SSOAR

Open Access Repository
www.ssoar.info

Autorennetzwerke: Verfahren der Netzwerkanalyse als Mehrwertdienste für Informationssysteme

Mutschke, Peter

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Mutschke, P. (2004). *Autorennetzwerke: Verfahren der Netzwerkanalyse als Mehrwertdienste für Informationssysteme.* (IZ-Arbeitsbericht, 32). Bonn: Informationszentrum Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-50747-9

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

GESIS

IZ InformationsZentrum Sozialwissenschaften

IZ-Arbeitsbericht Nr. 32

Autorennetzwerke: Verfahren der Netzwerkanalyse als Mehrwertdienste für Informationssysteme

Peter Mutschke

April 2004

InformationsZentrum
Sozialwissenschaften

Lennéstraße 30
D-53113 Bonn
Tel.: 0228/2281-0
Fax.: 0228/2281-120
E-Mail: iz@bonn.iz-soz.de
Internet: http://www.gesis.org/IZ/index.htm

ISSN: 1431-6943

Herausgeber: Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)

Druck u. Vertrieb: Informationszentrum Sozialwissenschaften, Bonn
Printed in Germany

Das IZ ist Mitglied der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS).
Die GESIS ist Mitglied der Leibniz-Gemeinschaft.

Inhalt

1 Einleitung 4
2 Autorennetzwerke und die Theorie sozialer Netzwerke 7
3 Methodische Grundlagen 10
3.1 Soziale Netzwerke 10
3.2 Zentralität 11
3.3 Skalierung von Autorennetzwerken 14
3.3.1 k-cores 14
3.3.2 m-paths 15
3.4 Propagierung von Autorennetzwerken 16
3.4.1 Propagierung auf der Basis von Resultsets 17
3.4.2 Propagierung persönlicher Netzwerke von Autoren 19
4 Eigenschaften von Autorennetzwerken 22
4.1 Evolution und Dynamik 22
4.2 Topologie 26
4.3 Zusammenfassung und Schlussfolgerungen 29
5 Nutzung von Autorennetzwerken in Informationssystemen 31
5.1.1 Ranking von Dokumenten nach Akteurszentralität 31
5.1.1.1 Ex-Post-Ranking 34
5.1.1.2 Ex-Ante-Ranking 36
5.1.2 Suche nach zentralen Akteuren (Expertensuche) 37
5.1.2.1 Zentrale Akteure in einer Dokumentenmenge 37
5.1.2.2 Zentrale Akteure im Netzwerk eines Autors 38
5.1.3 Alerting mit Autorennetzwerken 42
6 Heuristische Evaluation von Autorennetzwerk-Retrievalmodellen 43
7 Zusammenfassung und Ausblick 45
8 Literatur 47

1 Einleitung

Virtuelle Bibliotheken enthalten eine Fülle an Informationen, die in ihrer Vielfalt und Tiefe von Standardsuchmaschinen nicht erschöpfend erfasst wird. Traditionelle Retrievalsysteme sind in der Regel strikt dokumentorientiert. Der Informationsgehalt der Ergebnisse wird auf die „sichtbaren“ Teile einzelner Dokumente reduziert. Dem Benutzer ist es nicht möglich, die volle Komplexität der gespeicherten Information zu explorieren. Bibliographische Daten, zum Beispiel, bieten reichhaltige Informationen über die Entwicklung und Struktur einer wissenschaftlichen Community. Die Basisinformationen, wie z.B. Koautorenschaften, sind in den bibliographischen Datensätzen zwar enthalten, Dokumentgrenzen überschreitende Zusammenhänge werden von traditionellen Informationssystemen jedoch nicht erkannt und dem Benutzer somit auch nicht zugänglich gemacht. Dies betrifft v.a. Linkstrukturen zwischen Wissenschaftlern, die z.B. in Koautoren- oder Zitationsrelationen repräsentiert sind, insbesondere aber globalere Eigenschaften der Akteure, wie deren strategische Position in wissenschaftlichen Kommunikations- und Kooperationsstrukturen.

Ein weiteres Problem ist das kontinuierliche Wachstum an online verfügbarer Information, das in verstärktem Maße dazu führt, dass der Benutzer immer mehr mit irrelevanten Informationen überhäuft wird, sich andererseits aber relevante Informationen über heterogene Datenquellen und –services verteilen (wie bibliographische Nachweisdatenbanken, Zitationsnachweise, Volltextdienste usw.).

Virtuelle Bibliotheken sind daher nur dann sinnvoll nutzbar, wenn sowohl hochwertige Suchservices, die vorhandene Informationsstrukturen voll ausschöpfen, als auch Dienste bereitgestellt werden, welche die Fülle und Komplexität der in Datenbanken abgelegten Informationen auf hochrelevante Objekte reduzieren. Ein Benutzer, der sich über ein bestimmtes wissenschaftliches Thema informieren möchte, wird nicht nur nach Literatur(nachweisen) suchen, sondern sich auch für menschliche Experten und die sozio-kognitive Struktur des Gebietes interessieren. Die zunehmende Informationsflut einerseits und die lauter werdende Forderung der Benutzer nach qualitativ hochwertigen Informationen andererseits (Krause 2003) legen daher die Entwicklung von Retrieval- und Analyseverfahren nahe, die über die herkömmlichen Retrievalmodelle hinausgehen.

Der vorliegende Bericht informiert über Entwicklungen am IZ, die darauf abzielen, Wissen über das Interaktionsgeschehen in wissenschaftlichen Communities und den sozialen Status ihrer Akteure für das Retrieval auszunutzen. Grundlage hierfür sind *soziale Netzwerke*, die sich durch Kooperation der wissenschaftlichen Akteure konstituieren und in den Dokumenten der Datenbasis z.B. als Koautorbeziehungen repräsentiert sind. Diese Netzwerke werden im folgenden *Autorennetzwerke* genannt. Kernanliegen der am IZ entwickelten Autorennetzwerkmodelle ist die Suche nach **Experten** und das **Ranking** von Dokumenten auf der Basis von Akteurszentralität, d.h. Autoren, deren strategische Positionierung in ihren sozialen Netzwerken auf eine besondere Relevanz des Autors schließen lässt. Hierfür wurden datenbankbasierte Komponenten in Java entwickelt, die Autorennetzwerke und Akteurszentralität im online-Zugriff auf (relationale) Datenbanken auf der Basis einer Recherche berechnen.

Die Bemühungen am IZ gehen auf das Projekt AKCESS zurück, in dem netzwerkanalytische Verfahren für die Suche nach Experten, unseres Wissens erstmalig, für Retrievalzwecke erprobt wurden (Mutschke 1994, 1996). Die im AKCESS-Projekt begonnenen Arbeiten wurden in dem Projekt DAFFODIL[1] weiterentwickelt und dort erstmals in einer größeren Retrievalumgebung eingesetzt (Mutschke 2001, Fuhr et al. 2002). Gegenwärtig werden sie in den Informationsverbund infoconnex[2] (Ballay et al. 2004) integriert, wo sie als zusätzliche Mehrwertdienste z.B. für das Dokumentenranking genutzt werden sollen.

Die Verfahren sind unter Verwendung genereller Konzepte der Graphentheorie und der Theorie sozialer Netzwerke und teilweise deren Weiterentwicklung entstanden. Sie wurden exemplarisch am Beispiel von Koautoren-Netzwerken entwickelt, d.h. Autorennetzwerken, die sich auf der Basis von Koautorenschaften konstituieren. Sie sind aber auf ohne weiteres auf andere Kooperationsbeziehungen (wie z.B. Ko-Projektmitarbeiter-Relationen) sowie andere Arten vernetzter Strukturen anwendbar (wie z.B. Institutionen-, Zitations- oder Begriffsnetzwerke).

Die Relevanz von Autorennetzwerken und Akteurszentralität für Retrievalzwecke konnte in einer Reihe von empirischen Studien belegt werden. Zwei am IZ durchgeführte Studien unter Verwendung von AKCESS und Cognitive-Mapping-Verfahren wiesen einen starken statistischen Zusammenhang zwi-

---

[1] www.daffodil.de

[2] www.infoconnex.de

schen der Zentralität von Themen in Cognitive Maps und der von Akteuren in Koautoren-Netzwerken nach. Die Studien zeigen, dass Mainstream-Themen eines Forschungsfeldes in starkem Maße von zentral positionierten Akteuren besetzt werden (Mutschke & Renner 1995), wohingegen Innovationen eher von Akteuren mittlerer Zentralität auszugehen scheinen (Mutschke & Quan Haase 2001). Die Aussagekraft wissenschaftlicher Kooperationsnetzwerke wurde darüber hinaus in einer Reihe szientometrischer Studien untersucht. Stellvertretend seien hier die neueren Studien von Güdler (2003), Newman (2001a-c, 2004) und Barabasi et al. (2002) genannt.

Graphentheoretische Konzepte wurden bisher nur in sehr eingeschränktem Umfang in Informationssystemen eingesetzt. Zu nennen sind hier insbesondere die Referralsysteme, wie z.B. REFERRALWEB (Kautz et al. 1997), wo Empfehlungspfade von einem Benutzer (oder Ausgangsakteur) zu einem menschlichen Experten in einer gegebenen Netzwerkstruktur evaluiert werden. Diese Systeme beschränken sich jedoch auf die Analyse von lokalen Akteursbeziehungen und ziehen die Eingebettetheit der Akteure in die Gesamtstruktur nicht in Betracht. Ähnliches gilt für die bei Kleinberg (1999) und in GOOGLE[3] verwendeten Zentralitätskonzepte. Beide Systeme bestimmen die Relevanz von Web-Seiten aufgrund ihrer Vernetztheit. Allerdings schöpfen sowohl Kleinbergs Hubs und Authorities als auch Google's PageRank nicht die volle Netzwerkstruktur aus, sondern betrachten lediglich die Zahl der direkten Nachbarn eines Knotens im Netzwerk und reduzieren Zentralität damit auf ein rein lokales Attribut (vgl. Brandes & Cornelsen 2003).

Der Ansatz der vom Autor entwickelten Autorennetzwerk-Komponenten ist es dagegen, Zentralität über die gesamte Netzwerkstruktur zu evaluieren, um somit zu Aussagen über die globale Bedeutung eines Akteurs für die betrachtete Community zu gelangen. Das theoretische Fundament dazu liefert die *Theorie sozialer Netzwerke* (Kapitel 2). Die methodischen Grundlagen der in diesem Bericht diskutierten Autorennetzwerkmodelle werden in Kapitel 3 beschrieben. Kapitel 4 skizziert generelle Eigenschaften von Autorennetzwerken anhand empirischer Untersuchungen zur Evolution und Topologie von Autorennetzwerken in verschiedenen Forschungsfeldern. In Kapitel 5 werden Szenarios diskutiert, die beschreiben, wie Autorennetzwerke und hier insbesondere das Konzept der Akteurszentralität für die Informationssuche in Datenbanken sinnvoll genutzt werden können. Kapitel 6 beschreibt eine heuristische Evaluation von Autorennetzwerken in Informationssystemen. Der Bericht schließt mit einer Zusammenfassung und einem Ausblick.

---

[3] www.google.de

2 Autorennetzwerke und die Theorie sozialer Netzwerke

In der Soziologie hat sich ein Theoriemodell durchgesetzt, das nicht nur für die Diagnose menschlicher Gesellschaften, sondern – in Verbindung mit der in der Theoretischen Physik formalisierten *Small-World*-Theorie (Watts 1999; Newman 2001a-c, 2004; Barabasi 2002) – mittlerweile sogar für die Erklärung von Ökosystemen und den Ausfall von Power Grids herangezogen wird: Die *Theorie sozialer Netzwerke*. Dieses Theoriemodell macht für die Handlungsmöglichkeiten individueller oder korporativer Akteure vor allem deren *Eingebettetsein* in soziale Kontexte verantwortlich (vgl. v.a. Coleman et al. 1966, Granovetter 1985, Wellman 1988). Diese konstituieren sich aus den sozialen Beziehungen zwischen den Akteuren, und genau sie sind es, so die Theorie, die zur Emergenz sozialer Systeme entscheidend beitragen, d.h. den Kräften sozialen Handelns, die weder von allgemeinen Merkmalen der Gesellschaft (wie z.B. deren Werte und Normen), noch von Eigenschaften der Individuen (Geschlecht usw.) abhängig sind, sondern vielmehr von deren *Interaktion* (vgl. Jansen 2003, 11). Demnach sind auch wissenschaftliche Communities also nicht einfach nur die Summe der Autoren in einem bestimmten Fachgebiet oder die Summe ihrer Publikationen, sondern vielmehr die Summe der sozialen Beziehungen zwischen den Akteuren des Feldes.

Die Theorie sozialer Netzwerke setzt sich damit als *strukturelle Handlungstheorie* des Eingebettetseins grundsätzlich von voluntaristischen und neoklassischen Theorietraditionen ab (Hobbes, Adam Smith, Parsons), die in ihren Akteurmodellen von atomisierten Individuen ausgehen, die ihre Entscheidungen in sozialer Isoliertheit treffen (vgl. Jansen 2003, 18). Ressourcen und Interessen der Akteure hingen vielmehr weitgehend von deren struktureller Einbettung ab (Burt 1982). Um das Handeln von Individuen verstehen und erklären zu können, muss man aus der Perspektive der Netzwerkanalyse also das Ganze untersuchen, d.h. das Netzwerk, in das die Individuen eingebettet sind. Aus dieser Perspektive stehen also nicht die individuellen, sondern die *relationalen* Merkmale der Akteure und ihr sozialer Status in der Gesamtstruktur im Mittelpunkt der Analyse (vgl. Jansen 2003, 18; Wasserman & Faust 1994).

Der Grundansatz der Netzwerkanalyse ist dabei, nicht nur die direkten Beziehungen der Akteure, sondern gerade auch die *indirekten* Beziehungsmuster zu berücksichtigen, um die Eingebettetheit der Akteure zu evaluieren (vgl. ebd.). Denn die sozialen Beziehungen zwischen den Akteuren in einer Gesellschaft sind nicht alle gleichwertig. Es gibt Beziehungen, die im Kontext der globalen Struktur einen "wichtigeren" Stellwert haben als andere, weil sie im Zentrum

des Netzwerkes lokalisiert sind, also dort, wo viele Beziehungen "zusammenlaufen", während andere eher an der Peripherie der Struktur angesiedelt sind. Demnach haben die Mitglieder eines Netzwerkes einen unterschiedlichen sozialen Status in der Gesamtstruktur, je nachdem ob sie eher im Zentrum oder eher in der Peripherie der Struktur positioniert sind. Dieses Modell deckt sich mit unserer Alltagserfahrung: Jeder, der schon einmal eine Familienfeier oder wissenschaftliche Tagung besucht hat, weiß, dass man die "wichtigen Leute" sehr schnell von den "unwichtigen" anhand der Zahl ihrer "Besucher" unterscheiden kann.

Der soziale Status der Akteure in einem Netzwerk ist insofern relevant, als nach Burt (1982) die Akteure die Möglichkeit der Rückwirkung auf die Struktur haben. Burts Mikro-Makro-Modell enthält somit nicht nur eine strukturbezogene Komponente, wonach die durch die Position in der Sozialstruktur geformten Interessen eines Akteurs zwar die *constraints* für dessen Handlungsmöglichkeiten sind. Burts Modell propagiert auch eine *akteurbezogene* Komponente, wonach die Akteure durch ihre Handlungen die Struktur auch reproduzieren und – sofern sie strategisch günstig positioniert sind – sogar modifizieren können. Der Status der Akteure, und hier insbesondere deren zentrale Positionierung in der (entstehenden) Netzwerkstruktur spielt also für deren Evolution eine entscheidende Rolle. Dahinter steht die Annahme, dass zentrale Akteure an vielen Beziehungen im Netzwerk beteiligt, daher einen leichteren Zugang zu Netzwerkressourcen (v.a. Informationen) haben und Interaktionsprozesse in der Community deshalb besser kontrollieren können (Knoke & Burt 1983, vgl. Jansen 2003, 29ff., 127).

Netzwerkpositionen mit hoher Zentralität werden dabei um ihrer selbst Willen zum Ziel von positiven Kooperationsangeboten anderer Akteure (Jansen 2003, 31). Dieses Phänomen, also die Neigung (menschlicher) Akteure primär Beziehungen mit Akteuren einzugehen, die bereits zentrale Positionen in der Sozialstruktur einnehmen, wird *preferential attachment* genannt (Barabasi et al. 2002, Newman 2001c, 2004). Preferential Attachment bewirkt sich selbst verstärkende *Anlagerungsprozesse*, die zu einer „Stratifizierung der Akteure“ führen (Jansen 2003, 31). Preferential Attachment wurde in einigen, v.a. von Newman (2001c, 2004) und Barabasi et al. (2002) für den naturwissenschaftlichen Bereich durchgeführten Studien auch für wissenschaftliche Communities nachgewiesen. Es gibt also so etwas wie eine verstärkte Neigung wissenschaftlicher Akteure, primär mit denen zu kooperieren, die bereits viele Koautoren haben.

