Standardabw. fuer die einzelnen Jahre
218 for (j in 1:(length(timespan))) {
219   mEvec=c(mEvec,mean(Evec[((j-1)*SampleS)+1):(j*SampleS)))

```

```

218     sEvec=c(sEvec, sd(Evec[(((j-1)*SampleS)+1):(j*SampleS)]))
219     mRvec=c(mRvec, mean(Rvec[(((j-1)*SampleS)+1):(j*SampleS)]))
220     mUvec=c(mUvec, mean(Uvec[(((j-1)*SampleS)+1):(j*SampleS)]))
221   }
222
223 # Bibliothek mit Plotbefehl fuer Mittelwert und Standardabweichung
224 library(gplots)
225
226 # Ergebnisdaten fuer spaeteren Gebrauch in Textdatei schreiben.
227 erg=data.frame(mEvec=mEvec, sEvec=sEvec, mRvec=mRvec, mUvec=mUvec)
228 write.table(erg, "/home/mitesser/images/zeitreihe.txt")
229 write.table(data.frame(SampleN=SampleN, SampleS=SampleS, entmax=entmax, tmin=timespan[1],
230                       tmax=timespan[length(timespan)]), "/home/mitesser/images/zeitreihe.para.txt")
231
232 # Wiedereinlesen der Daten.
233 # Dies hier nur der Vollstaendigkeit halber, so dass dieser letzte Abschnitt
234 # zum Erzeugen der Abbildung auch unabhaengig vom Rest genutzt werden kann
235 # wenn man z.B. nur die Abbildung neu machen will, ohne nochmal rechnen zu muessen
236 param=read.table("/home/mitesser/images/zeitreihe.txt", header=T)
237 param=read.table("/home/mitesser/images/zeitreihe.para.txt", header=T)
238
239 # Abbildung
240 postscript("/home/mitesser/images/zeitreihe.eps", width = 8.80, height = 5.80)
241 par(mfrow=c(1,2))
242 timespan=param$tmin:param$tmax
243 plotCI(timespan, erg$mEvec, uiw=erg$sEvec, xlab="Jahre", ylab="Entropie", main=paste(c("
244   Stichprobenumfang: ", param$SampleN)), sub=paste("Max. Ent. = ", format(param$entmax,
245   digits=3)), cex.sub=0.7)
246 plot(timespan, erg$mRvec, xlab="Jahre", ylab="Mittlere Anzahl von Ref. pro Artikel")
247 points(timespan, erg$mUvec, pch=2)
248 dev.off()

```

Listing 2: Quellcode für PLSA-Auswertung (exemplarisch für Abbildung 3.17)

```

1 # Funktion zur Berechnung der PLSA
2 PLSA=function(m, NTops){
3 # m: uebergebene Referenz-Dokument-Matrix
4 # NTops: uebergebene Anzahl an Themen
5
6 # Additionskonstante zur Vermeidung von Division durch Null
7 eps=0.00000001
8
9 # akzeptierte Differenz fuer Konvergenz
10 diffmax=0.00001
11
12 # Initialisiere Vektor fuer Entropiewerte
13 ents=c()
14
15 # Initialisiere Vektor fuer Log_likelihooods zur abschliessenden Wahl des Maximums
16 LLs=c()
17
18 # Bestimme Anzahl der Dokumente aus Referenz-Dokument-Matrix
19 NDocs=length(m[,1])
20
21 # Bestimme Anzahl der Referenzen aus Referenz-Dokument-Matrix
22 NRefs=length(m[,1])
23
24 # Iteration zur Bestimmung von lokalen Maxima
25 for (t in 1:50){
26
27 #Initialisiere Matrizen fuer die bedingten Wahrscheinlichkeiten:
28 # P1 Relevanz einer Referenz in einem Thema;
29 # P2 Relevanz eines Themas fuer ein Dokument
30 P1=matrix(runif(NRefs*NTops), nrow=NRefs, ncol=NTops)
31 P1=P1/matrix(apply(P1, MARGIN=2, sum), ncol=NTops, nrow=NRefs, byrow=T)
32 P2=matrix(runif(NDocs*NTops), nrow=NTops, ncol=NDocs)
33 P2=P2/matrix(apply(P2, MARGIN=2, sum), ncol=NDocs, nrow=NTops, byrow=T)
34
35 # Verwende EM-Algorithmus fuer PLSA Analyse
36 # Initialisiere diff: Differenz zw. alten und neuen Werten zur Kontrolle auf
37 # Konvergenz
38 diff=1
39
40 # Iteriere bis Konvergenz erreicht
41 while (diff>diffmax){
42   P1old=P1
43   P2old=P2
44   P2=P2*(t(P1)%*(m+eps)/(P1%*P2+eps))
45   P2=P2/matrix(apply(P2, MARGIN=2, sum), ncol=NDocs, nrow=NTops, byrow=T) # Normierung
46   P1=P1*((m+eps)/(P1%*P2+eps))%*t(P2)

```