Nach Güdler (2003) u.a. spielen diese Akteure in wissenschaftlichen Communities die Rolle *zentraler Vermittler*: Zentrale Vermittler agieren als Wis-

senschaftsmanager, die Wissen selektieren, Spezialistenwissen vernetzen und über das Netzwerk transportieren. Sie manifestieren (neue) Forschungsfelder durch Kooperation mit anderen Experten und tragen umgekehrt, durch Nicht-Kooperation, zur Marginalisierung „unwichtiger" Bereiche bei. Bestätigt werden diese Tendenzen durch Untersuchungen von Mutschke, Renner und Quan Haase, die einen starken statistischen Zusammenhang zwischen der Zentralität von Themen und der von Autoren in Wissenschaftsgemeinschaften nachwiesen (Mutschke & Renner 1995, Mutschke & Quan Haase 2001). Wissenschaftlernetzwerke kanalisieren also in asymmetrischer Weise den Zugang zu hochbewerteten Ressourcen (Informationen) und beeinflussen damit sowohl kooperatives als auch konkurrierendes Verhalten unter wissenschaftlichen Akteuren (vgl. Jansen 2003, 22).

Was dies für die Evolution wissenschaftlicher Communities bedeutet, liegt, zumindest theoretisch, auf der Hand: Soziale Netzwerke sind nicht notwendigerweise auch effizient. Ihr Entstehen und Funktionieren hängt entscheidend von der *Interaktionsorientierung* der beteiligten Akteure ab, die wiederum vom Vertrauen der Akteure untereinander abhängig ist (vgl. Jansen 2003, 12). Güdler (2003) bestätigt diesen Zusammenhang anhand eines Phasenmodells der Evolution von wissenschaftlicher Communities am Beispiel der Sozialwissenschaften: In der Entstehungsphase eines Forschungsfeldes sind die Akteure nicht oder nur schwach vernetzt. Die Akteure sind primär auf die inhaltliche Konzeptualisierung „ihres" neuen Feldes fokussiert und dabei überwiegend auf sich selbst gestellt. In der zweiten Phase der Entwicklung gehen die Akteure erste Kooperationsbeziehungen ein, die sich jedoch noch auf institutionell geprägte Kontexte beschränken. Es entstehen kleinere Forschungscliquen, die sich typischerweise um eine Person herum gruppieren („Schulen"). Um das Forschungsfeld durchzusetzen, müssen die Akteure allerdings Kooperationen mit „entfernteren" Kollegen, sog. *weak ties* (Granovetter 1973) eingehen (vgl. Jansen 2003, 31). In dieser Phase kommt es zu einer zunehmenden Vernetzung des Forschungsfeldes. Es bilden sich sog. *Invisible Colleges* (Crane 1972) heraus, die über Instituts-, Fach- und Ländergrenzen hinausgehen.

Kooperation hat also strukturbildende Eigenschaften. Dem besonderen Stellenwert von Kooperation wird auch auf Seiten der Wissenschaftspolitik Rechnung getragen. Die Sonderforschungsbereiche der Deutschen Forschungsgemeinschaft (DFG) und die Rahmenprogramme der EU z.B. sind auf eine „Zusammenarbeit von Wissenschaftlern im Rahmen eines fächerüber-

greifenden Forschungsprogramms“ (DFG)[4] und eine „Bündelung und Integration der Forschung“ (Sechstes Europäisches Forschungsrahmenprogramm)[5] angelegt.

Es bietet sich an, diese Erkenntnisse auch für die Informationssuche in wissenschaftlichen Literaturdatenbanken zu nutzen. Dies ist das Ziel der hier beschriebenen Autorennetzwerk-Retrievalmodelle. Im folgenden werden ihre methodischen Grundlagen beschrieben.

3 Methodische Grundlagen

3.1 Soziale Netzwerke

Ein soziales Netzwerk (hier: Autorennetzwerk) wird in unserem Modell in Anlehnung an die klassische Graphentheorie (s. Palmer 1985) als ein Graph $G = (V, E)$ beschrieben, der aus einer Menge V von Knoten (*vertices*) und einer Menge E von Kanten (*edges*) besteht. In dem hier beschriebenen Autorennetzwerkmodell werden die Knoten durch Autoren und die Kanten durch Koautorenschaften repräsentiert. Da es sich um ganz generelle Komponenten handelt, können aber auch andere Relationen, z.B. Ko-Mitarbeiter-Beziehungen oder Begriffsbeziehungen, verwendet werden. Abbildung 1 visualisiert einen Graphen mit neun Knoten und zehn Kanten.

Die Struktur eines Netzwerkes erschließt sich über sog. Wege (*walks*) oder Pfade (*paths*), mit denen v.a. die Beziehungen zwischen entfernteren Akteuren beschrieben werden können. Ein *walk* in dem Graph von $s \in V$ nach $t \in V$ ist eine Sequenz von verbundenen Knoten, beginnend mit s und endend in t, so dass s und t verbunden sind. Ein Graph ist *verbunden*, wenn jeder Knoten durch jeden anderen über einen *walk* erreicht werden kann. Ein maximal verbundener Teilgraph wird *Komponente* genannt. Ein *path* ist ein *walk*, in dem alle Knoten und Kanten distinkt sind. Die *Länge* eines Pfades wird durch die Anzahl seiner Kanten bestimmt. Die Nähe zwischen zwei Punkten in dem Graphen wird durch die *kürzeste Pfaddistanz* ausgedrückt. Das ist die Länge des kürzesten Pfades (*geodesic*) zwischen zwei Punkten. Je kürzer die Pfadlänge zwischen zwei Akteuren in einem sozialen Netzwerk, desto weniger störungsanfällig ist ihre Beziehung und desto schneller werden Ressourcen,

---

[4] http://www.dfg.de/forschungsfoerderung/koordinierte_programme/sonderforschungsbreiche/

[5] http://www.rp6.de/inhalte

z.B. Informationen, von einem Akteur zum anderen übertragen.[6] Der kürzeste Pfad z.B. von B nach H in dem Graphen von Abbildung 1 führt über C, E und G und hat die Länge 4.

Auf sozialen Netzwerken können eine Reihe Berechnungen durchgeführt werden. Auf der Ebene des Netzwerkes interessieren hier v.a. dessen Größe und Dichte sowie die Zahl seiner Komponenten. Auf der Ebene des Akteurs ist v.a. dessen *Zentralität* interessant. Die Zentralität eines Akteurs beschreibt dessen strategische Position im Netzwerk, seine „Prominenz“ und seinen Grad an Einfluss im Netzwerk. Die in unserem Autorennetzwerkmodell verwendeten Zentralitätskonzepte werden im folgenden beschrieben.

3.2 Zentralität

Ein wichtiges Strukturmerkmal der Akteure eines Netzwerkes ist der Grad ihrer Zentralität. Zentralität ist ein netzwerkanalytisches Konzept, das den Grad der Eingebundenheit eines Akteurs in einer Sozialstruktur und somit so etwas wie dessen „Prominenz“ in der betrachteten Community beschreibt. Dahinter steht die Annahme, dass zentrale Akteure Zugang zu vielen Informationsquellen haben, deshalb von Innovationen als erste erfahren und zugleich deren Verbreitung aktiv beeinflussen können (vgl. Jansen 2003, 127ff.).

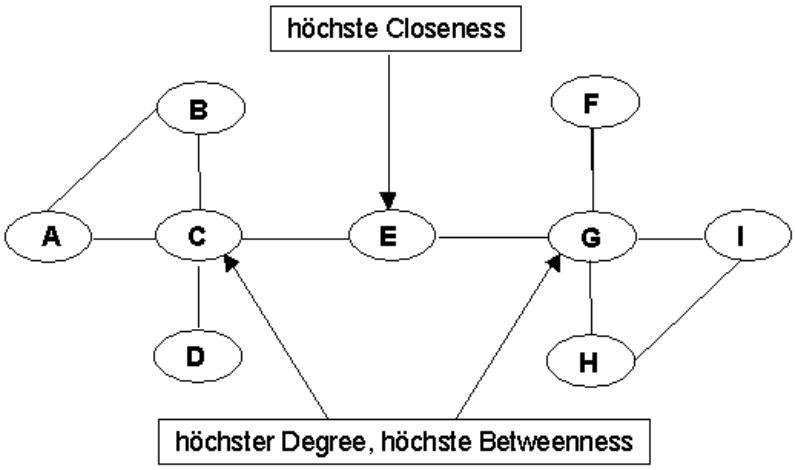

Abb. 1: Zentralität in einem Graphen mit neun Knoten und zehn Kanten

In unserem Modell werden drei grundlegende Zentralitätsmaße aus der Netzwerkanalyse verwendet: Das einfachste Maß zur Charakterisierung der Zentralität eines Akteurs ist die Zahl seiner direkten Nachbarn (*degree centrality*).

---

[6] vgl. Jansen 2003, Wasserman & Faust 1994

Zentral ist nach diesem Zentralitätskonzept der Akteur, der viele direkte Beziehungen hat. Der Grad der Eingebundenheit eines Akteurs erschließt sich aber nicht nur über die Zahl seiner direkten Verbindungen, seines *Degrees*, sondern auch, und vielmehr, über seine *indirekten* Verbindungen. Hier sind zwei Maße relevant, die auf dem Konzept des kürzesten Weges basieren, aber jeweils etwas ganz anderes über die Zentralität eines Akteurs aussagen: *Closeness centrality* misst die Nähe eines Akteurs zu entfernteren Akteuren im Netzwerk[7] und *betweenness centrality* (Freeman 1979) die Zahl der durch einen Akteur verbundenen Akteure[8].

Zentral nach *Closeness* ist der Akteur, der über viele *kurze* Verbindungen zu *allen* anderen Akteuren im Netzwerk verfügt. Ein Closeness-zentraler Akteur ist relativ selten auf die Vermittlung durch andere Akteure angewiesen. Aufgrund seiner kurzen Verbindungen zu anderen kommen Informationen bei ihm ohne große Verzerrungen an. Zentral nach *Betweenness* ist der Akteur, der *zwischen* vielen Akteurspaaren im Netzwerk auf deren kürzesten Verbindungen positioniert ist. Ein Betweenness-zentraler Akteur verbindet also viele Akteure im Netzwerk, wird deshalb häufig von anderen Akteuren als „Makler“ benutzt und kann deshalb viele Aktivitäten im Netzwerk kontrollieren.[9]

In dem Beispielgraphen von Abbildung 1 haben C und G mit jeweils vier direkten Verbindungen den höchsten Degree und zugleich die höchste Betweenness, weil C und G nicht nur A, B und D bzw. F, H und I mit dem Rest des Netzwerkes, sondern auch D mit A und B bzw. F mit H und I verbinden. Knoten E hat die höchste Closeness, weil E die Cliquen um C und G miteinander verbindet und deshalb die kürzesten Verbindungen zu allen Knoten im Netzwerk hat.

---

[7] Summe der Länge der kürzesten Pfade zwischen einem betrachteten Knoten $v \in V$ und allen $t \in V$ in G. Um Closeness auch in unverbundenen Graphen evaluieren zu können, wurden nach Tallberg (2000) die Closeness-Werte mit der Größe der jeweiligen Komponente gewichtet, so dass Knoten in größeren Komponenten einen vergleichsweise höheren Zentralitätswert erhalten.

[8] Summe der kürzesten Pfade zwischen allen $s \in V$ und allen $t \in V$, die den betrachteten Knoten $v \in V$ als Vorgänger- oder Nachfolgerknoten auf dem kürzesten, s und t in G verbindenden Pfad haben. In unserem Modell werden durchweg normalisierte Werte verwendet. Auf die Darstellung des Formelwerks zu diesen Algorithmen sei hier aber verzichtet. Es findet sich in der einschlägigen Literatur, z.B. in dem Standardwerk von Wasserman & Faust (2004). Für eine effiziente Berechnung von Zentralität auch in großen Netzwerken wurde eine Adaption des Betweenness-Algorithmus von Brandes (2001) verwendet, die auch Closeness berücksichtigt.

[9] vgl. Jansen 2003, 131

Es gibt einen wichtigen formalen Unterschied zwischen den drei Zentralitätsmaßen hinsichtlich der Zahl der jeweils betrachteten Knoten: Degree- und Closeness-Zentralität evaluieren jeweils alle Dyaden (Paarbeziehungen) mit dem betrachteten Akteur. Sie unterscheiden sich lediglich darin, ob nur direkte oder auch indirekte Beziehungen berücksichtigt werden. Betweenness hat dagegen eine prinzipiell andere Logik. Es betrachtet jeweils drei Akteure und evaluiert, ob der betrachtete Akteur auf dem kürzesten Pfad zwischen den beiden anderen Akteuren liegt und somit ein Mittler zwischen diesen Akteuren ist. Je häufiger ein Akteur zwischen anderen auf deren geodesics „vermittelt", desto zentraler ist er nach dem Betweenness-Maß.[10]

Mit diesem formalen Unterschied ist eine prinzipiell andere Interpretation des sozialen Status eines Akteurs im Netzwerk verbunden: Degree- und Closeness-Zentralität messen die Unabhängigkeit eines Akteurs von anderen. Ein zentraler Akteur nach Degree oder Closeness ist nicht oder selten auf andere angewiesen, weil er viele direkte bzw. kurze indirekte Beziehungen zu allen anderen Akteuren im Netzwerk unterhält. Betweenness misst dagegen, ob andere Akteure vom betrachteten Akteur abhängig sind, weil dieser als Schnittstelle zwischen vielen anderen Akteuren im Netzwerk fungiert. Betweenness misst somit die Kontrollmöglichkeiten, die dem Akteur aufgrund seiner strategischen Position im Netzwerk zufallen.[11]

Degree gilt daher auch als Maß für die soziale Aktivität eines Akteurs, Closeness als Maß für seine soziale Effizienz (im Sinne von Unabhängigkeit) und Betweenness als Maß für die Kontrolle von sozialen Beziehungen. Degree ist zugleich die Art von Zentralität, die – im Unterschied zu Closeness und Betweenness - für die anderen Akteure im Netzwerk auch am ehesten „sichtbar" ist. Die Zahl der Koautoren eines Wissenschaftlers kann man seinen Veröffentlichungen entnehmen, und in der Regel ist in einer wissenschaftlichen Community auch bekannt, wer mit wem kooperiert (hat). Das Preferential-Attachment-Phänomen bestätigt diese These. Die Closeness und Betweenness eines Akteurs ist dagegen quasi in der Struktur des Netzwerkes „versteckt" und daher ohne technische Hilfsmittel nicht sichtbar. Dies ist aber zugleich der Vorteil dieser Zentralitätsmaße: sie evaluieren die Eingebettetheit eines Akteurs in der *globalen* Struktur eines Netzwerkes, während Degree nur die lokale Bedeutung eines Akteurs misst.

---

[10] vgl. auch Jansen 2003, 134

[11] vgl. auch Jansen 2003, 134

3.3 Skalierung von Autorennetzwerken

Autorennetzwerke können sehr komplex und unübersichtlich sein. Die Reduktion der Komplexität von Autorennetzwerken ist deshalb nicht nur für deren Visualisierung relevant, sondern auch für eine zielgenauere Evaluation der Zentralität der Akteure. Komplexitätsreduktion kann sowohl bei den Akteuren als auch bei den Verbindungen ansetzen, indem „unwichtige“ Akteure oder „unwichtige“ Kanten aus den Netzwerk entfernt werden. Im folgenden werden zwei, in unserem Modell verwendete Verfahren beschrieben.

3.3.1 *k*-cores

Die *k*-core-Technik ist ein akteurbezogenes Reduktionsverfahren. Das Netzwerk wird auf Teilgraphen reduziert, in denen die Mitglieder mindestens *k* direkte Verbindungen haben. Der Parameter *k* gibt also die Mindestzahl der Mitglieder eines Teilgraphen an, die jeder Akteur direkt erreichen kann.[12] Ein 1-core-Graph entspricht demnach dem ursprünglichen Netzwerk. In einem 2-core-Netzwerk sind alle Akteure mit einer Null-Betweenness entfernt, d.h. der Graph enthält keine Akteure mehr, die nur eine Verbindung haben (sog. „hanger-ons“) und somit auch nicht zwischen anderen Akteuren „vermitteln“ können.

Die Methode hat den Vorteil, dass sich die Zentralität der Akteure, v.a. deren Betweenness, bei einem *k*-Wert größer 1 zugunsten von Cutpoint-Akteuren verschiebt, d.h. von Akteuren, die zwischen kohäsiven Subgruppen vermitteln. Zentralität in einem *k*-core-Netzwerk mit einem *k*-Wert größer gleich 2 hebt also tendenziell Akteure hervor, die zwar einen relativ geringen Degree haben, aber dennoch strukturell autonom sind, als Brückenpersonen „entferntere“ Subcommunities miteinander verbinden und somit zum Small-World-Charakter[13] vieler sozialer Netzwerke mehr beitragen als Akteure mit einem hohen Degree. Die 2-core-Version des Graphen in Abbildung 1 z.B. wäre um die Knoten D und F reduziert und die Betweenness würde sich zugunsten von E, dem Cutpoint-Akteur, verschieben.

Die *k*-core-Methode (und andere Cliquen-Analyse-Verfahren) sind akteurbezogene Skalierungsverfahren. Sie reduzieren das Netzwerk auf „wichtigere“ Akteure. Im folgenden wird ein *kantenbezogenes* Skalierungsverfahren vorge-

---

[12] Vgl. Wasserman & Faust 1994, Jansen 2003, 99f.

[13] s. Kapitel 4.2

stellt, das nicht „unwichtige“ Akteure, sondern „unwichtige“ Kanten eliminiert: das vom Autor vorgeschlagene Main-Paths-Modell (Mutschke 2001, 2003).

3.3.2 *m*-paths

Das Main-Paths-Modell (Mutschke 2001, 2003) [14], in der vom Autor vorgeschlagenen Version im folgenden *m*-paths genannt, geht davon aus, dass nicht alle Beziehungen in einem Netzwerk gleichwertig sind. Für einen Wissenschaftler ist die Kooperation mit einem „prominenten“ Autor seines Faches sicherlich wichtiger als die mit einem „gleichrangigen“ Kollegen. Das Preferential-Attachment-Phänomen (s. Kapitel 2) bestätigt diese Hypothese. Es liegt daher nahe, die Komplexität eines Autorennetzwerkes dadurch zu reduzieren, dass nur Kanten betrachtet werden, die zu zentralen Akteuren führen.

Das *m*-paths-Verfahren reduziert die Zahl der von einem Akteur ausgehenden Verbindungen auf die *m* „besten“ Beziehungen. Welches die „besten“ Kanten sind, ermittelt eine Evaluationsfunktion, die - dem Preferential-Attachment-Mechanismus folgend - dem Degree der jeweiligen Ko-Akteure entspricht. Die Beziehungen eines Akteurs werden also daraufhin evaluiert, ob sie zu Ko-Akteuren mit hohem Degree führen. Unter allen Beziehungen eines Akteurs werden dann diejenigen ausgewählt, welche zu den *m* degree-zentralsten Ko-Akteuren des betrachteten Akteurs führen. Diese „besten“ Verbindungen heißen *main paths*, in Anlehnung an die in der Zitationsanalyse gebräuchliche *Main Paths Analysis*, welche die von einem Artikel ausgehenden Hauptzitationspfade (*main paths*) beschreibt (Hummon & Doreian 1989, Carley et al. 1993).

Der Parameter *m* (für *main*) definiert den Mindestrang, gemessen an der Höhe der Degrees unter den Ko-Akteuren eines betrachteten Akteurs, den Ko-Akteure haben müssen, damit die Verbindung zu ihnen in das Netzwerk aufgenommen wird. Ein 1-path-Netzwerk enthält also nur Kanten zu dem Ko-Akteur (oder den Ko-Akteuren) eines betrachteten Akteurs, der unter allen seinen Ko-Akteuren den höchsten Degree aufweist. Ein 2-paths-Netzwerk „akzeptiert“ auch Verbindungen zu Ko-Akteuren mit zweithöchstem Degree usw. Der *m*-Wert entspricht also (bei ungleichen Degrees) der maximalen Zahl an „besten“ Ko-Akteuren, mit denen ein Akteur in einem *m*-paths-Netzwerk verbunden ist. Haben zwei Ko-Akteure den gleichen Degree-Wert, werden beide Verbindungen zu diesen Ko-Akteuren in das Netzwerk aufge-

[14] Dort noch „Extended Main Path Analysis“ genannt.

nommen. Die 1-path-Version des Graphen in Abbildung 1 z.B. wäre um die A und B sowie H und I verbindenden Kanten reduziert und würde deshalb die Betweenness von C und G verstärken.

In einem Autorennetzwerk kennzeichnen diese *main paths* die wichtigsten Kooperationsbeziehungen eines Autors. Bezogen auf das ganze Autorennetzwerk beschreiben sie die Hauptkooperationsverbindungen in einer wissenschaftlichen Community. Ein Netzwerk, das nur aus main paths besteht, heißt *m*-paths- oder Main-Paths-Netzwerk. Das Konzept eines „main path" entspricht der Vorstellung, dass solche Wege in einem Autorennetzwerk zu den *centers of excellence* in einem Forschungsfeld führen, also den Bereichen in einer wissenschaftlichen Community, wo sich die Hauptforschungsaktivitäten des Feldes abspielen.

3.4 Propagierung von Autorennetzwerken

Wie werden nun Autorennetzwerke auf der Basis von bibliographischen Datenbanken identifiziert? Unser Modell unterstützt hierzu zwei Varianten, die im folgenden vorgestellt werden: die Propagierung auf der Basis einer gegebenen Dokumentenmenge und die Propagierung von „persönlichen" Netzwerken eines gegebenen Autors.

Die technische Basis hiefür ist die am IZ entwickelte infoconnex-Software[15] sowie eine mit der infoconnex-Suchmaschine kommunizierende, vom Autor entwickelte Autorennetzwerk-Komponente in Java. Mit der infoconnex-Suchmaschine können gleichzeitig mehrere heterogene Datenquellen abgefragt werden. Eine Propagierung von Autorennetzwerken über die infoconnex-Suchmaschine ist zur Zeit allerdings nur für Datenbanken möglich, die über JDBC erreichbar sind und abfragbare Koautoren-Relationen in ihrer Datenbank vorhalten. Die Autorennetzwerk-Komponente unterstützt zur Zeit die IZ-Datenbanken SOLIS und FORIS[16] sowie die Datenbank FIS-Bildung vom Deutschen Institut für Pädagogische Forschung (DIPF). Eine Erweiterung des Protokolls, die auch eine Unterstützung von Nicht-JDBC-Datenquellen erlaubt, ist in Vorbereitung.

---

[15] www.infoconnex.de

[16] Da FORIS als Forschungsprojektdatenbank keine Koautoren-Relationen enthält, werden hier stattdessen Ko-Projektmitarbeiter-Relationen verwendet.

3.4.1 Propagierung auf der Basis von Resultsets

Das Vorgehen bei der Propagierung von Autorennetzwerken auf der Basis einer Dokumentenmenge sieht wie folgt aus (s. Abb. 2):

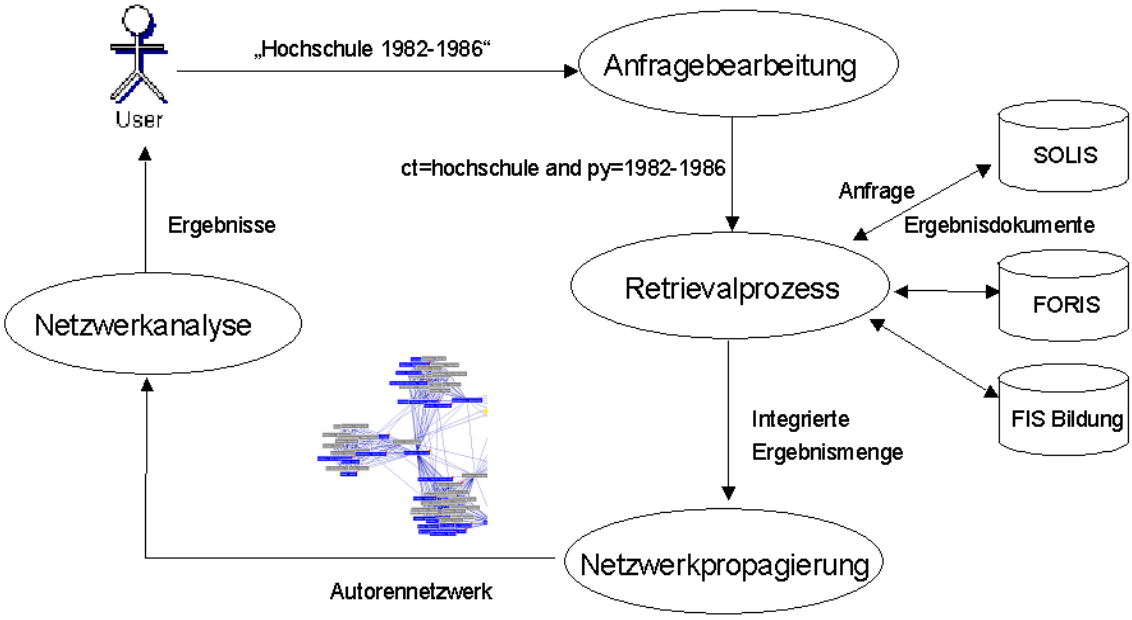

Abb. 2: Vorgehensmodell bei der Propagierung von Autorennetzwerken auf der Basis von Ergebnismengen zu einer Recherche

Der Benutzer führt eine Recherche mit der infoconnex-Suchmaschine aus, z.B. zum Thema ‚Hochschule 1982-1986’ in den Datenbanken SOLIS, FORIS und FIS-Bildung[17]. Die infoconnex-Suchmaschine konvertiert die von den abgefragten Datenbanken selektierten Dokumente zu einem integrierten XML-Resultset. Aus den in den Ergebnisdokumenten enthaltenen Angaben zu den Autoren werden Koautoren-Paare gebildet und in eine interne Datenstruktur der Autorennetzwerk-Komponente tupelweise eingestellt[18]. Diese Datenstruktur ist die interne Repräsentation des Autorennetzwerkes, auf dessen Basis dann Berechnungen z.B. zur Zentralität der Akteure durchgeführt werden. Die Ergebnisse der Analyse der Autorennetzwerke werden an den

[17] Das Beispiel ist der in Kapitel 4 beschriebenen Untersuchung zur Entwicklung von Vernetzung im Forschungsfeld Bildung entnommen.

[18] Alternativ wird die Generierung von Koautoren-Anfragen unterstützt, d.h. es werden für die Ergebnisdokumente Koautoren-Anfragen generiert und über die infoconnex-Suchmaschine an die Datenquellen gestellt. Die Ergebnistupel können dann direkt in die interne Netzwerkrepräsentation eingestellt werden. Diese Variante kommt zum Tragen, wenn die Suchmaschine nur Kurzinformationen zu den Trefferdokumenten zurückliefert, die keine Angaben zu den Autoren enthalten.

Benutzer zurückgeliefert. Kapitel 5.1.1 und 5.1.2.1 diskutieren, wie dieses Propagierungsmodell für die Suche nach Experten und relevanten Dokumenten genutzt werden kann.

Eine Besonderheit bei der Propagierung über mehrere Datenbanken ist, dass auf diese Weise Autorennetzwerke generiert werden können, die sich über mehrere Datenbestände, möglicherweise sogar über mehrere Fachgebiete, verteilen. Abbildung 3 zeigt einen Ausschnitt des sich über SOLIS, FORIS und FIS-Bildung erstreckenden Autorennetzwerkes zum Thema ‚Hochschule 1982-1986'.

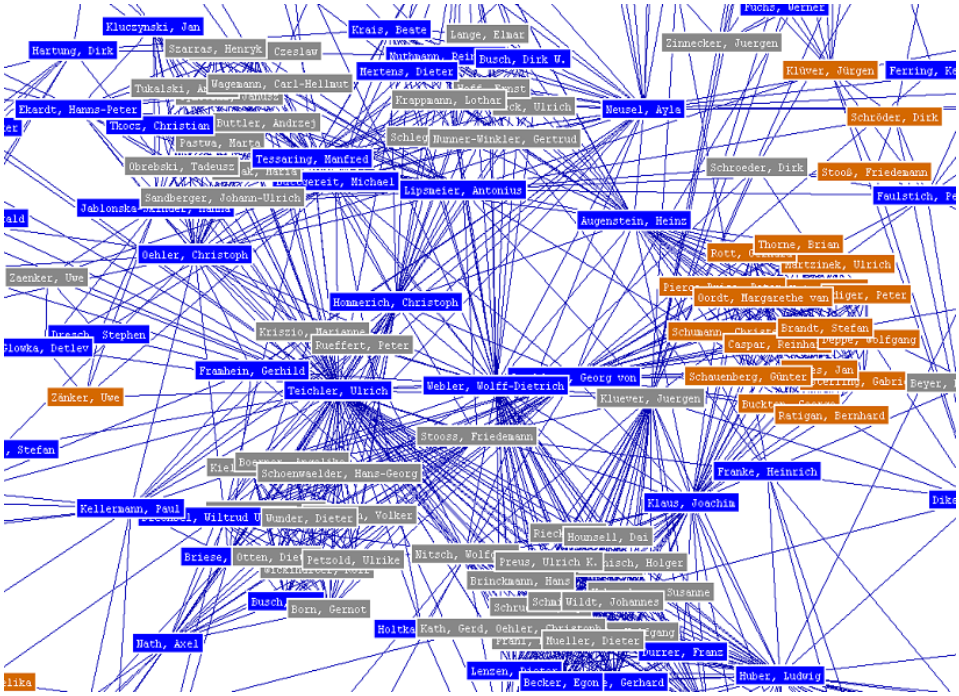

Abb. 3: Ein Bildungsforschernetzwerk in SOLIS, FORIS und FIS-Bildung (Ausschnitt). Die braunen Knoten sind Akteure, die nur in den IZ-Datenbanken SOLIS und FORIS vorkommen. Die grauen repräsentieren Akteure, die nur in FIS-Bildung vorkommen. Die blauen stellen Akteure dar, die sowohl in den IZ-Datenbanken als auch in FIS-Bildung erscheinen.[19]

[19] Die Farbcodierung ist in der Printversion des Arbeitsberichtes nicht ersichtlich. Die PDF-Version desselben ist zugänglich über die Homepage des Autors (http://www.social-science-gesis.de/IZ/Mutschke/) bzw. unter http://www.social-science-gesis.de/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm

3.4.2 Propagierung persönlicher Netzwerke von Autoren

Ein weiterer Ansatz für die Propagierung von Autorennetzwerken ist, von einem bestimmten Autor auszugehen und dessen „persönliches“ Autorennetzwerk zu generieren. Diese Variante entspricht den in der Netzwerkanalyse bekannten *Ego-zentrierten Netzwerken*. Ego-zentrierte Netzwerke sind Netzwerke, die die Beziehungen eines betrachteten Akteurs *Ego* mit allen seinen Ko-Akteuren, den *Alteri*, sowie deren Vernetzung untereinander darstellen.

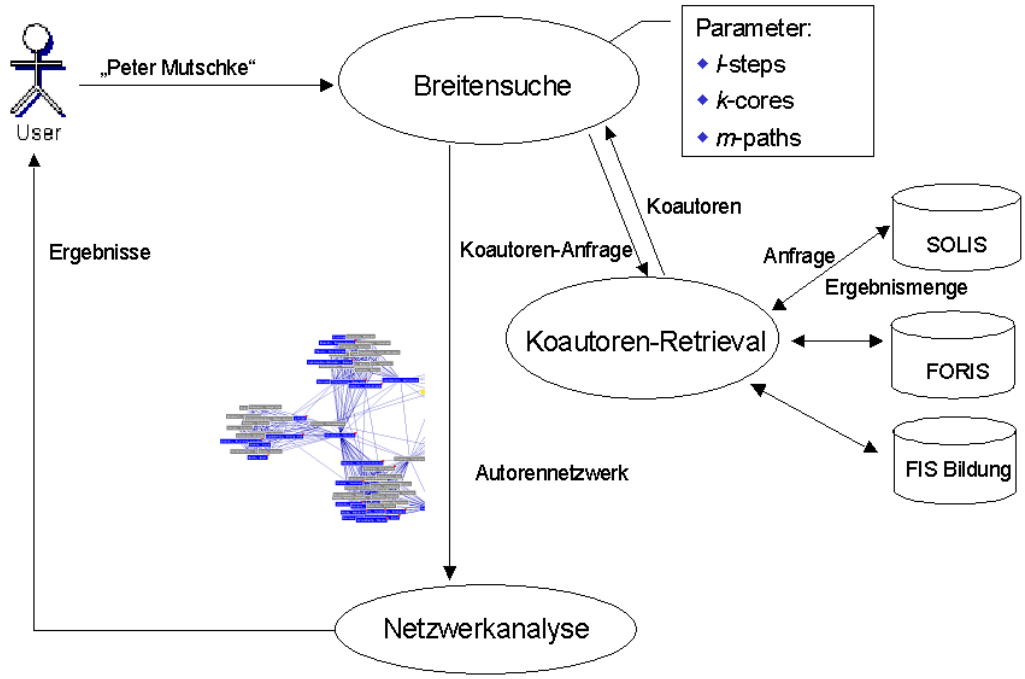

Abb. 4: Vorgehensmodell bei der Propagierung von erweiterten Ego-zentrierten Autorennetzwerken

Ego-zentrierte Netzwerke in der Netzwerkanalyse beschränken sich allerdings auf direkte Beziehungen von Ego zu den Alteri. Um auch das weitere strukturelle Umfeld eines Autors evaluieren zu können, erweitern wir dieses Modell um die *indirekten* Beziehungen von Ego, d.h. es werden auch Akteure in das Netzwerk von Ego aufgenommen, die zwei oder mehr Schritte von Ego entfernt sind. Bei einer Schrittweite von 2 werden also nicht nur die direkten Koautoren von Ego erfasst, sondern auch deren Ko-Akteure (sofern sie nicht selbst direkt mit Ego verbunden sind).

Das Verfahren für die Propagierung erweiterter Ego-zentrierter Autorennetzwerke erfolgt durch eine Breitensuche, bei der rekursiv dem transitiven Abschlusses der von Ego ausgehenden Koautoren-Relationen „nachgelaufen“ wird. Das Netzwerk wird dabei über *alle* Nachbarn eines gegebenen Knoten

propagiert, d.h. es werden zuerst alle Nachbarn des Ausgangsknoten in das Netzwerk eingefügt, dann deren Nachbarn usw. (s. Abb. 4).