```

46     P1=P1/matrix(apply(P1,MARGIN=2,sum),ncol=NTops,nrow=NRefs,byrow=T) # Normierung
47     diff=max(sum(abs(P1old-P1)),sum(abs(P2old-P2)))
48 }
49
50 # Bestimme Entropie
51 v=apply(P2,MARGIN=1,sum);v=v/sum(v); D=-sum(v*log2(v)); #print(D)
52
53 # Bestimme maximale Entropie
54 entmax=log2(NTops)
55
56 # Kette Entropiewerte aneinander
57 ents=c(ents,D)
58
59 # Bestimme loglikelihood LL0
60 LL0=0; for (j in (1:NDocs)) for (i in (1:NRefs)) LL0=LL0+m[i,j]*log2(sum(P1[i,]*P2[,j]
61 ))
62
63 # Kette alle LL0 zusammen
64 LLS=c(LLs,LL0)
65 #print(LL0)
66
67 # Suche Maximum durch Vergleich der Log-Likelihood-Werte
68 if (t==1 | is.nan(LLX)==T) {P1S=P1;P2S=P2;DS=D;LLX=LL0} else if (is.nan(LL0)==F & is.
69     nan(LLX)==F & LL0>LLX){P1S=P1;P2S=P2;DS=D;LLX=LL0}
70 # Set der Maximalwerte: P1S, P2S, DS, LLX
71 }
72 # Folgende Liste wird von Funktion zurueckgegeben
73 list(P1S,P2S,DS,LLX,LLs,ents,entmax)
74 }
75
76 # Lokalen Datenpfad definieren;
77 # muss auf anderen Rechnern entsprechend angepasst werden.
78 # st="I:\\Daten\\Bibliothek\\Masterarbeit\\data\\"
79 # st="/home/mitesser/data/szientometrie/"
80 # st="/media/TREKSTOR1/DateDsn/Bibliothek/Masterarbeit/data/"
81
82 # Auswahl des Datensets
83 # 1 = Szientometrie
84 # 2 = Elektrochemie
85 SelDataSet=1
86
87 # Dateinamen, die geladen werden sollen, definieren (Dateien muessen im eben
88 # definierten Verzeichnis liegen)
89 if (SelDataSet==1){
90     st="/home/mitesser/data/szientometrie/"
91     # st="/media/TREKSTOR1/Daten/Bibliothek/Masterarbeit/data/"
92     datapaths=c(
93         "STM80-06.txt",
94         "IPM62-07.txt",
95         "JOD73-07.txt",
96         "JIS68-07.txt",
97         "JAS56-07.txt"
98     )
99 }
100 if (SelDataSet==2){
101     st="/home/mitesser/data/elektrochemie/"
102     datapaths=c(
103         "ElChemXXBAB.txt",
104         "ElChemXXBIO.txt",
105         "ElChemXXCVD.txt",
106         "ElChemXXCOS.txt",
107         "ElChemXXELA.txt",
108         "ElChemXXELC.txt",
109         "ElChemXXECA1.txt",
110         "ElChemXXECA2.txt",
111         "ElChemXXAEC.txt",
112         "ElChemXXJEC1.txt",
113         "ElChemXXJEC2.txt",
114         "ElChemXXJEC3.txt",
115         "ElChemXXJES1.txt",
116         "ElChemXXJES2.txt",
117         "ElChemXXJES3.txt",
118         "ElChemXXJES4.txt",
119         "ElChemXXJES5.txt",
120         "ElChemXXPSF.txt",
121         "ElChemXXRJE.txt",
122         "ElChemXXSAC.txt",
123         "ElChemXXSSI1.txt",
124         "ElChemXXSSI2.txt",
125         "ElChemXXJPS1.txt",
126         "ElChemXXJPS2.txt"

```