Das Problem dieser Vorgehensweise ist jedoch zu bestimmen, wann der Breitensuchprozess terminieren soll. Eine naheliegende und in unserem Modell auch unterstützte Variante ist, den Benutzer einen Tiefenschwellwert (*l*-steps) vorgeben zu lassen, der die maximale Entfernung *l* zwischen Ego und seinen Alteri, d.h. die maximale Zahl der Kanten zwischen Ego und den Alteri definiert. Der Breitensuchprozess terminiert dann, wenn der durch den Benutzer gesetzte Tiefenschwellwert erreicht ist. Ein Tiefenschwellwert von z.B. 2 bedeutet, dass die maximale Entfernung zwischen den Alteri und Ego zwei Links ist. Das Netzwerk würde also alle direkten Nachbarn von Ego erfassen sowie deren Nachbarn. Der Tiefenschwellwert kann beliebig hoch gesetzt werden. Ein Wert von –1 bedeutet, dass der Propagierungsprozess nicht bei einer bestimmten Tiefe terminiert, sondern jeweils an terminalen Knoten, d.h. bei Autoren, die keine weiteren Koautoren haben, die nicht schon gefunden wurden. Die Propagierung erfolgt dann über den gesamten Korpus.

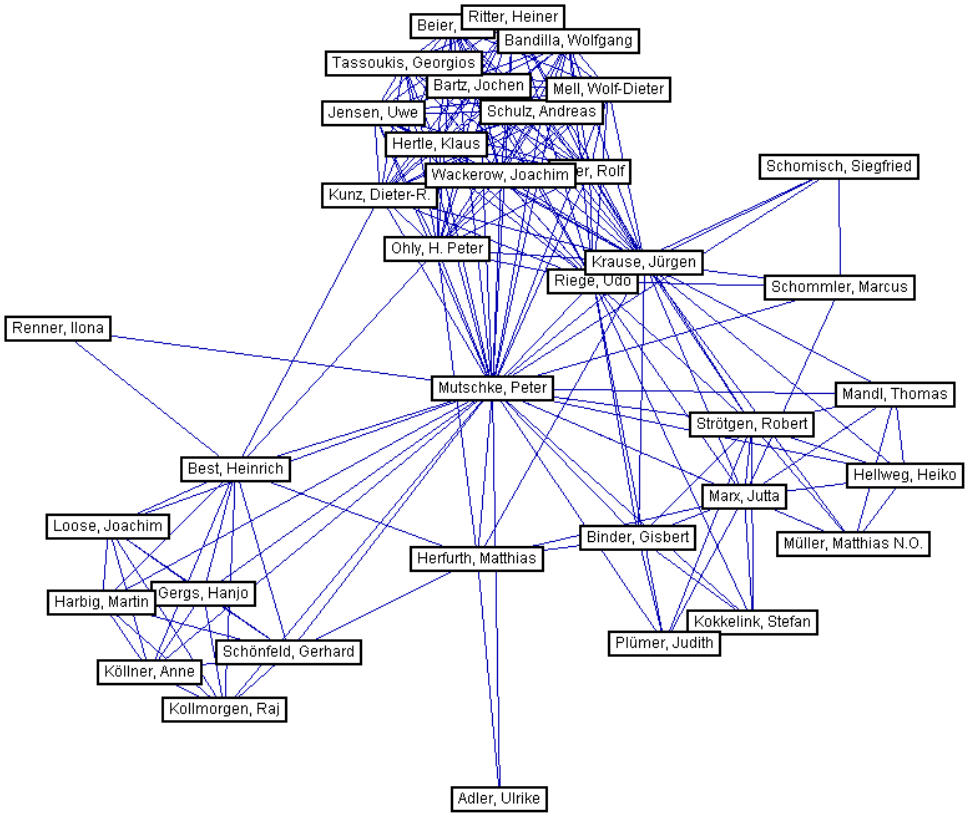

Abb. 5: Das Ko-Autorennetzwerk von ‚Peter Mutschke' in SOLIS und FORIS (*l*=1)

Ein Tiefenschwellwert ist allerdings notwendig, weil die Zahl der Alteri in sozialen Netzwerken i.d.R. mit jedem Schritt exponentiell anwächst und der Suchraum damit explodiert. Das Ego-zentrierte Autorennetzwerk des Autors in SOLIS und FORIS z.B. umfasst bei einer Schrittweite von 1 zunächst 36 Knoten, bei einer Schrittweite von 2 schon 256 Knoten und bei einer Schrittweite von 3 bereits 2629 Akteure. Abbildung 5 zeigt das Netzwerk des Autors in SOLIS und FORIS mit einer Schrittweite von 1.

Das konzeptuelle Problem eines Tiefenschwellwerts ist jedoch, dass die gesuchten zentralen Akteure außerhalb dieses Schwellwertes liegen können, von Ego aus betrachtet also tiefer in der Sozialstruktur „verborgen“ sind als der Tiefenschwellwert vorgibt. Einen hohen Tiefenschwellwert zu setzen, ist jedoch aufgrund des mit jeder Schrittweite verbundenen exponentiellen Wachstums des Netzwerks mit erheblichen Performanceproblemen verbunden, so dass der Benutzer gezwungen ist, einen niedrigen Schwellwert anzusetzen, um zu vermeiden, dass sich der Suchraum dramatisch ausweitet. Dies ist bereits bei einem Tiefenschwellwert größer 2 der Fall. Der Propagierungsprozess kann daher mit einigen Skalierungs-Parametern optimiert werden: den *k*-cores, d.h. die Propagierung mit Ko-Akteuren mit einem Mindestdegree von *k* (s. Kapitel 3.3.1) und vor allem den *m*-paths (s. Kapitel 3.3.2).

Das vom Autor vorgeschlagene *m*-paths-Modell (Mutschke 2001, 2003) ist für die Propagierung von Ego-zentrierten Netzwerken – in der um indirekte Beziehungen erweiterten Variante – deswegen besonders interessant, weil es die Suche nach zentralen Akteuren in von Ego aus betrachtet tieferen Schichten der Sozialstruktur effizient unterstützt. Die Komplexität der Netzwerke wird mit dem *m*-paths-Modell auf die jeweils „besten“ Beziehungen zwischen den Akteuren, die sog. *main paths* reduziert. Es werden also nicht alle, sondern nur die Verbindungen beschritten, die jeweils zu den *m* Ko-Akteuren mit dem höchsten Degree im Umfeld des betrachteten Akteurs führen. Der Prozess terminiert für jeden beschrittenen Pfad an einem lokalen Maximum, d.h. bei dem Akteur, der keine Ko-Akteure mehr hat, die einen höheren Degree als dieser selbst aufweisen. Dieses Vorgehen generiert eine Sequenz von Kanten, die einen bei Ego beginnenden und an einem lokalem Maximum endenden *main path* durch das Netzwerk beschreibt. Dieses Verfahren hat den Vorteil, dass auf eine nachträgliche Zentralitätsberechnung verzichtet werden kann, da die terminalen Knoten zugleich die zentralsten Knoten in Ego’s Umfeld sind.

Formal entspricht die *m*-paths-Technik somit einer um Hill-Climbing-Strategien und Backtracking erweiterten Prioritätssuche in gewichteten Graphen. Eine Backtracking-Strategie ist notwendig, weil es je nach Struktur des Netzwerkes sein kann, dass über den „besten“ Pfad allein möglicherweise

nicht alle lokalen Maxima um Ego herum gefunden werden. Deshalb muss es möglich sein, auch mit den „zweitbesten“ usw. Ko-Akteuren zu propagieren. Dies kann der Benutzer über den *m*-Parameter steuern. Kapitel 5.1.2.2 diskutiert einige Anwendungsfälle dieses Propagierungsverfahrens bei der Suche nach Experten im Umfeld einer Person.

4 Eigenschaften von Autorennetzwerken

Wie sehen aber nun Autorennetzwerke in wissenschaftlichen Domänen aus? Welche Eigenschaften haben sie und wie entwickeln sie sich über Zeit? Dies sind Fragen, von deren Beantwortung entscheidend abhängt, welche Relevanz die strategische Position wissenschaftlicher Akteure in Autorennetzwerken hat. Denn um die geht es primär in den auf Akteurszentralität basierenden Retrievalmodellen, die in Kapitel 5 vorgeschlagen werden. Um also etwas mehr über Autorennetzwerke zu erfahren, wurden am IZ einige empirische Untersuchungen durchgeführt, deren Hauptergebnisse im folgenden vorgestellt werden.

4.1 Evolution und Dynamik

Bei einer Untersuchung der Entwicklung von Vernetzung in vier Bildungs-Forschungsfeldern (Bildungssystem, Bildungsplanung, Hochschule/Studium und Schule) in dem Zeitraum von 1982-2001[20] zeigte sich ein erhebliches Vernetzungsniveau in allen vier Feldern in der letzten Untersuchungsperiode 1997-2001 (s. Tab. 1). Von den durchschnittlich etwa 6.000 Personen, die in den vier Feldern in diesem Zeitraum geforscht und geschrieben haben, unterhielten durchschnittlich 60% Kooperationsbeziehungen zu anderen Wissenschaftlern des Forschungsfeldes. Durchschnittlich etwa 700 Akteure (20% der Vernetzten) waren in einer einzigen Komponente, d.h. in einem verbundenen Teilgraph integriert. Bei einem Vergleich mit dem Vernetzungsniveau in anderen Forschungsfeldern, hier Gewalt, Jugend, Frau und Alter, in demselben Zeitraum, fand sich ein ähnlich hoher Anteil an vernetzten Personen (durchschnittlich 62%).

[20] Für die Propagierung der Netzwerke wurde das in Kapitel 3.4.1 beschriebene Verfahren verwendet.

Feld (1997-2001)	Personen insges.	Vernetzte Personen	% Personen	Größte Komponente	% Vernetzte
Bildungssystem	5487	3260	59	458	14
Bildungsplanung	4260	2326	55	739	32
Hochschule	7453	4867	65	924	19
Schule	7712	4747	62	625	13
Durchschnitt	**6228**	**3800**	**60**	**687**	**20**

Tab. 1: Vernetzungsniveau in vier Bildungsforschungsfeldern

Betrachtet man nun die relative Zunahme der Vernetzung von 1982-2001 anhand von vier Fünf-Jahres-Zeitscheiben (1982-1986, 1987-1991, 1992-1996, 1997-2001), dann lässt sich - ausgehend von den Anfangsgrößen im ersten Untersuchungszeitraum 1982-1986 - für den Bereich ‚Bildungssystem' ein geradezu dramatischer Vernetzungsprozess feststellen (s. Abb. 6): Während die Zahl der Personen insgesamt von 1982/86-1997/2001 lediglich um 54% zunahm, und die Zahl der vernetzten Personen nur um 80%, ist die Zahl der in der größten Netzwerkkomponente integrierten Personen um 718% angewachsen, d.h. die Größe dieser Komponente hat sich von der ersten bis zur letzten Untersuchungsperiode verachtfacht.

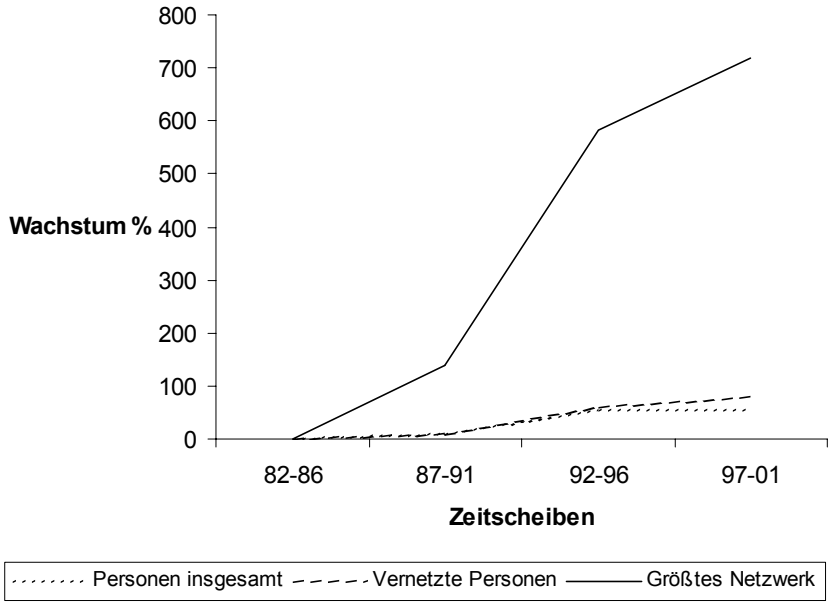

Abb. 6: Entwicklung der Vernetzung im Feld ‚Bildungssystem'

Ähnlich sieht es für den Bereich ‚Bildungsplanung' aus (s. Abb. 7), wenngleich in diesem Feld die Größenverhältnisse etwas geringer ausfallen.

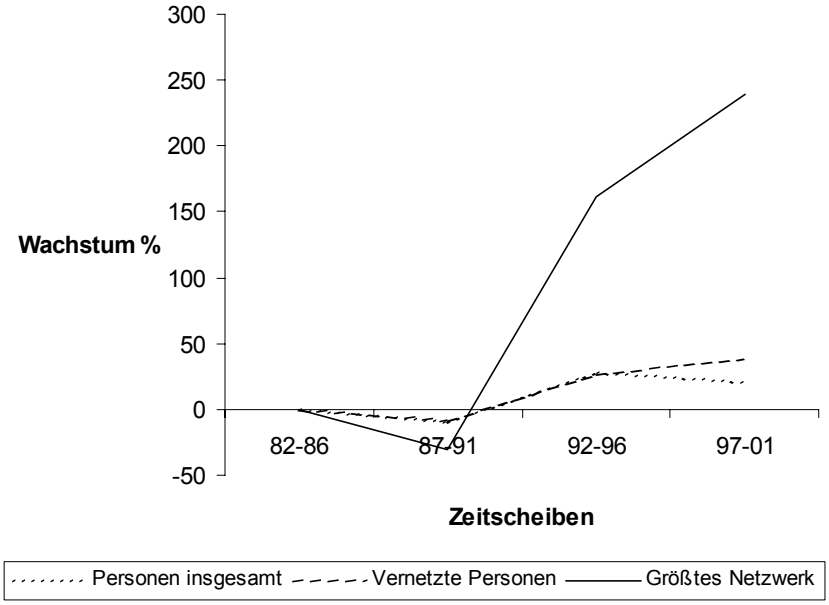

Abb. 7: Entwicklung der Vernetzung im Feld ‚Bildungsplanung'

Nicht ganz so ausgeprägt, aber von der Grundtendenz ähnlich fällt die Entwicklung der Vernetzung im Feld ‚Hochschule' aus (s. Abb. 8).

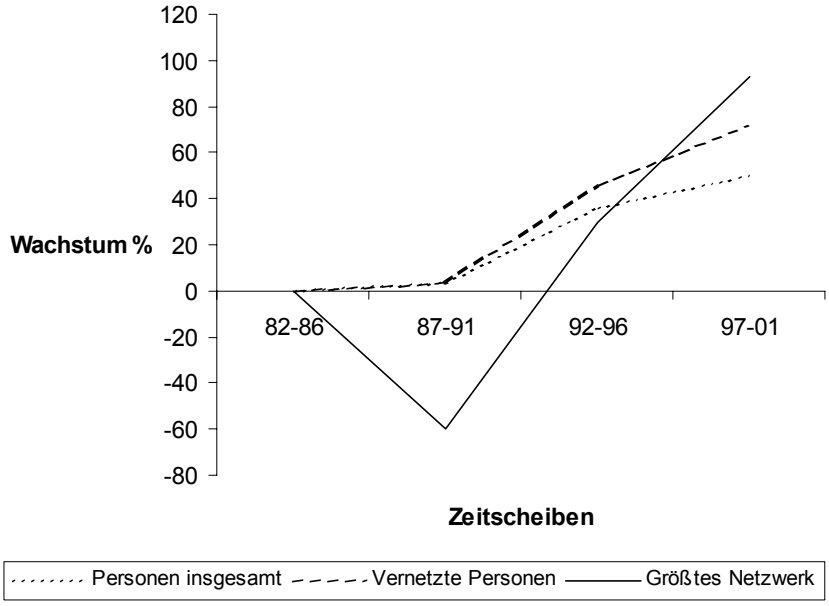

Abb. 8: Entwicklung der Vernetzung im Feld ‚Hochschule'

Besonders markant, und ähnlich ausgeprägt wie im Feld ‚Bildungssystem', ist die Entwicklung der Vernetzung wiederum im Feld ‚Schule': Während sich die Zahl der Personen lediglich um 57% erhöht, und sich die Zahl der kooperierenden Personen in etwa verdoppelt, verachtfacht sich die Größe des größten Komponente (s. Abb. 9).

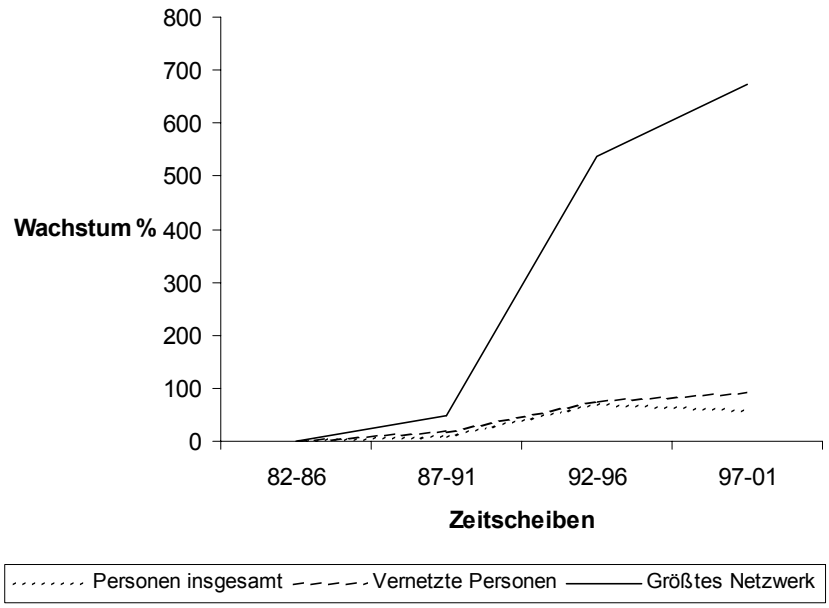

Abb. 9: Entwicklung der Vernetzung im Feld ‚Schule'

Bemerkenswert ist nun, dass sich trotz signifikant zunehmender Vernetzung die Struktur der Netzwerke über die Zeit nicht wesentlich verändert hat. Wenn wir einmal einen Blick auf die Entwicklung der Struktur der größten Netzwerkkomponente im Feld ‚Bildungssystem' werfen, dessen Größe sich ja verachtfacht hatte, dann sehen wir, dass sich grundlegende Netzwerkeigenschaften über die Zeit kaum verändert haben (s. Tab. 2):

Zeit-scheibe	Durch-messer[1]	Charakt. Pfadlänge[2]	Clustering-Koeffizient[3]
1982-1986	12	4,38	0,70
1987-1991	13	5,71	0,78
1992-1996	22	8,38	0,70
1997-2001	15	6,38	0,73

Tab. 2: Veränderung der Netzwerk-Topologie im Feld ‚Bildungssystem'

Der Durchmesser des Netzwerkes, d.h. die Länge des längsten kürzesten Pfades, steigt zwar zunächst von 12 im Zeitraum 1982-1986 auf 22 im Zeitraum 1992-1996 an, fällt dann aber trotz Anwachsens des Netzwerkes in der letzten Periode wieder auf 15 ab. Ähnlich sieht es mit der charakteristischen Pfadlänge aus, d.h. der durchschnittlichen Länge aller kürzesten Pfade: Die charakteristische Pfadlänge steigt zunächst von ca. 4 (1982-86) auf ca. 8 im Zeitraum 1992-96 an, fällt dann aber wieder auf etwa 6 ab. Gleichzeitig bleibt der Clustering-Koeffizient, d.h. der Grad, in dem die Nachbarn eines Akteurs selbst untereinander vernetzt sind, auf einem konstant hohen Niveau von durchschnittlich 73%.

Diese Entwicklung deutet darauf hin, dass in dem Zeitraum 1992-96 verstärkt zuvor isolierte Cliquen zu dem größten Netzwerk hinzugestoßen sind, in dem sie sich an „Außenstellen“ des Netzwerkes, d.h. an Akteuren, die in der Peripherie des Netzwerkes lokalisiert waren, angedockt haben. Dieser Prozess hat zunächst zu einer Abnahme der Dichte des Netzwerkes geführt hat, also einer Ausfransung desselben, die sich in längeren Wegen im Netzwerk ausdrückt. Im Zeitraum 1997-2001 ist es diesen Cliquen dann offensichtlich gelungen, stärker an Kommunikations- und Kooperationsprozesse zu partizipieren, die im Zentrum des Netzwerkes stattfanden. Akteure, die zunächst eher an der Peripherie angesiedelt waren, sind verstärkt Kooperationsbeziehungen mit zentraleren Akteuren eingegangen (preferential attachment). Diese Anlagerungsprozesse haben dann zu einer Verdichtung des Netzwerkes geführt, zu erkennen an einem kleineren Durchmesser und v.a. einer kürzeren charakteristischen Pfadlänge. Dieser Prozess entspricht dem oben beschriebenen Phasenmodell der Entwicklung wissenschaftlicher Communities.

4.2 Topologie

Wie wir gerade gesehen haben, war bei der Untersuchung der Entwicklung der Vernetzung im Forschungsfeld ‚Bildungssystem’ besonders auffallend, dass trotz der Größe der Population (5.487 Akteure im Untersuchungszeitraum 1997-2001, s. Tab. 1) und der Größe der größten Komponente (458 Akteure, s. ebd.) die Netzwerke relativ dicht „gestrickt“ sind: Die charakteristische Pfadlänge liegt bei durchschnittlich etwa sechs Links und die lokale Vernetzungsrate (Clustering-Koeffizient), bleibt in allen vier Untersuchungszeiträumen auf einem konstant hohen Niveau von durchschnittlich 73% (s. Tab. 2).

Wenn man sich die Struktur der größten Komponente in allen vier Bildungsfeldern in der letzten Untersuchungsperiode anschaut, dann findet man sehr ähnlich dichte topologische Eigenschaften vor (s. Tab. 3):

- relativ kurze Verbindungen von durchschnittlich 6,6 Links, d.h. alle Akteure in den Netzwerken im Bereich Bildung werden durchschnittlich über sechs bis sieben Zwischenschritte erreicht.
- ein relativ hohes lokales Vernetzungsniveau von durchschnittlich 76%.

Feld (1997-2001)	Charakt. Pfadlänge[1]	Clustering-Koeffizient[2]	Super-knoten[3]
Bildungssystem	6,38	0,73	38
Bildungsplanung	5,22	0,81	91
Hochschule	6,92	0,74	59
Schule	7,79	0,75	45
Durchschnitt	**6,58**	**0,76**	**58**

Tab. 3: Topologie der Netzwerke in vier Bildungsforschungsfeldern

Diese Strukturmerkmale weisen das Organisationsmuster dieser Netzwerke als *Small-World-Architektur* (Watts 1999) aus. Small Worlds zeichnen sich gerade dadurch aus, dass die Länge der Verbindungen in einem Netzwerk mit dessen Anwachsen nur unwesentlich, nämlich logarithmisch, zunimmt und gleichzeitig die lokale Vernetzungsrate relativ hoch ist.

Dieser Bauplan der Autorennetzwerke ist offensichtlich nicht nur unabhängig von Untersuchungszeitraum und Thema, sondern auch unabhängig von der Größe der Netzwerke, wie eine Analyse von Autorennetzwerken (hier: der größten Komponente) auf der Basis folgender vier Kollektionen in den Datenbanken SOLIS und FORIS zeigt: (A) Schlagwort = ‚Internet', (B) Schlagwort = ‚Gewalt', (C) Erscheinungjahr >= 1998, (D) komplette SOLIS/FORIS-Kollektion. Wie wir in Tabelle 4 sehen, weisen die größten Komponenten in allen vier Autorennetzwerken trotz ihrer gewaltigen Größenunterschiede (75 Akteure in Kollektion A gegenüber mehr als 100.000 Akteuren in der D-Kollektion) sehr ähnliche Small-World-Merkmale auf: relative kurze Wege (selbst in Kollektion D) von durchschnittlich etwa sieben Links[21] und eine hohe Clusterung von durchschnittlich ca. 60%.

[21] Ähnliche Verhältnisse hat Newman (2001a) für Kooperationsnetzwerke in den Bereichen Physik, Medizin und Informatik festgestellt.

SOLIS/FORIS-Kollektion	Größe	Charakt. Pfadlänge	Clustering-Koeffizient
(A) Internet	75	5.19	0.45
(B) Gewalt	469	6.76	0.75
(C) >= 1998	29.749	8.75	0.63
(D) ALLE	104.364	6.68	0.60
Durchschnitt		*6.85*	*0.61*

Tab. 4: Topologie von Autorennetzwerken unterschiedlicher Größe

Ein weiteres typisches Merkmal für Autorennetzwerke ist die Existenz von *Superknoten*, d.h. von Akteuren, von denen verhältnismäßig viele Verbindungen abgehen. In unseren vier Bildungsforschungsfeldern hatten die Akteure mit dem höchsten Degree durchschnittlich 58 Verbindungen (s. Tab. 3).

Wenn wir uns hierzu einmal die Verteilung der Zahl der Akteure auf die Zahl der Verbindungen am Beispiel ‚Bildungssystem 1997-2001' anschauen (s. Abb. 10), dann stellen wir tatsächlich fest, dass sehr viele Akteure, wie auch zu erwarten war, nur sehr wenige Kooperationsbeziehungen haben, während einige wenige Akteure sehr viele Verbindungen unterhalten und somit herausragend zentral sind. Ähnliche Verteilungen finden sich auch in anderen Forschungsfeldern.

Dieses Ergebnis zeigt, dass Autorennetzwerke in wissenschaftlichen Communities offensichtlich eine ausgeprägte *Zentrum-Peripherie-Topologie* haben (vgl. auch Jansen 2003, 33f.). Zentrum-Peripherie-Muster zeichnen sich gerade dadurch aus, dass einige der Akteure im Vergleich zu den anderen besonders viele Verbindungen im Netzwerk unterhalten.

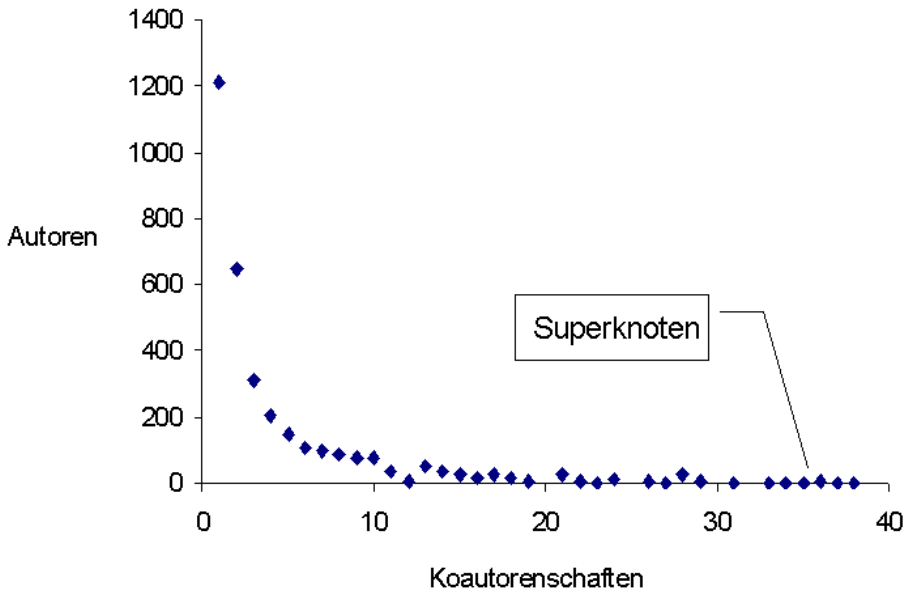

Abb. 10: Verteilung Akteure – Beziehungen im Forschungsfeld ‚Bildungssystem 1997-2001'

Dieses Ergebnis ist insofern interessant, als die Problemlösungsfähigkeit von Gruppen in stark zentralisierten Netzwerken offensichtlich höher ist als in weniger zentralisierten, wie empirische Studien nachwiesen. Untersuchungen in verschiedenen Domänen zeigten, dass Netzwerke (oder Teilgruppen), in denen es Akteure mit herausragender Zentralität gibt, offenbar über eine größere Kooperations- und Problemlösungskapazität verfügen als stark fragmentierte Strukturen (vgl. Jansen 2003, 128). Der Fragmentierungsgrad einer wissenschaftlichen Community ist deshalb ein wesentliches Qualitätskriterium für ihren „Reifegrad". Was dies für die Nutzung von Autorennetzwerken in Informationssystemen bedeutet, muss allerdings noch untersucht werden.

4.3 Zusammenfassung und Schlussfolgerungen

Die empirischen Untersuchungen zu Evolution und Topologie von Autorennetzwerken in den Sozialwissenschaften bestätigen, dass sich in wissenschaftlichen Communities nicht nur signifikant zunehmende Vernetzungsprozesse über Zeit und eine relativ hohe Vernetzungsrate nachweisen lassen, sondern auch ein bestimmtes Organisationsmuster von Vernetzung, nämlich Small-World-Architekturen „aristokratischen" Typs, die sich durch eine relativ hohe lokale Clusterdichte, global relativ kurze Verbindungen zwischen den Knoten im Netzwerk und durch die Existenz von Superknoten auszeichnen.

Wir haben es bei wissenschaftlichen Kooperationsnetzwerken also keinesfalls mit „Zufallsgraphen“ zu tun, sondern mit immer wiederkehrenden Organisationsmustern, die als grundlegendes Architekturprinzip die Struktur der Netzwerke prägen: der Small-World-Struktur. Dies bedeutet natürlich auch, dass die strategische Position der Akteure in diesen Netzwerken, mithin deren Zentralität, keine Zufallserscheinung ist, sondern dass Ergebnis von Small-World-Prozessen.

Ein Erklärungsmodell für dieses Phänomen ist der bereits genannte und empirisch bestätigte Preferential-Attachment-Mechanismus, also der Neigung wissenschaftlicher Akteure primär mit denen zu kooperieren, die bereits viele Koautoren haben (Newman 2001c, Barabasi et al. 2002). Die Analyse soziokognitiver Strukturen in sozialwissenschaftlichen Forschungsfeldern untermauert dieses Anlagerungs-Theorem: Zentrale Akteure neigen dazu, Mainstream-Themen zu manifestieren (Mutschke & Renner 1995, Mutschke & Quan Haase 2001). Die Vermutung liegt nahe, dass sie deshalb auch verstärkt dazu tendieren, Kooperationsbeziehungen mit ebenfalls zentralen Akteuren einzugehen. Gleichzeitig lässt sich nachweisen, dass soziale Aufsteiger in wissenschaftlichen Gemeinschaften, d.h. Akteure mittlerer Zentralität, eher zu innovativen Themen tendieren, um dann – so die Vermutung - die Aufmerksamkeit zentraler Akteure auf sich zu ziehen, sobald sich diese innovativen Felder anschicken, zum Mainstream der Forschung zu avancieren (Mutschke & Quan Haase 2001).

Nach Jansen (2003) kann aus einer zentralen Position in einem Zentrum-Peripherie-Netzwerk sogar der wissenschaftliche Erfolg einzelner Akteure oder Forschungseinrichtungen abgeleitet werden: „Die Akteure im Zentrum der Sozialstruktur ... sitzen im Zentrum des Informationsaustausches und der Kooperation. Dies erlaubt es ihnen, verschiedene wissenschaftliche Ergebnisse, Methoden und Ansätze zu kombinieren und zu bewerten und trägt zu ihrem künftigen wissenschaftlichen Erfolg bei.“ (Jansen 2003, 34). Dabei integrierten zentrale Akteure das Fachgebiet nicht nur, weil ihre wissenschaftlichen Leistungen anerkannt seien, sondern auch weil sie Kontakte zu den peripheren Gruppen unterhielten.

Ziel der Autorennetzwerk-Komponenten am IZ ist es, diese Akteure zu finden und für die Verbesserung der Suche in Datenbanken zu nutzen. Im folgenden werden einige Szenarios hierzu vorgestellt.

5 Nutzung von Autorennetzwerken in Informationssystemen

Wie wir gesehen haben, weisen Autorennetzwerke bestimmte Organisationsmuster auf, die auch die strategische Position der Akteure in diesen Netzwerken, mithin deren Zentralität, entscheidend prägen: Autorennetzwerke haben eine ausgeprägte Small-World-Architektur mit einer starken Tendenz zur Zentralisierung (Zentrum-Peripherie-Muster). Dies bedeutet natürlich auch, dass eine zentrale Position in einem Autorennetzwerk keine Zufallserscheinung ist, sondern dass Ergebnis von Small-World- und Preferential-Attachment-Prozessen. Zentrale Akteure in wissenschaftlichen Communities sind mithin genau die Akteure, die im Zentrum des Informationsaustausches positioniert sind, also dort, wo der Umsatz des Wissens stattfindet.

Es liegt daher nahe, Wissen über die globale Kommunikations- und Kooperationsstruktur in wissenschaftlichen Communities, und hier insbesondere Wissen über die Zentralität der Akteure in diesen Strukturen für die Informationensuche in Datenbanken auszunutzen. Für die Nutzung von Akteurszentralität als *Suchstrategie* bieten sich eine Reihe von Szenarios an, die wir im folgenden diskutieren. Ziel der Nutzung von Autorennetzwerken in Informationssystemen ist es, Strukturinformationen über das Interaktionsgeschehen in einer wissenschaftlichen Community als *Strategie* zu benutzen, um relevante Informationen in Datenbanken unter Vagheitsbedingungen besser auffinden zu können.

5.1.1 Ranking von Dokumenten nach Akteurszentralität

Allein die schiere Masse an Informationen, die einem Benutzer über bibliographische Datenbanken zugänglich sind, legt intelligente Verfahren nahe, die Ergebnismengen sinnvoll strukturieren und aus der Fülle der Daten die hochrelevanten Informationen herausfiltern. Für das Dokumentenretrieval sind hierfür quantitativ-statistische Rankingverfahren (wie z.B. das Vektorraummodell) entwickelt worden, die jedoch allein auf inhaltliche Entscheidungskategorien abgestimmt sind, d.h. auf das Vorkommen der Anfrageterme in den Dokumenten. Das Vektorraummodell z.B. liefert Dokumente, die den Suchbegriffen des Benutzers *inhaltlich* am ähnlichsten sind. Relevant sind hier Dokumente, die eine hohe Selektivität bzgl. der *inhaltlichen* Suchbegriffe der Anfrage haben. Die Relevanz der Autoren, also der eigentlichen Akteure wissenschaftlichen Outputs, bleibt in den herkömmlichen Standardretrievalverfahren jedoch vollkommen unberücksichtigt. Die Ursache für dieses Desiderat

ist die Fokussierung des Faches Information Retrieval auf das Wiederauffinden von Informationen durch inhaltliche Matching-Verfahren.