```

127 )
128 }
129
130 # Stichprobenumfang innerhalb eines Jahrgangs
131 SampleN=100
132
133 # Anzahl der gezogenen Stichproben
134 SampleS=20
135
136 # Lege Anzahl der Topics=Themen fest
137 NTops=10
138
139 # Ausgewerteter Zeitraum
140 timespan=1986:2006
141
142 # Anzahl der zu verarbeitenden Dateien bestimmen
143 SetNumber=length(datapaths)
144
145 # Variable zur Datenaufnahme leeren (falls vorher besetzt)
146 z=c()
147 zhlp=c()
148
149 # Auswahl der Spalten in den Textdateien, die verwendet werden sollen.
150 # Die anderen Spalten brauchen nicht im Arbeitsspeicher gehalten werden
151 fff=c(9,17,26,2,16)
152
153 # Wiederhole fuer jede einzulesende Datei
154 for (k in 1:SetNumber){
155
156 # Gib Dateinamen aus
157 print(datapaths[k])
158 # Lese eine einzelne Textdatei ein. Spezielles Datenformat erforderlich: mit Kopfzeile
159
160 # Trennung der Felder mit "/", Textindikator " "
161 zhlp=read.table(paste(st,datapaths[k],sep=""),sep="|",header=T,as.is=T,quote=" ")
162 # Kette Datensatze aneinander, aber nur die Spalten fff
163 z=rbind(z,zhlp[,fff])
164 # z=rbind(z,data.frame(DT=zhlp$DT,NR=zhlp$NR,PY=zhlp$PY,AU=zhlp$AU,CR=zhlp$CR))
165 }
166
167 # Reduziere Menge der Datensatze auf "Artikel", d.h. entferne alle Dokumente,
168 # die nicht Artikel sind, und entferne Artikel ohne Referenzen
169 z=z[z$DT=="Article" & z$NR>0,]
170
171 Ds=c()
172 # Analyse der Jahrgaenge
173 for (k in (timespan)){
174   DInYear=c()
175
176 # Waehle alle Artikel aus Jahrgang k aus
177 z0=z[z$PY==k,]
178 print(paste("nr.ofart.in",k,":",length(z0$AU)))
179
180 # Wiederhole Stichproben
181 for (huz in (1:SampleS)){
182   print(paste("huz",huz))
183
184 # Ziehe Stichprobe von Artikeln aus dem gesamten Jahrgang
185 z1=z0[sample(1:length(z0$AU),SampleN),]
186
187
188 n=length(z1$AU);# print(c("number of articles: ",n))
189
190 # Erzeuge Dataframe mit aufgetrennten Referenzenlisten
191 # RefList: Dataframe zur Aufnahme der einzelnen Referenzenstrings aus allen Artikeln
192 # (evtl. doppelte Referenzen!)
193 RefList=c();
194 # Wiederhole fuer alle Artikel
195 for (i in 1:n) {
196   # Relenzenfeld ($CR) wird bei den Strichpunkten zertrennt
197   y=split(z1[i,]$CR,split=";");
198   # Kette Referenzen aneinander
199   RefList=c(RefList,y)
200 }
201
202 # Verwandte Dataframe in Vektor mit allen im Datensatz verwendeten Referenzen
203 TList=c(); for (i in 1:n) TList=c(TList,RefList[[i]])
204
205 # Erzeuge Referenzenvektor, der alle verwendeten Referenzen
206 # genau einmal enthaelt.
207 # DList: Vektor mit Referenzen ohne Duplikate
208 DList=sort(unique(TList))

```

```

208 #DList=DList[sample(1:length(DList),SampleRef)]
209
210 # Bestimme Anzahl der unterschiedlichen Referenzen
211 N0=length(DList)
212
213 # Lege Refenz-Dokumentmatrix an. Dokumente sind hier noch Zeilen.
214 # m: Matrix mit Artikeldaten
215 m=matrix(rep(0,n*N0),nrow=n,ncol=N0)
216 for (i in 1:n) m[i,which(DList %in% RefList[[i]])]=1
217
218 # transponiere Matrix, so dass ein Dokument einer Spalte entspricht
219 m=t(m)
220
221 # entferne 0 Zeilen aus der Matrix
222 # sollte es eigentlich gar nicht mehr geben
223 m=m[apply(m,MARGIN=1,sum)!=0,]
224
225 # nur Umbenennungen
226 NRefs=N0
227 NDocs=n
228
229 # Rufe PLSA Algorithmus auf
230 RST=PLSA(m,NTops)
231
232 # Kette Entropiewerte aneinander
233 DInYear=c(DInYear,RST[3][[1]])
234 }
235 # Kette Entropievektoren aneinander
236 Ds=cbind(Ds,DInYear)
237 }
238
239 # Grafikausgabe
240 library(gplots)
241 pdf(paste("/home/mitesser/zeitreihePLSA",NTops,"-",SelDataSet,"Themes.eps",sep=""),
242     width = 8.80, height = 5.80)
243 mdata=apply(Ds,MARGIN=2,mean)
244 sdata=apply(Ds,MARGIN=2,sd)
245 #plot(timespan,mdata,sub=log2(NTops),xlab="Jahre",ylab="Entropie")
246 plotCI(timespan,mdata,uiw=sdata,xlab="Jahre",ylab="Entropie",cex.sub=0.7)
247 dev.off()

```

Dank

Ich möchte mich ganz herzlich bei Dr. Frank Havemann und Dipl. Math. Michael Heinz bedanken. Sie haben mir die Möglichkeit gegeben, diese Arbeit anzufertigen und standen mir stets mit Rat und Tat zur Seite.

Mein herzlicher Dank gilt auch Frau Dr. Ata Kaban, die mir als Expertin eine Reihe von Emails zur PLSA beantwortet hat.