Eine Berücksichtigung der Relevanz der Akteure wissenschaftlicher Arbeit beim Dokumentenranking ist allerdings in dreifacher Hinsicht interessant:

- Hohe Selektivität: Die Zahl der Akteure der in Datenbanken vorkommenden Autoren ist dramatisch hoch und sie nimmt mit dem Wachstum der Datenbanken kontinuierlich zu: Allein in SOLIS sind fast 150.000 Autorennamen verzeichnet. Das Verhältnis von Zahl der Dokumente in SOLIS (ca. 300.000) zu der Zahl der Autoren ist in SOLIS also in etwa 2:1. Dieses Verhältnis macht Autoren zu einem scharfen Selektionskriterium. Mit Autorennamen kann der Suchraum dramatisch eingeschränkt werden. Akteurszentralität wird nicht durch hochfrequente Autoren gestört - wie es etwa bei einem Begriffsnetzwerk der Fall wäre, wo es viele Begriffe gibt, die in sehr vielen Dokumente vorkommen.

- Status in der Sozialstruktur einer Community: Die Akteure wissenschaftlicher Arbeit sind keineswegs isolierte Individuen, die quasi „gleichrangig“ mit anderen agieren. Sondern sie sind eingebettet in Forschungskontexte, die aus einer Vielzahl von (wie auch immer) miteinander interagierenden Individuen bestehen, d.h. sie forschen und schreiben auf dem Hintergrund bestehender Kommunikations- und Kooperationsstrukturen, oftmals sogar in direkter Interaktion mit anderen Wissenschaftlern, oder zumindest mit Rekurs auf die wissenschaftlichen Arbeiten anderer. Dies bedeutet, dass sie bestimmte Rollen und Positionen im Interaktionsgeschehen einnehmen. Zentrale Autoren sind genau die Akteure, die den besten Zugang zu Ressourcen (Informationen) haben, Kommunikations- und Kooperationsprozesse steuern und somit die inhaltliche Konzeptualisierung der Community entscheidend mitprägen. Es liegt daher nahe, bei der Suche nach relevanten Dokumenten auch den sozialen Status der Autoren in der Community zu berücksichtigen.

- Resultset-Strukturierung: Benutzer werden oftmals mit Rechercheergebnissen beträchtlichen Umfanges konfrontiert, die eine gezielte Suche nach relevanten und qualitativ hochwertigen Informationen erschweren oder sogar unmöglich machen. Eine sinnvolle inhaltliche Einschränkung der Suche stellt den Benutzer oftmals vor unüberwindliche Probleme, ist oftmals auch gar nicht erwünscht oder mit unerfreulichen Informationsverlusten verbunden. Eine sinnvolle Strukturierung eines größeren Resultsets, die Informationen über die Struktur eines

Forschungsfeldes berücksichtigte, wie z.B. seine Kooperationsstruktur und den Status der Akteure in dieser Struktur, könnte die Qualität einer Recherche angesichts der schieren Masse an Informationen möglicherweise dramatisch erhöhen.

Die Berücksichtigung von Akteurszentralität bei der Suche oder beim Ranking hätte also nicht nur den Vorteil, den Suchraum drastisch einschränken zu können. Sie würde vielmehr auch die Chance erhöhen, Dokumente zu finden, die von *relevanten* Autoren verfasst wurden, also Veröffentlichungen von Autoren, die in ihrer Community eine bedeutende, vielleicht sogar herausragende Rolle spielen, und deshalb in der Community stärker rezipiert werden als Publikationen von Autoren, die eher an der Peripherie der Community angesiedelt sind.

Das auf Akteurszentralität basierende Rankingmodell zielt konzeptuell also darauf ab, Veröffentlichungen peripherer Autoren von denen zentraler Autoren zu unterscheiden und Dokumente zu favorisieren, bei denen von einer erhöhten Relevanz für die Community ausgegangen werden kann, weil sie von Autoren verfasst wurden, die in dem Forschungsfeld eine zentrale Rolle spielen (oder gespielt haben).

Verglichen mit traditionellen Rankingverfahren ist das hier vorgeschlagene Rankingmodell ein Paradigmenwechsel bei der Relevanzbewertung von Dokumenten, der nicht nur mit dem Relevanzkriterium selbst zu tun hat (hier die Bedeutung der Autoren, dort die inhaltliche Ähnlichkeit zwischen Dokument und Anfrage), sondern vor allem mit der gewählten Analyseebene: Während traditionelle Rankingverfahren beim Dokument ansetzen (um etwa die Häufigkeit eines Anfrageterms in einem Volltext festzustellen)[22], wird bei dem Zentralitätsmodell die *strukturelle* Bedeutung des Autors in der *globalen* Kommunikations- und Kooperationsstruktur seiner Community betrachtet.

Dies macht Akteurszentralität nicht einfach nur zu einem weiteren Selektionskriterium neben anderen, sondern zu einem Qualitätssicherungskonzept: Der Benutzer erhält nicht einfach nur Dokumente, die zu seiner Anfrage am besten passen, sondern – so die Kernthese des Modells – Dokumente, von denen zu erwarten ist, dass sie auch in der wissenschaftlichen Diskussion des Faches eine wichtige Rolle spielen (oder gespielt haben), weil sie von Auto-

[22] Bei der inversen Dokumenthäufigkeit wird zwar auch die Relevanz eines Terms in der Gesamtkollektion gemessen. Dies ist jedoch ein reines Häufigkeitskonzept, das die Bedeutung des Terms für die Community unter Berücksichtigung ihrer strukturellen Beschaffenheit nicht erfasst.

ren verfasst wurden, die eine herausragende Stellung in der Community einnehmen.

Grundsätzlich gibt es für ein auf Akteurszentralität basierendes Rankingmodell zwei Operrationalisierungsvarianten:

- Ex-Post-Ranking: Akteurszentralität wird auf der Basis der Ergebnismenge zu einer Recherche berechnet.
- Ex-Ante-Ranking: Akteurszentralität wird nicht auf der Basis von Rechercheergebnissen, sondern für eine Gesamtkollektion (bzw. für Subkollektionen) berechnet und für die Indexierung der Dokumente verwendet.

5.1.1.1 Ex-Post-Ranking

Bei der Ex-Post-Ranking-Variante wird unter Verwendung des in Kapitel 3.4.1 beschriebenen Propagierungsmodells die Akteurszentralität auf der Basis der Ergebnismenge zu einer Recherche berechnet. Der Benutzer führt also eine Datenbankrecherche durch. Auf der Basis der Ergebnisdokumente wird ein Autorennetzwerk propagiert. Die Zentralität der Autoren in dem Autorennetzwerk wird berechnet, wobei parametergesteuert entweder einer der drei in Kapitel 3.2 beschriebenen Zentralitätsmaße (Degree, Closeness oder Betweenness) verwendet wird oder ein gemischter Zentralitätsindex aus allen drei Zentralitätsmaßen. Die Ergebnisdokumente werden nach der Zentralität ihrer Autoren (und Erscheinungsjahr als sekundäres Sortierkriterium) absteigend sortiert an den Benutzer zurückgeliefert.

Eine Anfrage z.B. zum Thema ‚Elite'[23] in SOLIS und FORIS würde unter mehr als 1500 Trefferdokumenten die Arbeiten von Heinrich Best und Ursula Hoffmann-Lange „empfehlen". Durch das sekundäre Sortierkriterium Erscheinungsjahr würden die Veröffentlichungen „Der langfristige Wandel politischer Eliten in Europa 1867-2000: Auf dem Weg der Konvergenz?" (2003) von Heinrich Best und „Das pluralistische Paradigma der Elitenforschung" (2003) von Ursula Hoffmann-Lange ganz oben auf der Ergebnisliste rangieren.

Umgekehrt könnte natürlich auch Autorenzentralität als sekundäres und Erscheinungsjahr als primäres Sortierkriterium benutzt werden. Dann hätte der Benutzer immer die aktuellsten Dokumente oben in der Ergebnisliste, innerhalb der Jahresblöcke würden allerdings die Veröffentlichungen der jeweils

[23] Schlagwort = ‚Elite'

zentralsten Autoren oben rangieren. Da zentrale Autoren i.d.R. viel publiziert haben, kann die Ausgabe in der jetzigen Implementierung des Modells dahingehend gesteuert werden, dass nur die jeweils *n* aktuellsten Dokumente eines (zentralen) Autors erscheinen. Das Autorennetzwerk bietet dem Benutzer zugleich die Möglichkeit zu Koautoren zentraler Autoren zu verzweigen, um nach weiteren relevanten Dokumenten im Umfeld zentraler Autoren zu suchen.[24]

Ein besonderer Vorteil des Ex-Post-Verfahrens ist, dass mit diesem Modell eines retrospektiven Rankings ein datenbankübergreifendes Ranking durchgeführt werden kann, das von der Größe der beteiligten Datenbanken und der Häufigkeit des Vorkommens von Termen unabhängig ist. Uns ist kein System bekannt, das ein Ranking über mehrere Datenquellen unterstützt. Denn das Mischen probabilistischer Textretrieval-Rankings aus unterschiedlichen Datenquellen ist ein in der Information-Retrieval-Forschung bislang ungelöstes Problem. Mit dem auf Autorennetzwerken basierenden Ex-Post-Ranking-Verfahren liegt jedoch ein Modell vor, dass von datenquellenspezifischen Rankings unabhängig ist und ohne nachträgliche Indexierung integrierter Ergebnismengen auskommt.

Ein Nachteil dieser Vorgehensweise ist jedoch, dass aus allen beteiligten Datenquellen alle Koautoren-Relationen abgefragt werden müssen. Dies kann, gerade bei entfernten Datenquellen, u.U. zu erheblichen Performance-Einbußen führen. Rein technisch bietet es sich jedoch ohne weiteres an, bei jeder beteiligten Datenquelle einen Autorennetzwerk-Agenten zu installieren, der lokal ein Autorennetzwerk für das gesuchte Thema propagiert und dieses an den zentralen Autorennetzwerk-Agenten schickt, wo alle hereinkommenden Autorennetzwerke integriert werden.

Für das Ex-Post-Ranking von Rechercheergebnissen nach Akteurszentralität wurde ein Retrieval-Test auf der Basis von SOLIS durchgeführt (s. Kapitel 6). Der Test ergab, dass mit dem auf Akteurszentralität basierenden Rankingmodell eine deutliche höhere Precision erzielt werden konnte als mit der Standardausgabe von SOLIS (nach Erscheinungsjahr absteigend) und traditionellen Rankingverfahren wie der inversen Dokumenthäufigkeit (TF-IDF). Ein

---

[24] Eine weitere Option des Modells ist, dass sich der Benutzer z.B. auch Dokumente von Autoren mittlerer Zentralität anzeigen lassen könnte, von denen nach Mutschke & Quan Haase (2001) ja am ehesten „neue“ Ideen auszugehen scheinen. Wie sich diese Variante auf die Recherchequalität auswirken würde, müsste allerdings noch untersucht werden. Gleiches gilt für „Einzelautoren“ (Autoren mit einem Degree von 0), die durch das Zentralitätsmodell ebenfalls ausgewiesen werden.

weiteres interessantes Ergebnis dieses Tests war, dass Rankings mit Autorenzentralität offenbar ganz andere Dokumente favorisieren als auf Termvorkommen abstellende Rankingverfahren. Die Top-20-Dokumente der nach Akteurszentralität gerankten Menge überschnitten sich gut wie gar nicht mit der nach TF-IDF gerankten Menge. Das Autorenzentralität-Rankingmodell bietet dem Benutzer also eine ganz andere Sicht auf die Datenbank als herkömmliche Retrievalmodelle. Es liegt daher nahe, ein auf Zentralität in Autorennetzwerken basierendes Retrievalmodell dem Benutzer als alternativen Zugang zu einer Datenquelle neben herkömmlichen Verfahren zur Verfügung zu stellen.

Das Ex-Post-Rankingmodell mit Akteurszentralität ist in DAFFODIL[25] bereits realisiert. Die Übernahme in infoconnex[26] ist in Bearbeitung.

5.1.1.2 Ex-Ante-Ranking

Das Ex-Ante-Modell berechnet Akteurszentralität nicht retrospektiv auf der Basis von Rechercheergebnissen, sondern prospektiv für eine Gesamtkollektion und indexiert die Dokumente der Kollektion mit dem Zentralitätswert ihrer jeweils zentralsten Autoren. Bei einer Recherche würden dann Dokumente mit einem höheren Zentralitätsindex stärker gewichtet. Der Vorteil dieses Verfahrens im Unterschied zu dem Ex-Post-Ranking-Modell wäre, dass die Ausnutzung von Akteurszentralität ohne Performance-Einbußen bei der Suche verbunden wäre. Wie sich solch ein Modell auf die Retrievalqualität auswirken würde, muss allerdings noch untersucht werden.

Da eine Propagierung von Autorennetzwerken auf der Basis großer Kollektionen vermutlich zu wenig kohäsiven Netzwerken führt (s. das 100.000-Mitglieder-Netzwerk in Tab. 4), wäre sicherlich die Anwendung elaborierter Skalierungsverfahren erforderlich, die hochkomplexe Netzwerke auf kohäsive Subgruppen reduzieren können. Ein Alternative dazu wäre, Autorennetzwerke auf der Basis von Subkollektionen zu propagieren, z.B. anhand einer Klassifikation. Da ein Dokument bei dieser Variante zu mehreren Subkollektionen gehören kann (weil es z.B. mehrere Klassenbezeichner hat) und sein(e) Autor(en) daher zu mehreren Autorennetzwerken, wären für diese Variante geeignete Normalisierungsverfahren zu entwickeln, um Zentralitätswerte aus mehreren Autorennetzwerken vergleichen zu können. Ob die Propagierung mit (sinnvollen) Subkollektionen im Unterschied zum Ex-Post-Ranking sich nachteilig auf die Retrievalqualität auswirken würde oder möglicherweise so-

---

[25] www.daffodil.de

[26] www.infoconnex.de

gar ein Vorteil wäre, weil auf ganzen (Sub-)Kollektionen basierende Autorennetzwerke real existierende wissenschaftliche Communities möglicherweise korrekter abbilden als auf Rechercheergebnisse aufbauende Autorennetzwerke, müsste ebenfalls erst noch genauer untersucht werden.

5.1.2 Suche nach zentralen Akteuren (Expertensuche)

Aus den grundsätzlichen Überlegungen zu einem auf Akteurszentralität basierenden Retrievalmodell in Kapitel 5.1.1 bietet sich unmittelbar an, Autorennetzwerke und das Konzept der Zentralität von Autoren auch direkt für die gezielte Suche nach relevanten Personen zu nutzen. Grundidee ist, menschliche Experten (hier: zentrale Autoren) für ein bestimmtes Thema bzw. in einer bestimmten Community zu finden.

Für die Expertensuche auf der Basis von Autorennetzwerken bieten sich zwei Szenarios unterschiedlicher Zielrichtung an:

- die Suche nach zentralen Akteuren zu einem bestimmten Thema
- die Suche nach zentralen Akteuren im Netzwerk eines gegebenen Autors.

5.1.2.1 Zentrale Akteure in einer Dokumentenmenge

Grundidee dieses Szenarios ist es, Experten (zentrale Autoren) zu einem bestimmten Thema zu finden. Diese werden auf der Basis der Ergebnisdokumente zu einer Recherche evaluiert. Im Unterschied zu dem in Kapitel 5.1.1 beschriebenen Rankingmodell werden dem Benutzer aber nicht Dokumente angezeigt, sondern Autoren, die in dem untersuchten Forschungsfeld zentrale Positionen einnehmen.

Grundlage dieses Szenarios ist wiederum das in Kapitel 3.4.1 beschriebene Propagierungsmodell: Der Benutzer startet mit seiner Anfrage, z.B. Schlagwort = ‚Elite'. Aus den Ergebnismengen der abgefragten Datenbanken werden die Koautoren-Relationen für die Propagierung des Autorennetzwerkes gewonnen. Die Zentralität der Akteure wird berechnet und dem Benutzer werden Autoren angezeigt, die zentrale Positionen in der jeweils untersuchten Community einnehmen. Für das Beispiel ‚Elite' (auf der Basis von SOLIS und FORIS) wären das z.B. Heinrich Best und Ursula Hoffmann-Lange.

In einem um zusätzliche High-Level-Services erweiterten Szenario wäre denkbar, dass der Benutzer sich das wissenschaftliche Profil zentraler Autoren anzeigen lassen kann, das neben den Veröffentlichungen auch Links auf die

Homepage und Konferenzen des Autors sowie Beschreibungen zu dessen Forschungsaktivitäten enthalten könnte.

5.1.2.2 Zentrale Akteure im Netzwerk eines Autors

Experten lassen sich aber nicht nur themenbezogen auf der Basis einer Dokumentenmenge ermitteln, sondern auch themenunabhängig im sozialen Netzwerk eines bestimmten Autors. Die Kernthese hier ist, dass Autoren über ihre Kooperationsbeziehungen in ihrem näheren oder weiteren strukturellen Umfeld mit Experten ihres Faches verbunden sind. Das Konzept dieses Szenarios ist es also, diese Experten zu finden. Solch ein Szenario erscheint sinnvoll, wenn der Benutzer einen Autor kennt und zentralere Akteure im Umfeld dieses Autors finden möchte, ohne bestimmte Inhalte spezifizieren zu müssen wie in dem vorherigen Szenario.

Ein methodischer Grundansatz zur Operrationalisierung dieses Szenarios sind die aus der sozialen Netzwerkanalyse bekannten *Ego-zentrierten Netzwerke*, die hier in ihrer um indirekte Beziehungen erweiterten Form verwendet werden. Der Benutzer startet also mit einem Autorennamen (Ego). Die Generierung des Autorennetzwerkes von Ego erfolgt gemäß dem in Kapitel 3.4.2 beschriebenen Propagierungsmodell per Breitensuche. Der Breitensuchprozess stellt für Ego eine Koautoren-Anfrage an die infoconnex-Suchmaschine, die diese auf den jeweils angeschlossenen Datenbanken zur Ausführung bringt und das Ergebnis, also die Koautoren von Ego, an den Breitensuchprozess zurückgibt. Die Koautorenrelationen werden in das Koautorennetzwerk von Ego eingepflegt und es werden neue Koautorenanfragen für die Koautoren von Ego generiert. Dieses Verfahren wiederholt sich bis der Prozess gemäß der Parametrisierung durch den Benutzer terminiert.

Zur Zeit werden folgende Terminierungsparameter unterstützt:

- Tiefenschwellwert (l-steps): Der Benutzer definiert die maximale Entfernung l (=Zahl der Links) von Alter-Akteuren zu Ego (s. Kapitel 3.4.2). Der Standardwert für l ist 1, d.h. es werden alle direkten Koautoren von Ego sowie deren Verbindungen untereinander identifiziert. Der Tiefenschwellwert kann beliebig hoch gesetzt werden. Ein Wert von –1 bedeutet, dass über den gesamten Korpus propagiert wird. Ein Wert größer 2 führt aufgrund des i.d.R. exponentiellen Wachstums Ego-zentrierter Autorennetzwerke zu erheblichen Performanceeinbußen. Daher kann der Propagierungsprozess mit zusätzlichen Parametern optimiert werden, den k-cores und den m-paths.

- *k*-cores: Mit diesem Parameter spezifiziert der Benutzer die minimale Anzahl *k* an Verbindungen, die ein Alter-Akteur von Ego haben muss. Es wird also nur mit Autoren weiterpropagiert, die ihrerseits mindestens *k* Koautoren außer Ego haben (s.a. Kapitel 3.3.1).

- *m*-paths (Mutschke 2001, 2003): Dieser Parameter stößt eine Prioritätssuche an, die nur mit den jeweils *m* besten Koautoren weiterpropagiert und damit *main paths* der Kooperation durch ein Autorennetzwerk beschreibt. Die besten Koautoren sind jeweils diejenigen, die unter den Koautoren eines betrachteten Autors den höchsten Degree, also die meisten Verbindungen, haben (s.a. Kapitel 3.3.2). Da der Suchraum mit diesem Parameter drastisch eingeschränkt wird, kann der Benutzer ohne Performanceverlust beliebig tief in die Struktur eines Autorennetzwerkes vorstoßen und somit auch entferntere zentrale Akteure finden. Da die Main-Paths-Propagierung an lokalen Maxima terminiert, d.h. bei Akteuren, die keinen „besseren" (bisher noch nicht gefundenen) Koautoren haben, muss bei Verwendung des *m*-paths-Parameter kein Tiefenschwellwert mehr gesetzt werden.

Auf der Basis des so propagierten Autorennetzwerkes wird die Zentralität der Akteure berechnet. Der Benutzer erhält eine Liste der Akteure im Umfeld von Ego, nach deren Zentralität absteigend sortiert. Im Falle des Ego-zentrierten Netzwerkes des Autors in SOLIS und FORIS mit einer Schrittweite von 1 (s. Abbildung 5 in Kapitel 3.4.2) wäre dies z.B. Jürgen Krause. Die Zentralität von Jürgen Krause ist in diesem Netzwerk nicht unbedingt zu erkennen, da das Netzwerk aufgrund seiner Vollständigkeit relativ unübersichtlich ist. Mit den Skalierungsoptionen *k*-cores und *m*-paths kann die Komplexität der Netzwerke erheblich reduziert werden kann, so dass auch das weitere Umfeld von Ego auf performante Weise propagiert und dargestellt werden kann. Abbildung 11 visualisiert das 2-steps/6-core-Autorennetzwerk des Autors[27] und weist nun Heinrich Best als zentralsten Akteur im weiteren Kooperationsumfeldes von Ego (‚Peter Mutschke') aus.

[27] Das Netzwerk enthält also alle direkten Koautoren von Ego sowie deren Koautoren, sofern sie mindestens sechs Verbindungen haben.

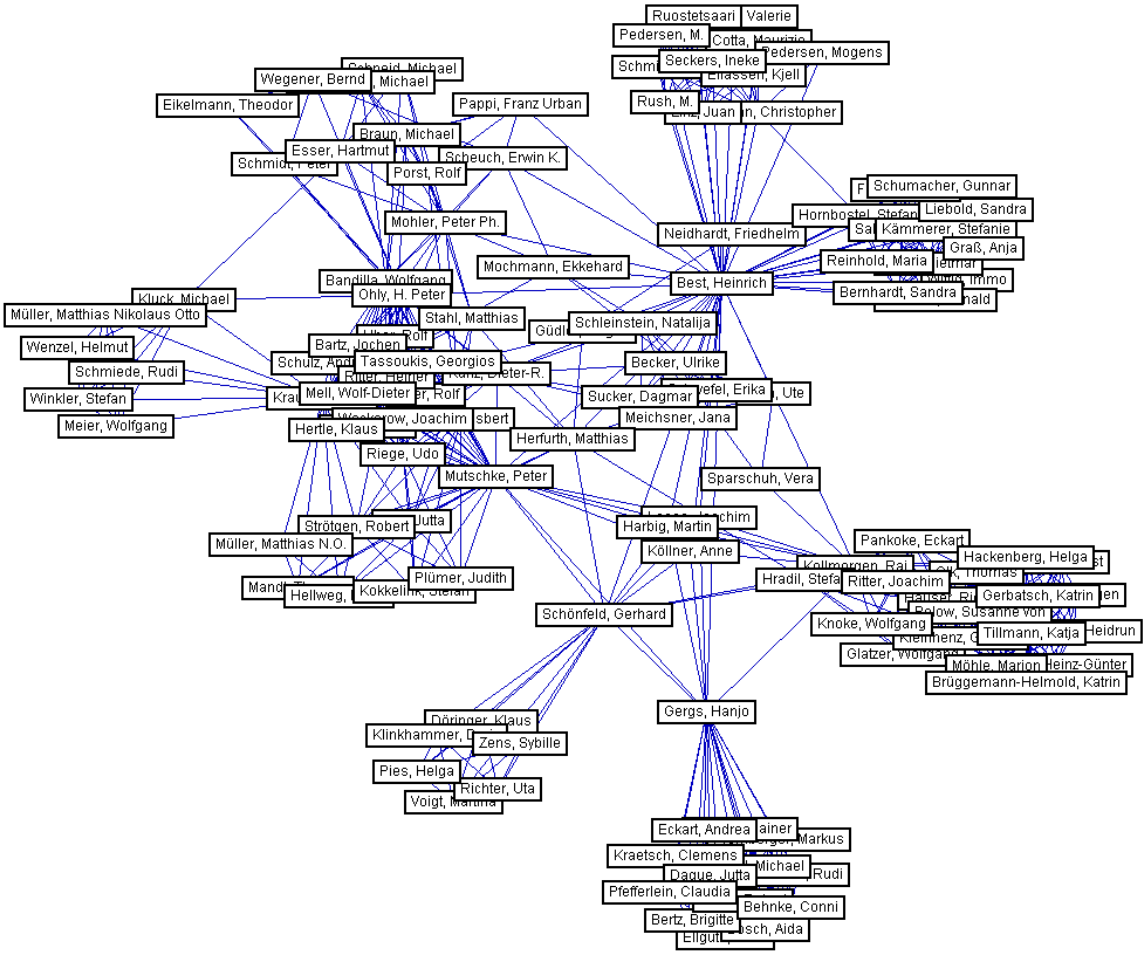

Abb. 11: Ein erweitertes Ego-zentriertes Autorennetzwerk von ‚Peter Mutschke' in SOLIS und FORIS (l=2, k=6)

Abbildung 12 stellt das 2-paths-Koautorennetzwerk[28] von ‚Peter Mutschke' in SOLIS und FORIS dar. Das Netzwerk hat jetzt mit einem Durchmesser von 4 eine viel tiefere Struktur und enthält jetzt auch von Ego weiter entfernte prominente Autoren wie Erwin K. Scheuch, Walter Müller, Renate Mayntz und Jürgen Friedrichs, die bei einem Tiefenschwellwert von 1 oder 2 nicht gefunden worden wären. Die terminalen Knoten sind hier schwarz umrandet dargestellt (Walter Müller, Josef Schmid, Jürgen Friedrichs).

[28] Das Netzwerk enthält also, ausgehend von ‚Peter Mutschke', nur noch die jeweils zwei degree-zentralsten Koautoren eines betrachteten Autors.

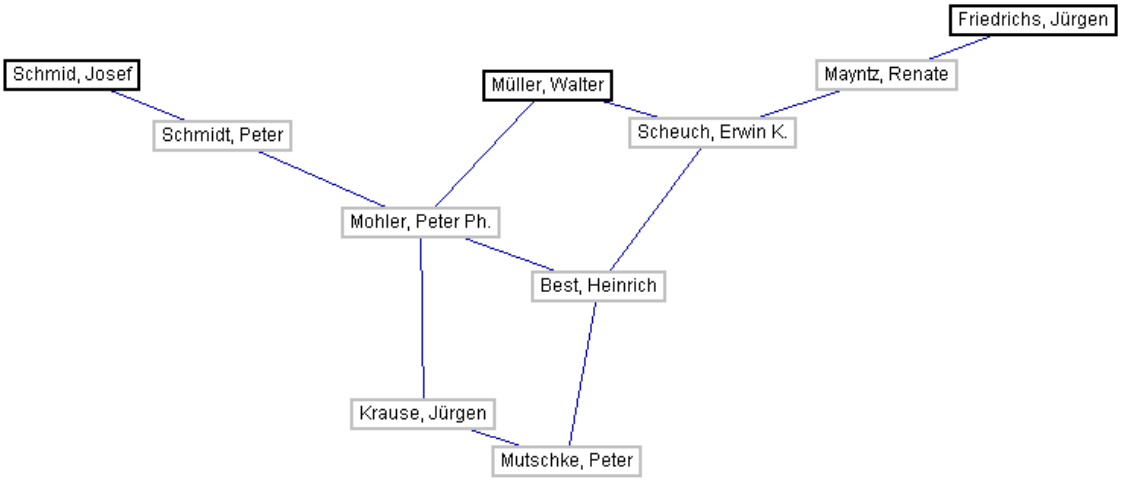

Abb. 12: Das 2-paths-Autorennetzwerk von ‚Peter Mutschke' in SOLIS und FORIS

Die Zentralität spielt in dem *m*-paths-Graphen jedoch eine untergeordnete Rolle, da das Netzwerk ohnehin nur noch zentrale Akteure enthält. Das Beispiel demonstriert, dass das *m*-paths-Verfahren eine performante Möglichkeit bietet, direkt zu den Hauptakteuren einer wissenschaftliche Community vorzustoßen, auch wenn sie aus Ego's Perspektive tief in der Struktur „verborgen" sind.

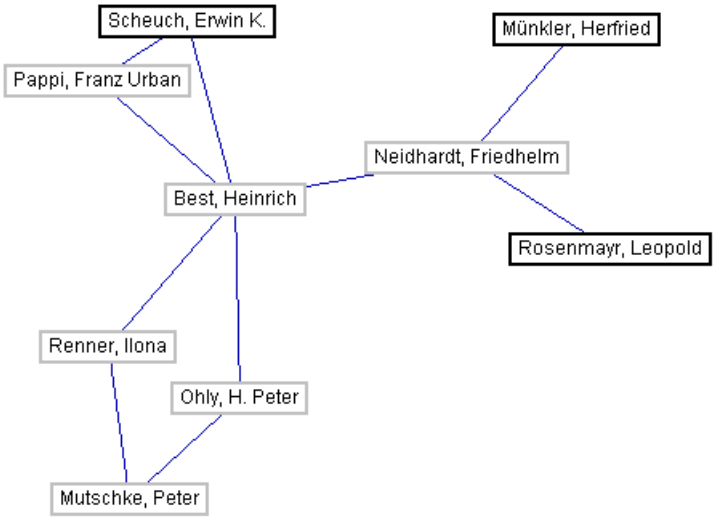

Abb. 13: Das 3-paths-Autorennetzwerk von ‚Peter Mutschke' zum Thema ‚Wissenschaft' in SOLIS und FORIS

Die Suche nach zentralen Akteuren im Umfeld eines bestimmten Ego-Autors kann auch mit inhaltlichen Suchbedingungen verknüpft werden. Die Netzwerkpropagierung erfolgt dann immer nur mit jeweils den Koautoren, die diese Suchbedingungen erfüllen, also z.B. mit Autoren, die in einem bestimmten Forschungsfeld tätig sind. Abbildung 13 visualisiert das 3-paths-Netzwerk

von ‚Peter Mutschke' zum Thema ‚Wissenschaft'[29] in SOLIS und FORIS und weist u.a. Erwin K. Scheuch, Friedhelm Neidhardt und Leopold Rosenmayr als Experten zum Thema ‚Wissenschaft' im weiteren Umfeld von ‚Peter Mutschke' aus.

Die Umsetzung dieses Szenarios ist in DAFFODIL[30] realisiert.

5.1.3 Alerting mit Autorennetzwerken

Autorennetzwerke lassen sich aber nicht nur für die Suche in Datenbanken nutzen, sondern auch für die Unterstützung von Zusammenarbeit zwischen Wissenschaftlern nutzen. Ein individueller Alerting-Service auf der Basis von Autorennetzwerken könnte z.B. dafür sorgen, dass ein Benutzer nicht über alle für ihn inhaltlich interessanten Neuzugänge in einer Datenbank benachrichtigt wird, sondern nur über die, die aus seinem *Netzwerk* stammen (s. Abb. 14).

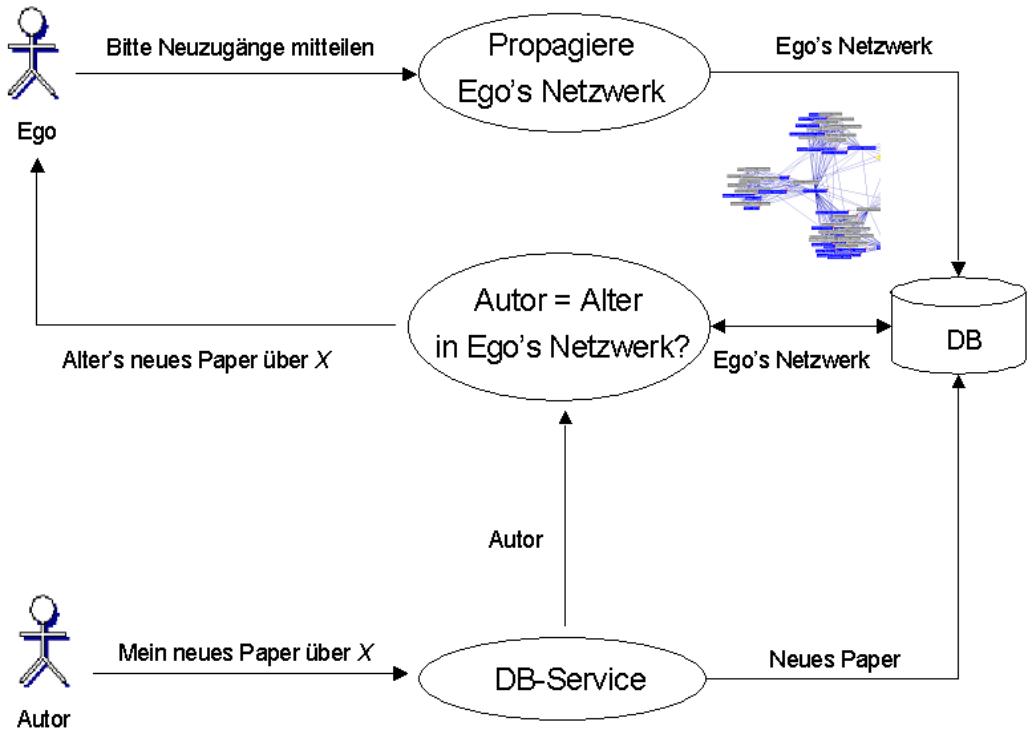

Abb. 14: Alerting unter Nutzung von Ego-zentrierten Autorennetzwerken

Das Ego-zentrierte Netzwerk des Benutzers hätte in diesem Fall eine Filterfunktion bezüglich der für ihn potentiell interessanten Neuzugänge (social

[29] Schlagwort = Wissenschaft* (rechtstrunkiert)

[30] www.daffodil.de

filtering). Im Unterschied zu traditionellen, allein an Inhalten orientierten Benachrichtigungsdiensten würde ein auf Ego-zentrierten Netzwerken basierender Alerting-Dienst auch noch die Möglichkeit bieten, dass ein Benutzer von *allen* Neuerungen erfährt, die in seinem *Netzwerk* stattfinden.

Eine innovative Perspektive dieses Modells wäre, auf der Basis eines um indirekte Beziehungen erweiterten Ego-zentrierten Netzwerkes auch weiter entfernte Alteri in das Netzwerk eines bestimmten Benutzers (Ego) aufzunehmen. Inwieweit dies sinnvoll ist, müsste jedoch noch untersucht werden.

6 Heuristische Evaluation von Autorennetzwerk-Retrievalmodellen

Die Leistungsfähigkeit des hier vorgeschlagenen Autorennetzwerkmodells und seines Zentralitätskonzeptes für die Suche in Datenbanken wurde anhand eines Retrieval-Tests unter Verwendung des Dokumentenranking-Szenarios (s. Kapitel 5.1.1.1) getestet. Der Retrievaltest misst die *Precision* einer nach Autorenzentralität gerankten Ergebnismenge, also den Grad, in dem die Ergebnismenge einer Recherche nur *relevante* Dokumente liefert.

Der Retrievaltest wurde für zehn Boolesche Schlagwortanfragen (s. Tab. 5) auf der Basis der IZ-Datenbank SOLIS durchgeführt[31]. Alle zehn Anfragen waren von der Form (a_1 or a_2 ... or a_n) and (b_1 or b_2 ... or b_n), d.h. sie bestanden aus zwei mit UND verknüpften Konzepten, die ihrerseits aus einer Serie mit ODER verknüpfter Terme aufgebaut waren.

Um einen Vergleich zu ermöglichen, wurde jede Ergebnismenge der zehn Anfragen nach folgenden drei Verfahren aufbereitet:

- Sortierung der Ergebnisdokumente nach Erscheinungsjahr absteigend, die Standardausgabe von SOLIS (PY)
- Ranking der Dokumente nach Closeness-Zentralität ihrer Autoren im Autorennetzwerk der Ergebnismenge gemäß Szenario 5.1.1.1 (ACL)

[31] Es sind einige Fragen aus dem GIRT-Test übernommen worden. Da jedoch nicht alle GIRT-Fragen hinreichend große Ergebnismengen lieferten, wurden Anfragen aus dem Erfahrungshorizont des Autors hinzugenommen.

- Ranking der Dokumente nach inverser Dokumenthäufigkeit (IDF), einem Standard-Rankingverfahren im Text-Retrieval, das auf die Häufigkeit von Anfragetermen im Dokument in Relation zu ihrem Vorkommen im Gesamtkorpus abstellt. Hierfür wurde der contains-Operator von ORACLE TEXT in der *accumulate*-Scoring-Variante verwendet, der dem IDF-Verfahren entspricht. Durchsucht wurden Titel und Abstracts. Der accumulate-Operator sucht nach Dokumenten, die mindestens einen Anfrageterm enthalten und rankt die gefundenen Dokumente gemäß des Gewichts der Anfrageterme in den Dokumenten.

Die Precision der Ergebnismengen wurde anhand der Zahl hochrelevanter Dokumente unter den ersten 20 Dokumenten der gerankten Trefferliste evaluiert. Dividiert man diese Zahl durch 20, erhält man einen Wert zwischen 0 und 1. Je höher dieser Wert, desto besser die Precision des Resultsets. Als Evaluationskriterium für die Relevanz eines Dokumentes wurde das Vorkommen der Anfrageterme, einschließlich ihrer Synonyme, im Titel des Dokumentes verwendet.

Query	Ergebnismenge sortiert nach			Information. Mehrwert
	PY	*IDF*	*ACL*	*ACL*
Jugend – Gewalt	0.25	0.60	0.55	92
Rechtsextremismus – Ostdeutschland	0.35	0.45	0.60	122
Frau – Personalpolitik	0.35	0.60	0.65	100
Widerstand – Drittes Reich	0.40	0.65	0.95	138
Zwangsarbeit – II. Weltkrieg	0.55	0.65	0.70	92
Eliten – BRD	0.40	0.70	0.85	107
Armut – Stadt	0.30	0.35	0.55	157
Arbeiterbewegung – 19./20. Jahrh.	0.55	0.55	0.90	164
Wertewandel – Jugend	0.40	0.50	0.30	50
Terrorismus - Demokratie	0.20	0.35	0.60	129
Durchschnitt	**0.38**	**0.54**	**0.67**	**115**

Tab. 5: Precision von zehn Resultsets aus SOLIS (PY = Erscheinungsjahr, IDF = Inverse Dokumenthäufigkeit , ACL = Autor-Closeness)

Wie in Tabelle 5 ersichtlich, erreichten in fast allen Fällen die nach Autorenzentralität gerankten Resultsets eine deutlich höhere Precision als die nach Erscheinungsjahr sortierten Ergebnismengen und die nach IDF gerankten Resultsets: Die ACL-Rankings erzielten durchschnittlich eine Precision von 0.67, während die PY-Mengen lediglich eine Precision von 0.38 im Durchschnitt erreichten. Durch das Ranking nach Autorenzentraliät konnte die Retrievalqualität gegenüber der Standardausgabe (PY) also um 76% verbessert werden. Nicht ganz so ausgeprägt ist der Unterschied zwischen den ACL- und

den IDF-Rankings: Aber auch hier beträgt die Verbesserung immerhin 24%. Weitere, v.a. umfangreichere Tests sind hier allerdings erforderlich.

Ein weiteres, möglicherweise viel interessanteres Ergebnis des Tests ist, das Autorenzentralität basierte Rankings offenbar ganz andere Dokumente favorisieren als die traditionellen IDF-Rankings: Die meisten Dokumente unter den ersten 20 in den ACL-Rankinglisten tauchten in den IDF-Top-20 nicht auf. Der relative Anteil der relevanten Dokumente unter den ACL-Top-20, die in den IDF-Top-20 nicht auftauchten, betrug – bezogen auf die Zahl der relevanten Dokumente unter den IDF-Top-20 – mehr als 100%. Die Unterschiedlichkeit der Top-Dokumente weist darauf hin, dass offensichtlich kein Zusammenhang besteht zwischen der Relevanz von Anfragetermen für Dokumente und der Relevanz der Autoren in einer Community. Die Unterschiedlichkeit der Rankings dürfte deshalb umso größer sein, je umfangreicher eine Ergebnismenge ist, da die Wahrscheinlichkeit, dass unter den Top-Dokumenten eines IDF-Rankings ein nach Autorenzentralität relevantes Dokument dabei ist, mit der Größe des Resultsets sinkt.

Autorennetzwerk-Retrievalmodelle bieten also offensichtlich eine ganz andere Sicht auf eine Dokumentkollektion als traditionelle termbasierte Verfahren, deren Ergebnismengen sich oftmals erheblich überlappen. Dieses Ergebnis zeigt, dass Retrievalmodelle, die auf Zentralität in Autorennetzwerken basieren, einen erheblichen informationellen Mehrwert haben können. Ein Ranking nach Autorenzentralität ist demnach eine sinnvolle Alternative bzw. *Ergänzung* zu traditionellen Retrievalmodellen. Es bietet sich daher an, dem Benutzer eines Informationssystems mehrere, unterschiedliche Sichten auf die Datenbanken bietende Retrievalverfahren zugänglich zu machen.

7 Zusammenfassung und Ausblick

Die in diesem Bericht beschriebenen empirischen Studien zu Autorennetzwerken in wissenschaftlichen Communities haben gezeigt, dass diese Netzwerke ein erhebliches Potential für Informationssysteme haben. Die Untersuchungen zu Evolution und Topologie von Autorennetzwerken (Small-World-Architekur, Zentrum-Peripherie-Muster, Preferential-Attachment) unterstreichen die Leistungsfähigkeit des Zentralitätskonzeptes auch für soziale Netzwerke in wissenschaftlichen Kooperationsstrukturen. Autorennetzwerke lassen sich daher nicht nur zum Browsen in den Kooperationsstrukturen eines Forschungsfeldes verwenden, sondern auch für die Evaluation der Relevanz der Autoren aufgrund ihrer strategischen Position in der globalen sozialen Struktur ihrer Community (Zentralität).

Das Konzept der datenbankbasierten Analyse von Autorennetzwerken als Mehrwertdienst für Informationssysteme, und hier primär das Konzept der Zentralität der Autoren in solchen Netzwerken, wurde am Beispiel einiger Anwendungsszenarios exemplifiziert. Autorenzentralität lässt sich sinnvoll nutzen für das Ranking von Rechercheergebnissen und für die Suche nach Experten in einem bestimmten Forschungsfeld oder im strukturellen Umfeld einer gegebenen Person. Hierfür wurden verschiedene Varianten vorgestellt. Weitere Anwendungsmöglichkeiten für die Nutzung von Autorennetzwerken in Informationssystemen wären z.B. die Erkennung von Brückenpersonen und Forschercliquen oder die Erweiterung von Ergebnismengen über Koautorenschaften.

Die Ergebnisse der heuristischen Evaluation geben erste Hinweise, dass Rankings auf der Basis von Autorenzentralität die Retrievalqualität erheblich verbessern können. Weitere Untersuchungen hierzu sind allerdings noch erforderlich. Der Test zeigte auch, dass die Nutzung von Autorenzentralität beim Ranking den Benutzer offenbar zu ganz anderen relevanten Dokumenten hinführt als traditionelle termbezogene Rankingverfahren wie z.B. das Vektorraummodell. Es bietet sich daher an, dem Benutzer ein auf Autorenzentralität basierendes Ranking als Rechercheoption neben den herkömmlichen Retrievalmodellen an die Seite zu stellen.

Autorennetzwerke ließen sich aber nicht nur für die Suche in Datenbanken nutzen, sondern – wie das Alerting-Szenario zeigt – auch für die Unterstützung von Zusammenarbeit zwischen Wissenschaftlern. Ein Alerting-Dienst könnte z.B. auch so konfiguriert werden, dass ein Benutzer nur über Neuerungen aus seinem *Netzwerk* benachrichtigt wird (social filtering). Auch hier sind weitere Szenarios denkbar.

Eine offene Frage ist nach wie vor, wie man Autorennetzwerke in einem Informationssystem so visualisiert, dass ein (unkundiger) Benutzer mit der Visualisierung auch *arbeiten* kann. Die herkömmlichen Graphenlayout-Verfahren (spring embedder usw.) reichen hier sicherlich bei weitem nicht aus (zumal diese Verfahren vermutlich auch kaum noch zu verbessern sind). Ähnliches gilt für Verfahren der Komplexitätsreduktion von Graphen. Hier ist weitere Forschung notwendig, die geeignete Ansätze zur Komplexitätsreduktion mit Konzepten aus der Softwareergonomie und des Grafikdesigns integrieren müsste.

Ein weiteres offenes Problem betrifft die Entwicklung von Kriterien für die *Adäquatheit* von Netzwerkmodellen im Information Retrieval. Relevante Konzepte hierfür wären die Größe der Netzwerke, ihre interne Kohärenz, ihre

Robustheit und vor allem ihr Community-Charakter. Auch hier ist noch weitere Forschung notwendig.

Autorennetzwerke sind allerdings nur *ein* Beispiel für die Nutzung von Netzwerkanalyseverfahren in Informationssystemen. Weitere interessante Anwendungsperspektiven wären die netzwerkanalytische Betrachtung von Kooperationsbeziehungen zwischen Institutionen, von Shared-Interest-Netzwerken (z.B. anhand gemeinsam verwendeter Begriffe bei Autoren und Institutionen), von Begriffsnetzwerken (um z.B. die Zentralität von Thesaurustermen für das Ranking von Dokumente zu benutzen), oder von Linkstrukturen in Clearinghouse-Daten.

8 Literatur

Ballay, A., Markham, R., Mutschke, P., Stempfhuber, M. (2004): Der Informationsverbund Bildung-Sozialwissenschaften-Psychologie (infoconnex). Education, Research and New Media. Chances and Challenges for Science. 10. IuK-Frühjahrstagung, Darmstadt, 15. – 18. März 2004

Barabasi, A.-L. et al. (2002): Evolution of the social network of scientific collaborations. Physica A **311**, 590-614.

Brandes, U. (2001): A faster algorithm for betweenness centrality. Journal of Mathematical Sociology **25**, 163-177

Brandes, U., Cornelsen, S. (2003): Visual Ranking of Link Structures. Journal of Graph Algorithms and Applications, vol. 7, no. 2, pp. 181–201

Burt, R.S. (1982): Towards a structural theory of action. New York: Academic Press

Carley, K., Hummon, N., Harty, M. (1993): Scientific influence. An analysis of the main path structure in the journal of conflict resolution. Knowledge: Creation, Diffusion, Utilization **14**, 417–447

Coleman, J.S., Katz, E., Menzel, H. (1966): Medical Innovation: A Diffusion Study. Bobbs Merrill, New York, 1966

Crane, D. (1972): Invisible Colleges: Diffusion of Knowledge in Scientific Communities, University of Chicago Press, Chicago

Freeman, L. C. (1979): Centrality in social networks: Conceptual clarification. Social Networks **1** 215-239

Fuhr, N., Schaefer, A., Klas, C.-P., Mutschke, P.: Daffodil (2002): An Integrated Desktop for Supporting High-Level Search Activities in Federated Digital Libraries. In: Agosti, M., Thanos, C. (eds.): Research and Advanced Technology for Digital Libraries. 6th European Conference, EDCL 2002, Proceedings. Lecture Notes in Computer Science, Vol. 2458. Springer-Verlag, Berlin - Heidelberg - New York 597-612

Granovetter, M. (1973): The strength of weak ties. American Journal of Sociology 78, 1360-1380

Granovetter, M. (1985): Economic action and social structure. The problem of emdeddedness. American Journal of Sociology, 91, 481-510

Güdler, J. (2003): Kooperation in der Soziologie. Langfristige Entwicklungen, strukturbildende Wirkungen, individuelle Platzierungseffekte. Dissertation. Forschungsberichte, Band 5. Informationszentrum Sozialwissenschaften

Hummon, N., Doreian, P. (1989): Connectivity in a citation network: The development of DNA theory. Social Networks 11, 39–63

Jansen, D. (2003): Einführung in die Netzwerkanalyse. Leske + Budrich

Krause, Jürgen (2003): Suchen und "Publizieren" fachwissenschaftlicher Informationen im WWW. In: Audiovisuelle Medien online: Tagung: "Audiovisuelle Wissensmedien online"; Informationsveranstaltung der IWF Wissen und Medien gGmbH, Göttingen, 03.12. - 04.12.2002. Wien: Lang. (IWF: Menschen, Wissen, Medien)

Kautz, H., Selman, B., Shah, M. (1997): The Hidden Web. AI Magazine 18 (2), 27-36

Kleinberg, J. M. (1999): Authoritative sources in a hyperlinked environment. Journal of the Association for Computing Machinery **46**(5) 604-632

Knoke, D., Burt, R.S. (1983): Prominence, in: Burt, R.S., Minor, M.J. (Hrsg.): Applied network analysis. Beverly Hills: Sage, S. 195-224

Mutschke, P. (1994): Processing Scientific Networks in Bibliographic Databases. In: Bock, H.H., et al. (eds.): Information Systems and Data Analysis. Prospects-Foundations-Applications. Proceedings 17th Annual Conference of the GfKl 1993. Springer-Verlag, Heidelberg Berlin 127-133

Mutschke, P., Renner, I. (1995): Wissenschaftliche Akteure und Themen im Gewaltdiskurs. Eine Strukturanalyse der Forschungslandschaft. In: Mochmann, E., et al. (eds.): Gewalt in Deutschland. Soziale Befunde und Deutungslinien. Oldenbourg Verlag, München 147-192

Mutschke, P. (1996): Uncertainty and Actor-Oriented Information Retrieval in μ-AKCESS. An Approach Based on Fuzzy Set Theory. In: Bock, H.-H. et al. (eds.): Data Analysis and Information Systems. Statistical and conceptual approaches. Springer-Verlag, Berlin-Heidelberg 126-138

Mutschke, P., Quan Haase, A. (2001): Collaboration and Cognitive Structures in Social Science Research Fields: Towards Socio-Cognitive Analysis in Information Systems. Scientometrics **52** (3) 487-502

Mutschke, P. (2001): Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems. In: Constantopoulos, P., Sölvberg, I.T. (eds): Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Proceedings. Lecture Notes in Computer Science; Vol. 2163. Springer-Verlag, Berlin – Heidelberg – New York, 287-299

Newman, M.E.J. (2001a): Who is the best connected scientist? A study of scientific co-authorship networks. Phys. Rev. **E64** 016131

Newman, M.E.J. (2001b): The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98** 404-409

Newman, M.E.J. (2001c): Clustering and preferential attachment in growing networks, Phys. Rev. **E64**, 025102.

Newman, M.E.J. (2004): Coauthorship networks and patterns of scientific collaboration, Proc. Natl. Acad. Sci. USA, in press

Palmer, E. (1985): Graphical Evolution. Wiley, New York

Tallberg, C. (2000): Centrality and random graphs. Technical Report 7, Stockholm University, Department of Statistics

Wassermann, S., Faust, K. (1994): Social Network Analysis: Methods and Applications, Cambridge University Press, New York

Watts, D.J. (1999): Small Worlds. Princeton: Princeton University Press

Wellman, B. (1988): Structural analysis: From method and metaphor to theory and substance. In: Wellmann, B., Berkovitz (Hrsg.): Social structures: A network approach. Cambridge: Cambridge University Press, 2. Auflage, S. 19-61